Tracing Causal Paths from Experimental and Observational Data*

Xiang Zhou Teppei Yamamoto

Harvard University MIT

November 16, 2020

Abstract

Much of political science involves the study of causal mechanisms, and causal mediation analysis has grown rapidly across different subfields over the past decade. 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.

*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 excellent research assistance.

1 Introduction

Much of political science involves the study of causal mechanisms. 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. Baker 2015; Brader et al. 2008; Druckman and Nelson 2003; Nelson et al. 1997; 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; Nunn and Wantchekon 2011). 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 theoretical models 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; Robins 2003; VanderWeele 2015), which are also known as the average direct effect (ADE) and average causal mediation effect (ACME), respectively (Imai et al. 2010, 2011).

Despite its conceptual simplicity, this approach faces an important limitation when the effect of the treatment on the outcome involves multiple, potentially overlapping, causal pathways—a common scenario in 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, the last assumption is especially restrictive because it requires that there must not be any posttreatment variables that affect both the mediator and outcome, whether

they are observed or not.

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 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 refine, reformulate, and reassess the "competing hypotheses" of underlying processes. The prevailing practice of treating causally dependent mediators as independent can thus 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 literature on causal mediation analysis with multiple mediators. 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; Lin and VanderWeele 2017; Miles et al. 2020). Our general formula nonparametrically identifies the $K+1$ PSEs 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 causal relationships, a standard identification assumption or causal effects in observational studies (Pearl 2009).

Second, we develop a new method for estimating the PSEs identified by the decomposition formula. 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 is well suited to handle 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 variables. 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, only involves modeling 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 only needs correct modeling assumptions 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 new sensitivity analysis for the conditional ignorability assumptions required for the identification of PSEs. Although these assumptions are standard in the 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 the ignorability assumption for the mediator-outcome relationships — a key assumption that could be violated 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 and a finer decomposition of the treatment effect into multiple PSEs, each corresponding to one of the mediators, as well as providing a sensitivity analysis for the key identification assumptions. Applied researchers can thus adopt our framework to make richer inferences about how the treatment affects the outcome of interest through multiple possible causal pathways. To aid the application of our framework in empirical research, we offer an open source R package, `paths`, which implements all of the proposed methods.

In the rest of the paper, we first introduce two empirical examples where researchers have endeavored to disentangle causal pathways in the presence of multiple, causally dependent mediators. Then, for ease of exposition, we start with the case of two causally ordered mediators, where we present a decomposition of the total effect of treatment into a set of PSEs, review 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 and, for this general setting, describe the bias factor approach to sensitivity analysis. Finally, we illustrate these methods using our two empirical examples.

2 Two Empirical Examples

2.1 Issue Framing Effects

In political psychology, scholars have long been interested in 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. 1997). 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. 1997; Nelson and Oxley 1999; Slothuus 2008; Zaller et al. 1992). 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 differently 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) points out on the basis of her experimental study, "individuals use information obtained from the media to evaluate how important issues are" (p. 711) and "when media exposure to an issue causes negative emotional reactions about the issue, increased importance judgments will follow" (p. 712). Moreover, Imai and Yamamoto's (2013) reanalyses of Slothuus's data suggest that the independence assumption is unlikely to hold in this application (p. 153). If this is the case, the ACME of the importance mediator cannot be nonparametrically identified. Yet, as we will show, we can still identify the strength of the causal path *issue frame*$\rightarrow$*importance*$\rightarrow$*support for welfare reform*, i.e., 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.

2.2 The Legacy of Political Violence

In recent years, political scientists have begun to study the legacy of war and violence on contemporary political attitudes and behavior (e.g., Balcells 2012; Lupu and Peisakhin 2017; Rozenas et al. 2017). While empirical evidence on long-term impacts of political violence is growing, few studies have demonstrated how these legacies are transmitted across generations through the family. To bridge this gap, Lupu and Peisakhin (2017) 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" (p. 837). Leveraging this variation in violent victimization, the authors found that the grandchildren of individuals who suffered more deaths of family members support more strongly the Crimean Tatar political leadership, hold more hostile attitudes toward Russia, and participate more in politics.

To investigate the 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. However, as we will show, we can still identify and estimate a set of PSEs via the identities of first-, second-, and third-generation respondents, which, together with a "direct effect," constitute the total effect of ancestor victimization.

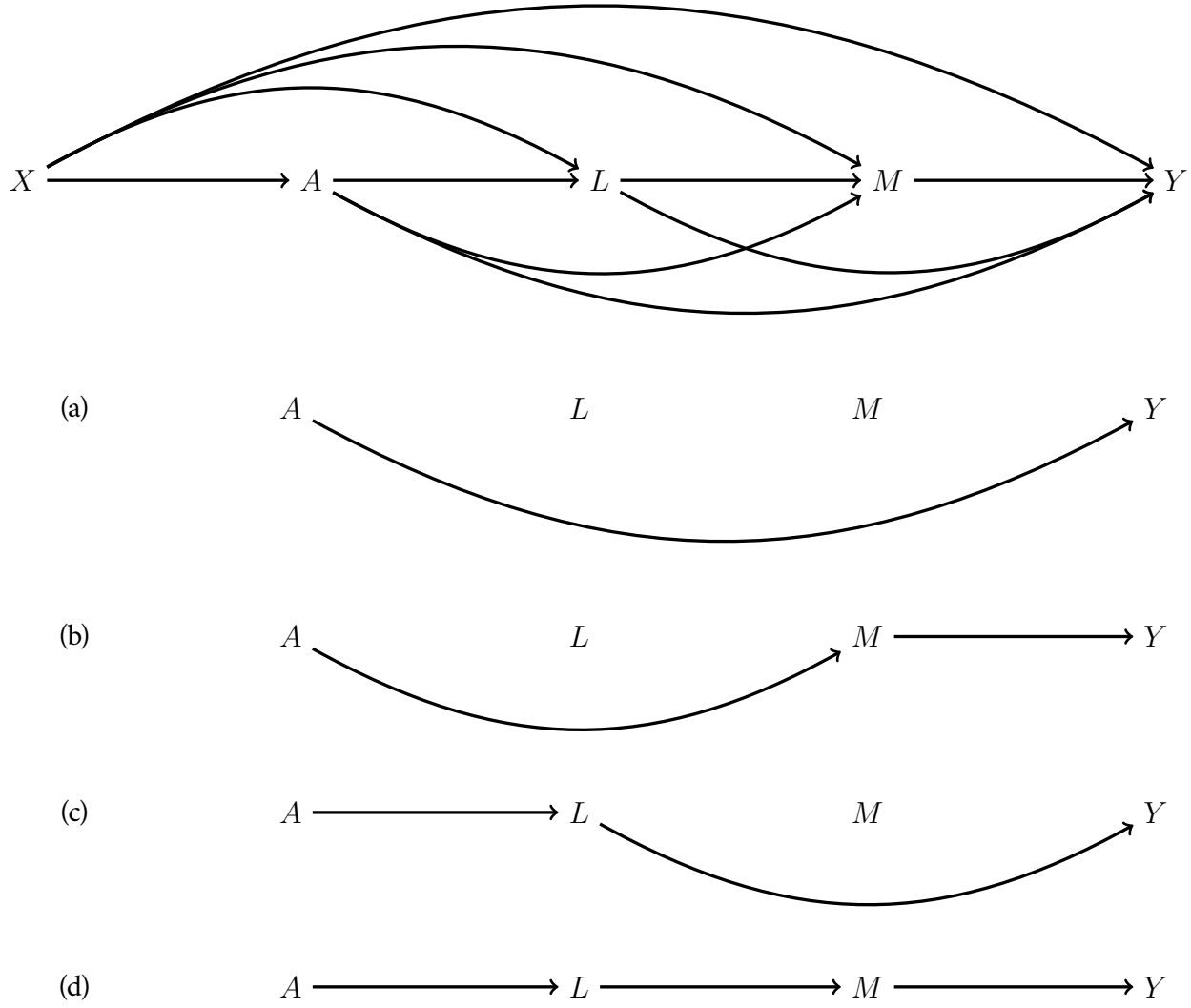

Figure 1: Causal Relationships with Two Causally Dependent Mediators Shown in a DAG.

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

3 Path-Specific Causal Effects

3.1 Definition of Path-Specific Effects

We use A to denote a binary treatment, Y to denote the observed outcome, and X to denote a vector of observed pretreatment confounders. Although our framework can accommodate any number of

mediators, let us first consider the case where two (sets of) mediators, L and M, lie on the causal paths from A to Y, for the ease of exposition. We assume that L precedes M, such that no component of M can causally affect any component of L.[1] A causal DAG that is consistent with the relationships between these variables is shown in the top panel of Figure 1. For example, in Slothuus's (2008) study on issue framing effects, A denotes the issue frame presented to the respondent, Y denotes the respondent's support for the proposed welfare reform, L denotes the respondent's beliefs about why some people receive welfare benefits, and M denotes the respondent's perceived importance of competing considerations.

In this causal DAG, four possible paths exist from the treatment to the outcome, as shown in the bottom panels of Figure 1: (a) $A \rightarrow Y$; (b) $A \rightarrow M \rightarrow Y$; (c) $A \rightarrow L \rightarrow Y$; and (d) $A \rightarrow L \rightarrow M \rightarrow Y$. If the mediators L and M are causally independent, i.e., if they do not affect each other, the last path does not exist. In this case, the total effect of A on Y can be partitioned into the effect operating through L ($A \rightarrow L \rightarrow Y$), the effect operating through M ($A \rightarrow M \rightarrow Y$), and a "direct" effect not operating through L or M ($A \rightarrow Y$) (Imai and Yamamoto, 2013). However, in the general case where L and M are causally dependent, it is not possible to partition the mediating effects of L and M into their respective components, since some of the total effect of A on Y goes through both L and M.

To define the PSEs formally, we use the potential outcomes notation for both the outcome and the mediators. Specifically, we use $Y(a, l, m)$ to denote the potential outcome under treatment status a and mediator values $L = l$ and $M = m$, $M(a, l)$ to denote the potential value of the mediator M under treatment status a and mediator value $L = l$, and $L(a)$ to denote the potential value of the mediator L under treatment status a.[2] This notation allows us to define nested counterfactuals.[3]

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

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

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

For example, $Y\big(1, L(0), M(0, L(0))\big)$ represents the potential outcome in the hypothetical scenario where the subject was treated but the mediators L and M were set to values they would haven taken if the subject had not been treated. Further, if we let $Y(a)$ denote the potential outcome when treatment status is set to a and the mediators L and M take on their "natural" values under treatment status a (i.e., $L(a)$ and $M(a, L(a))$), we have $Y(a) = Y\big(a, L(a), M(a, L(a))\big)$ by definition.[4]

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

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

where $a \neq a^* \in \{0, 1\}$.[5] We define the three terms in equation (1) as the PSEs for the causal paths $A \to Y$, $A \to M \to Y$, and $A \to L \rightsquigarrow Y$, respectively, with a straight arrow denoting a single direct path and a squiggly arrow representing a combination of multiple paths. Specifically, the first term ($\tau_{A \to Y}(a^*)$) corresponds to the amount of treatment effect if the mediators L and M were set to values they would haven taken under treatment status $A = a^*$ for each unit, representing the causal path $A \to Y$. The second term ($\tau_{A \to M \to Y}(a)$) corresponds to the amount of treatment effect

[4]This is sometimes referred to as the "composition" assumption (VanderWeele 2009a).

[5]Note that equation (1) holds both when $(a, a^*) = (1, 0)$ and when $(a, a^*) = (0, 1)$, representing two alternative decompositions. In general, when there is an interaction effect between the treatment and the mediators on the outcome, the PSEs defined by the two decompositions will be different. We focus on the case of $(a, a^*) = (1, 0)$ in the main text and illustrate the alternative decomposition in Part E of the Supporting Information.

operating through the mediator M under treatment status $A = a$ and mediator status $L = L(a^*)$, representing the causal path $A \rightarrow M \rightarrow Y$. Finally, the last term ($\tau_{A \rightarrow L \rightsquigarrow Y}(a)$) corresponds to the amount of treatment effect operating through the mediator L under treatment status $A = a$. It represents the causal path $A \rightarrow L \rightsquigarrow Y$, or the combination of the causal paths $A \rightarrow L \rightarrow Y$ and $A \rightarrow L \rightarrow M \rightarrow Y$.

Thus, in the issue framing example, $\tau_{A \rightarrow Y}(a^*)$ 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; $\tau_{A \rightarrow M \rightarrow Y}(a)$ reflects the effect of issue framing operating only through changing the respondent's perceived importance of competing considerations; and $\tau_{A \rightarrow L \rightsquigarrow Y}(a)$ reflects the effect of issue framing operating through changing the respondent's beliefs about the issue, whether or not the modified beliefs subsequently change the perceived importance of competing considerations.

3.2 Identification Results

Following Pearl (2009), we use a DAG to denote a nonparametric structural equation model with independent errors. In this framework, the top panel of Figure 1 corresponds to a set of nonparametric structural equations that imply our key identification assumption: no confounding exists for any of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships after conditioning on their antecedent variables (see part A of the Supporting Information).

Under these assumptions, it can be shown that the PSEs defined by equation (1) are nonparametrically identified (Avin et al. 2005). In fact, to identify the components of equation (1), it suffices to identify the counterfactual expectation $\mathbb{E}[Y\big(a, L(a^*), M(a^{**}, L(a^*))\big)]$ for any combination of a, a^*, a^{**}. This latter quantity can be expressed as a function of observed variables:

$$\mathbb{E}[Y\big(a, L(a^*), M(a^{**}, L(a^*))\big)] = \int_{x,l,m} \mathbb{E}[Y|x, a, l, m] f(m|x, a^{**}, l) f(l|x, a^*) f(x) dx dl dm, \quad (2)$$

where $f(\cdot)$ denotes a probability density/mass function. This equation generalizes Pearl's (2001) me-

diation 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., $\tau_{A\to L\rightsquigarrow Y}(a)$, reflects the combination of the causal paths $A \to L \to Y$ and $A \to L \to M \to Y$. Without additional assumptions, the PSEs for the paths $A \to L \to Y$ and $A \to L \to M \to Y$ cannot be separately identified. In the issue framing study, for example, we can identify the mediating effect via the respondent's beliefs about the issue ($A \to L \rightsquigarrow Y$), but we cannot pinpoint how much of this mediating effect further operates through the perceived importance of competing issue-related considerations ($A \to L \to M \to Y$).

3.3 Comparison with Existing Approaches

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

$$\begin{aligned}\mathbb{E}[Y(a) - Y(a^*)] &= \underbrace{\mathbb{E}[Y(a, L(a), M(a^*, L(a^*)))] - \mathbb{E}[Y(a^*, L(a^*), M(a^*, L(a^*)))]}_{A\to Y;A\to L\to Y} \\ &\quad + \underbrace{\mathbb{E}[Y(a, L(a), M(a, L(a)))] - \mathbb{E}[Y(a, L(a), M(a^*, L(a^*)))]}_{A\to M\to Y;A\to L\to M\to Y} \\ &\equiv \text{ADE}_M(a^*) + \text{ACME}_M(a), \end{aligned} \tag{3}$$

where $a \neq a^* \in \{0, 1\}$. In this decomposition, $\text{ACME}_M(a)$ reflects the amount of treatment effect operating through M under treatment status $A = a$, whether the effect also operates through L or not. Similarly, $\text{ADE}_M(a^*)$ reflects the amount of treatment effect that does not operate through M, regardless of L.

This decomposition is useful when the researcher's substantive interest lies solely in the mediator M, whereas the other mediator L is purely a nuisance that needs to be accounted for due to the confounding it causes between M and Y. However, a drawback of this approach is that neither the ACME nor the ADE can be nonparametrically identified in the presence of posttreatment confounding of the mediator-outcome relationship. Unfortunately, this is true even if all posttreatment

confounders of the mediator-outcome relationship are observed. Thus, if we think of the mediator L as a posttreatment confounder of the relationship between M and Y, then neither $\text{ACME}_M(a)$ nor $\text{ADE}_M(a^*)$ is identified. Moreover, empirical researchers are often in a situation where both L and M are of substantive interest, making it inappropriate to treat the mediator L as purely a nuisance.

Our proposed approach can be understood in terms of the following alternative decomposition:

$$\begin{aligned}\mathbb{E}[Y(a) - Y(a^*)] &= \underbrace{\mathbb{E}[Y\big(a, L(a^*), M(a, L(a^*))\big) - Y\big(a^*, L(a^*), M(a^*, L(a^*))\big)]}_{A\to Y;\, A\to M\to Y} \\ &\quad + \underbrace{\mathbb{E}[Y\big(a, L(a), M(a, L(a))\big) - Y\big(a, L(a^*), M(a, L(a^*))\big)]}_{A\to L\to Y;\, A\to L\to M\to Y} \\ &\equiv \text{ADE}_L(a^*) + \text{ACME}_L(a), \end{aligned} \tag{4}$$

where the two terms reflect the ADE and ACME with respect to L, rather than M. Equation (4) makes it clear that $\text{ACME}_L(a) = \tau_{A\to L\rightsquigarrow Y}(a)$ and $\text{ADE}_L(a^*) = \tau_{A\to Y}(a^*) + \tau_{A\to M\to Y}(a)$. Thus, the proposed approach allows us to estimate the amount of treatment effect that operates through L (i.e., $\text{ACME}_L(a)$), and, importantly, to further decompose the ADE for L into the effect operating through M but not through L ($\tau_{A\to M\to Y}(a)$) and the effect operating neither through L nor through M ($\tau_{A\to Y}(a^*)$). Table 1 summarizes how the PSEs defined in equation (1) relate to the ACMEs and ADEs for L and M.

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	ACME for M
ADE for L	PSE for $A \to Y$ ($\tau_{A\to Y}(a^*)$)	PSE for $A \to M \to Y$ ($\tau_{A\to M\to Y}(a)$)
ACME for L ($\tau_{A\to L\rightsquigarrow Y}(a)$)	PSE for $A \to L \to Y$	PSE for $A \to L \to M \to Y$

Note: Each of the ADEs and ACMEs is the sum of the two component PSEs belonging to the same column (or row) in the table. For example, $\text{ADE}_L(a^*) = \tau_{A\to Y}(a^*) + \tau_{A\to M\to Y}(a)$ (top row).

From Table 1, we can also see that if the mediators L and M are causally independent, i.e., if the causal path $A \to L \to M \to Y$ does not exist, then the PSEs will be equivalent to the ACMEs, as

$\tau_{A\to L\rightsquigarrow Y}(a) = \text{ACME}_L(a)$ and $\tau_{A\to M\to Y}(a) = \text{ACME}_M(a)$. The prevailing practice of treating different mediators as causally independent can thus be seen as a special case of our approach. Therefore, even in applications where the analyst is willing to assume different mediators to be 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 ACMEs.

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, a task essential to the study of causal mechanisms.

4 Estimating Path-Specific Effects

4.1 An Imputation Approach

To date, most methods for causal mediation analysis have focused on the setting where the researcher is interested in 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. A variety of methods have been proposed to estimate this quantity (e.g., Imai et al. 2010; Tchetgen Tchetgen and Shpitser 2012; VanderWeele 2009b). 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 exactly 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.

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

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

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

A proof of these equations is given in Part B of the Supporting Information. Thus, to evaluate these nested counterfactuals, we need only to estimate (a) the conditional means $\mathbb{E}[Y|X, A = 1, L, M]$ and $\mathbb{E}[Y|X, A = 1, L]$, and (b) their own conditional means given the pretreatment confounders X among the untreated units ($A = 0$). After these estimates are obtained, the outermost expectations in equations (5) and (6) can be estimated using their sample analogs.

Alternatively, the nested counterfactuals above can be written as (see also Part B of the Supporting Information)

$$\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)] = \mathbb{E}\left[\mathbb{E}[Y|X, A=1, L, M]\frac{\mathbb{P}[A=0]}{\mathbb{P}[A=0|X]}\Big|A=0\right] \tag{7}$$

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

These equations suggest that to evaluate the nested counterfactuals, we need only to estimate $\mathbb{E}[Y|X, A=1, L, M]$, $\mathbb{E}[Y|X, A=1, L]$, and the probability ratio $\mathbb{P}[A=0]/\mathbb{P}[A=0|X]$. After these estimates are obtained, the outer expectation in equations (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 $\mathbb{E}[Y\big(1, L(0), M(0, L(0))\big)]$ and $\mathbb{E}[Y\big(1, L(0), M(1, L(0))\big)]$. They can be seen as extensions of Vansteelandt et al.'s (2012) and Albert's (2012) imputation-based estimators for the ACME, respectively, to the estimation of PSEs. 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 the 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 L and M are multivariate and/or continuous. Moreover, the proposed approach only requires the analyst to correctly specify models for the outcome, not any of the mediators. This will likely reduce the possibility of model misspecification, since researchers often have better substantive understanding of the generative process for the outcome variable itself than for the mediators. Below, we provide a step-by-step guide on the implementation of these estimators in experimental and observational studies.

4.2 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

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

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

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

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

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

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

In practice, to reduce model dependence, data-adaptive methods such as Gradient Boosting Machines (GBM) or Bayesian Additive Regression Trees (BART) can be used to fit the outcome models in steps 2 and 3. This can be useful for mitigating bias due to model misspecification, especially when nonlinear

and/or interaction effects are likely to exist (Glynn 2012; Montgomery and Olivella 2018). 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 $\hat{\mathbb{E}}[Y|A=0,X]$ and $\hat{\mathbb{E}}[Y|A=1,X]$ among all units, respectively.

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

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

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

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

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

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

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

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

Again, to reduce model dependence, data-adaptive 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.

4.3 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 L and M 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 $f(m|x, a, l)$ and $f(l|x, a)$ 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.

More recently, 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). Yet, further work is still needed to extend this approach to the estimation of more general PSEs and to settings with more than two mediators.

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

So far, we have assumed that two mediators, L and M, lie on the causal paths from A to Y. The definition, identification, and estimation of PSEs can be generalized to the setting where the effect of treatment 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 average total effect of A on Y can now be decomposed as

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

where $a \neq a^* \in \{0, 1\}$. We assume that the variables $A, M_1, \dots M_K, Y$ follow a DAG that encodes a nonparametric structural equation model with independent errors, such that no unobserved confounding exists for any of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships.

To identify the components of equation (9), it suffices to identify the counterfactual expectation $\mathbb{E}[Y\big(a, \mathcal{M}_k(a^*)\big)]$ for any k and any combination of a, a^*. Similar to the two-mediator setting, this

counterfactual expectation can be expressed as a function of observed variables:

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

$$= \mathbb{E}\big[\mathbb{E}[Y|X, A = a, \mathcal{M}_k]\frac{\mathbb{P}[A = a^*]}{\mathbb{P}[A = a^*|X]}|A = a^*\big]. \tag{11}$$

Equations (10) and (11) correspond to the pure imputation estimator and the imputation-based weighting estimator, respectively, for the PSEs defined in equation (9). The algorithms for implementing these estimators are detailed in Part C of the Supporting Information. In Section 7.2, we illustrate the case of three causally ordered mediators ($K = 3$) by tracing the intergenerational pathways through which exposure to political violence reduces descendants' regime support.

6 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, mediator-mediator, or mediator-outcome relationships. In observational studies where treatment is not randomly assigned, all of these assumptions must be carefully 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-mediator and 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.

Let us consider the general case where the treatment effect operates through K causally ordered mediators $M_1, M_2, \ldots M_K$. Suppose 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

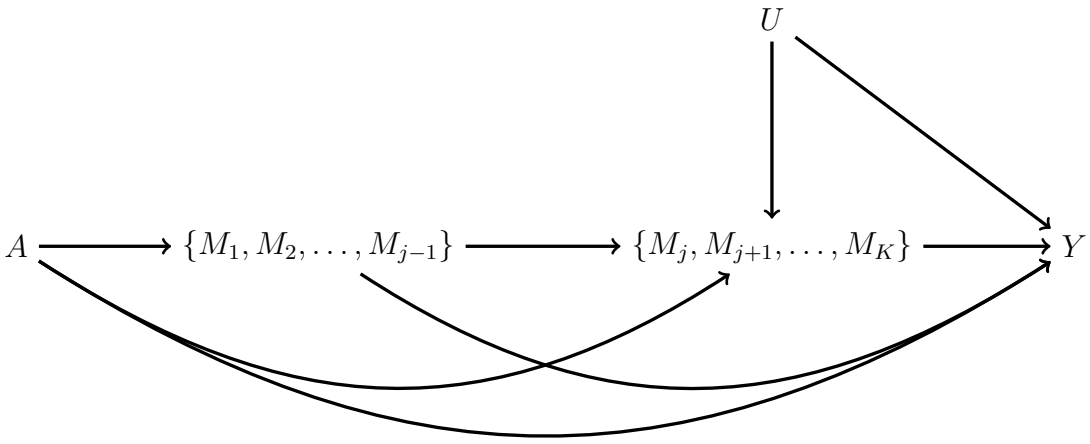

Figure 2: Causal Relationships with K Causally Ordered Mediators where Unobserved Confounding Exists for the Relationship between Mediators $\{M_j, \ldots M_K\}$ and outcome Y.

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

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 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 $M_j, M_{j+1}, \ldots M_K$. 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, 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 difference in the prevalence of U between treated and untreated units as η_k:

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

$$\eta_k = \mathbb{P}[U = 1|X, A = 1, \mathcal{M}_k] - \mathbb{P}[U = 1|X, A = 0, \mathcal{M}_k].$$

It can then be shown that estimates of the direct and path-specific effects without adjusting for U are

subject to the following biases:

$$\text{Bias}[\tau_{A \to Y}(a^*)] = \gamma_K \eta_K; \tag{12}$$

$$\text{Bias}[\tau_{A \to M_j \rightsquigarrow Y}(a)] = -\gamma_j \eta_j; \tag{13}$$

$$\text{Bias}[\tau_{A \to M_k \rightsquigarrow Y}(a)] = \gamma_{k-1}\eta_{k-1} - \gamma_k \eta_k, \quad \text{for any } k > j. \tag{14}$$

A proof of these formulas is provided in Part D of the Supporting Information. These bias formulas allow us to construct a range of bias-adjusted estimates for $\tau_{A \to Y}(a^*)$ and $\tau_{A \to M_k \rightsquigarrow Y}(a)$ across plausible values of $(\gamma_j, \gamma_{j+1}, \ldots \gamma_K)$ and $(\eta_j, \eta_{j+1}, \ldots \eta_K)$. In empirical applications, we may focus on the estimands that are of particular substantive interest. For example, if we are primarily interested in the robustness of the estimated PSE via M_j, i.e, $\hat{\tau}_{A \to M_j \rightsquigarrow Y}(a)$, we can identify the values of γ_j and η_j 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_K and η_K that would suffice to reduce $\hat{\tau}_{A \to Y}(a^*)$ to zero. In the next section, we illustrate this approach with our two empirical examples.

7 Empirical Illustrations

7.1 Issue Framing Effects

We first reanalyze Slothuus's (2008) data to trace the causal pathways through which issue framing affects public opinion. Using a survey experiment on a sample of Danish students, Slothuus 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 importance mediator are causally independent. However, as noted previously, it is likely that respondents' beliefs about welfare recipients 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 latter assumption, the causal 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 L 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 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.[6] We control for a set of pretreatment covariates because, although treatment is randomly assigned, the mediator-mediator and mediator-outcome relationships may still be confounded by 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 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 $A \rightarrow Y$, $A \rightarrow M \rightarrow Y$, and $A \rightarrow L \rightsquigarrow Y$ using the imputation approach described in Section 4.2. 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, $\{L, M\}$ and $\{L\}$).[7] The results are shown in Figure 3. The estimated PSE via the importance mediator ($A \rightarrow M \rightarrow Y$) is 0.19 (95% CI: [0.01, 0.36]), implying that the perceived

[6] Detailed definitions of these variables can be found in Slothuus (2008).

[7] We use the R function `BART::wbart()` (with default settings) to fit the outcome models. Alter-

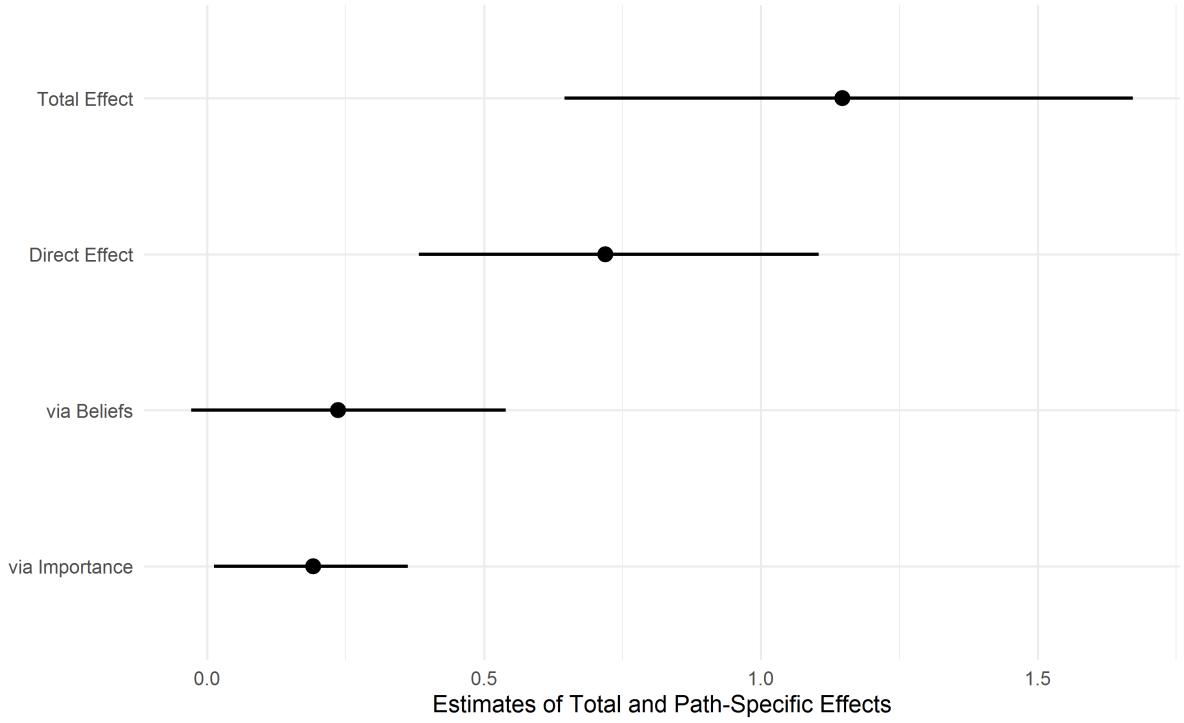

Figure 3: Estimates of Total and Path-Specific Effects of Issue Framing on Policy Support. Note: Error ranges correspond to 95% bootstrapped confidence intervals (1,000 iterations).

importance of competing considerations plays an independent, albeit small, role in transmitting the effect of issue framing on policy support. The estimated PSE via the belief mediator ($A \rightarrow L \rightsquigarrow Y$) is similar in magnitude but not statistically significant. By contrast, over half of the total effect appears to be "direct," i.e., operating neither through the belief mediator nor through the importance mediator.

To examine the robustness of the above conclusion to unobserved confounding, we 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 (L and M), 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 (L and M)

native methods (GLM, GBM) produce similar results (see Part E of the Supporting Information).

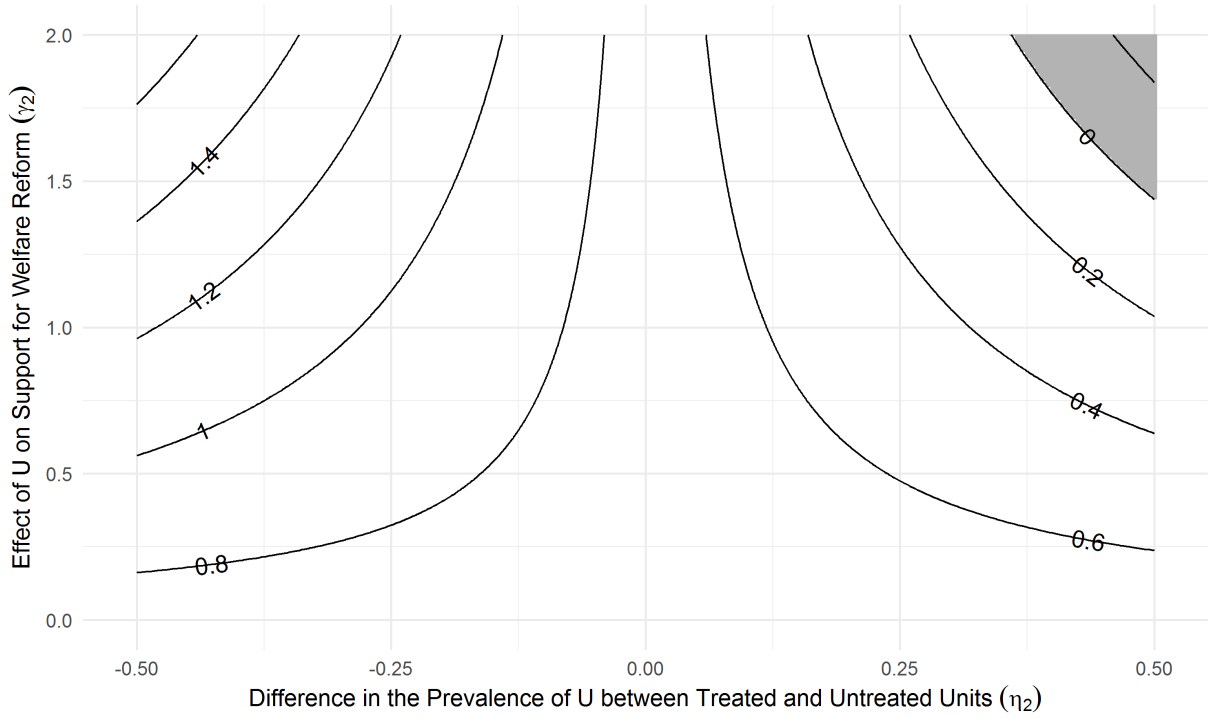

Figure 4: Bias-adjusted Estimates of the Direct Effect of Issue Framing on Policy Support. Note: The grey area shows the range of values of γ_2 and η_2 that would suffice to reverse the sign of the estimated direct effect.

and the baseline covariates (X). To get 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 (L) and perceived importance of competing considerations (M), the conditional association between A and U given L, M, and X can deviate from zero. Specifically, because L and M 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. Thus our finding that most of the framing effect does not operate through the belief mediator or the importance mediator appears to be robust. Figure 4 shows a range of bias-adjusted estimates of the direct effect across plausible values of γ_2 and η_2. 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 large.

7.2 The Legacy of Political Violence

Next, we illustrate the imputation approach for tracing causal paths from observational data. We re-analyze Lupu and Peisakhin's (2017) data to examine the intergenerational pathways through which exposure to political violence shapes descendants' political attitudes. In contrast to the authors' own 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 5. In this DAG, ancestor victimization (i.e., the treatment) denotes whether any family member of the first-generation respondent died during or shortly after the deportation due to poor conditions; the political identities of first-, second-, and third-generation respondents (i.e., the mediators) are measured by the intensity of their attachment to the Crimean Tatars as a social group, their association of that group with victimhood, and their perception of the threat posed by Russia; regime support (i.e., the outcome) denotes whether the third-generation respondent supported Russia's annexation of Crimea; finally, the pretreatment covariates include measures of the first generation respondent's family wealth, religiosity, attitudes toward the Soviet Union, and experience with persecution by state authorities prior to deportation. These covariates are used to control for potential confounding of the treatment-mediator, treatment-outcome, mediator-mediator, and mediator-outcome relationships.

We then estimate the 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 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).[8] The results, as shown in Figure 6, are

[8]We use the R function `BART::pbart()` (with default settings) to fit the outcome models and

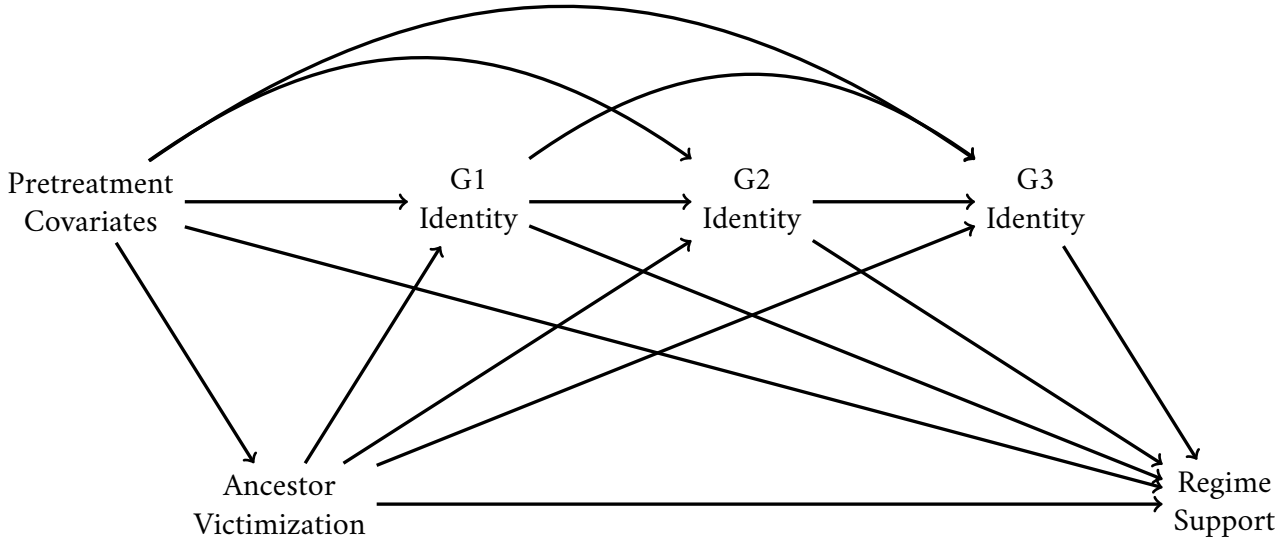

Figure 5: Causal Pathways from Ancestor Victimization to Descendants' Regime Support.

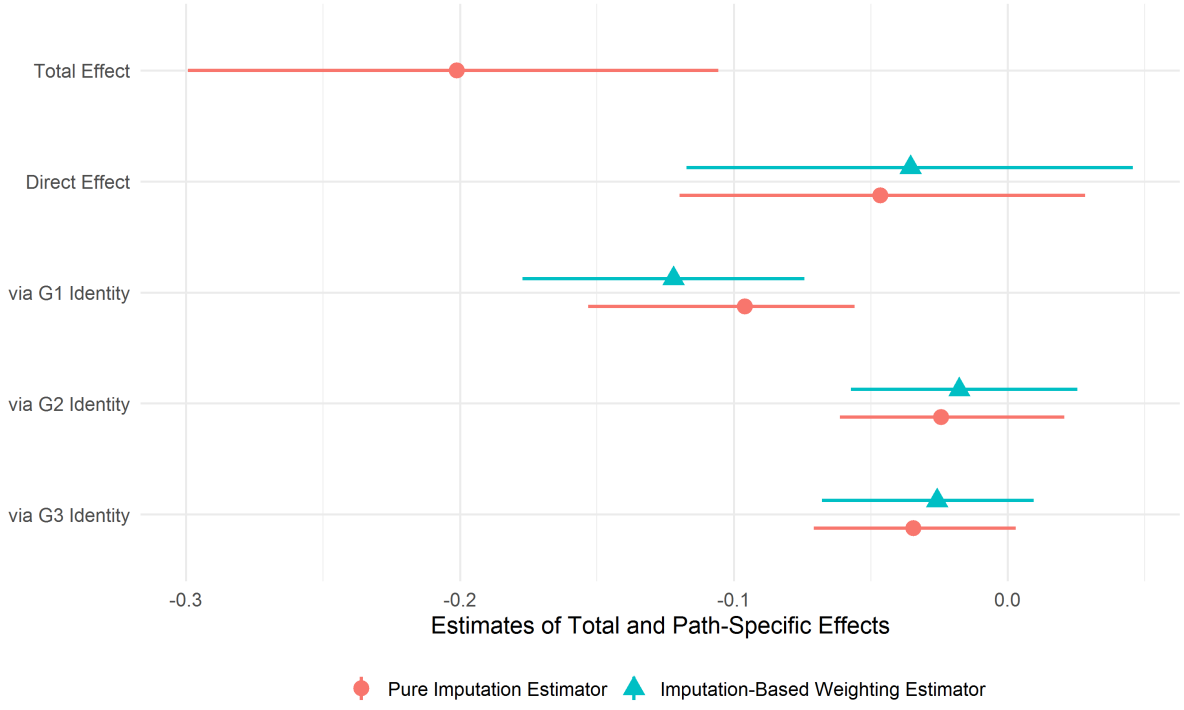

Figure 6: Estimates of Total and Path-Specific Effects of Ancestor Victimization on Support for Russia's Annexation of Crimea.

Note: Error ranges correspond to 95% bootstrapped confidence intervals (1,000 iterations).

the R function `twang::ps()` to fit the propensity score model. Alternative outcome and propensity score models (GLM, GBM) produce similar results (see Part E of the Supporting Information).

Figure 7: Bias-adjusted Estimates of the Path-Specific Effect of Ancestor Victimization on Regime Support via G1 Identity.
Note: The unadjusted estimate comes from the pure imputation estimator. The grey area shows the range of values of γ_1 and η_1 that would suffice to reverse the sign of the estimated path-specific effect via G1 identity.

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 on the probability scale (from 0.64 to 0.44). The "direct effect" is about -0.05 (pure imputation estimator), meaning that most of the total effect operates through the political identities of first-, second-, and third-generation respondents. Moreover, 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 provides strong support to Lupu and Peisakhin's original hypothesis that exposure to political violence affects the identities of first-generation respondents and that *they* transmit these through the family line to shape the political attitudes of their descendants.

To assess the robustness of the above finding to unobserved confounding of the mediator-

outcome relationships, we apply the bias formulas introduced in Section 6 for the PSE via G1 identity ($\tau_{A \to M_1 \rightsquigarrow Y}(a)$). Suppose there exists a binary unobserved confounder U (e.g., presence 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 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, however, can be driven by two opposing forces. On the one hand, if loss of family members during the deportation had also fostered this personality trait among G1 respondents, the association between A and U might be positive. On the other hand, 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. The sign of η_1, therefore, would be the net result of these two opposing forces. If $\eta_1 > 0$, $-\gamma_1\eta_1$ will be positive, suggesting an underestimate of the (negative) PSE via G1 identity. If $\eta_1 < 0$, $-\gamma_1\eta_1$ will be negative, suggesting an overestimate of the (negative) PSE via G1 identity. Figure 7 shows a range of bias-adjusted estimates of the PSE via G1 identity across plausible values of γ_1 and η_1. 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 negative and sizable (e.g., when $\gamma_1 = \eta_1 = -0.32$).

8 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 involves multiple, potentially overlapping causal pathways. In particular, the average causal mediation effect

(ACME) cannot be nonparametrically identified if the mediator-outcome relationship is confounded by other variables that are causally posterior to the treatment, 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 standard ignorability assumptions of causal mediation analysis.

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 different models of the expected outcome can be used to check the robustness of results. In part E of the Supporting Information, we illustrate this point by showing that for our two empirical examples, estimates of the PSEs are consistent 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, mediator-mediator, and mediator-outcome relationships have been observed and adjusted for. Although standard in studies of causal mediation, these assumptions must be scrutinized against the research design and subject matter knowledge in each empirical application. In experimental studies where treatment is randomly assigned, the assumptions of no unobserved treatment-outcome or treatment-mediator confounding are met by design, but the mediator-mediator and mediator-outcome relationships 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 have developed 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 Part F of the Supporting Information, we provide R code illustrating the use of `paths` for our two empirical examples.

References

Acharya, A., M. Blackwell, and M. Sen (2016a). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3), 512–529.

Acharya, A., M. Blackwell, and M. Sen (2016b). The political legacy of american slavery. *The Journal of Politics 78*(3), 621–641.

Albert, J. M. (2012). Mediation analysis for nonlinear models with confounding. *Epidemiology 23*(6), 879.

Albert, J. M. and S. Nelson (2011). Generalized causal mediation analysis. *Biometrics 67*(3), 1028–1038.

Avin, C., I. Shpitser, and J. 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.

Baker, A. (2015). Race, paternalism, and foreign aid: Evidence from us public opinion. *American Political Science Review 109*(1), 93–109.

Balcells, L. (2012). The consequences of victimization on political identities: Evidence from spain. *Politics & Society 40*(3), 311–347.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*(6), 1173.

Brader, T., N. A. Valentino, and E. Suhay (2008). What triggers public opposition to immigration? anxiety, group cues, and immigration threat. *American Journal of Political Science 52*(4), 959–978.

Bullock, J. G., D. P. Green, and S. E. Ha (2010). Yes, but what's the mechanism?(don't expect an easy answer). *Journal of personality and social psychology 98*(4), 550.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics 4*(1), 266–298.

Chong, D. and J. N. Druckman (2007). Framing theory. *Annual Review of Political Science 10*, 103–126.

Cole, S. R. and M. A. Hernán (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology 168*(6), 656–664.

Daniel, R., B. De Stavola, S. Cousens, and S. Vansteelandt (2015). Causal mediation analysis with multiple mediators. *Biometrics 71*(1), 1–14.

Druckman, J. N. and K. R. Nelson (2003). Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science 47*(4), 729–745.

Glynn, A. N. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science 56*(1), 257–269.

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

Imai, K., L. Keele, D. Tingley, and T. Yamamoto (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review 105*(4), 765–789.

Imai, K., L. Keele, T. Yamamoto, et al. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

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

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Lin, S.-H. and T. VanderWeele (2017). Interventional approach for path-specific effects. *Journal of Causal Inference 5*(1).

Lupu, N. and L. Peisakhin (2017). The legacy of political violence across generations. *American Journal of Political Science 61*(4), 836–851.

Mazumder, S. (2018). The persistent effect of us civil rights protests on political attitudes. *American Journal of Political Science 62*(4), 922–935.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods 9*(4), 403.

Miles, C. H., I. Shpitser, P. Kanki, S. Meloni, and E. 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*(520), 1443–1452.

Miles, C. H., I. Shpitser, P. Kanki, S. Meloni, and E. J. Tchetgen Tchetgen (2020). On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika 107*(1), 159–172.

Miller, J. M. (2007). Examining the mediators of agenda setting: A new experimental paradigm reveals the role of emotions. *Political Psychology 28*(6), 689–717.

Montgomery, J. M. and S. Olivella (2018). Tree-based models for political science data. *American Journal of Political Science 62*(3), 729–744.

Nelson, T. E., R. A. Clawson, and Z. M. Oxley (1997). Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review 91*(3), 567–583.

Nelson, T. E. and Z. M. Oxley (1999). Issue framing effects on belief importance and opinion. *The Journal of Politics 61*(4), 1040–1067.

Nelson, T. E., Z. M. Oxley, and R. A. Clawson (1997). Toward a psychology of framing effects. *Political Behavior 19*(3), 221–246.

Nunn, N. and L. Wantchekon (2011). The slave trade and the origins of mistrust in africa. *American Economic Review 101*(7), 3221–52.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Ridgeway, G., D. McCaffrey, A. Morral, L. Burgette, and B. A. Griffin (2017). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation*.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*, 70–81.

Rozenas, A., S. Schutte, and Y. Zhukov (2017). The political legacy of violence: The long-term impact of stalin's repression in ukraine. *The Journal of Politics 79*(4), 1147–1161.

Slothuus, R. (2008). More than weighting cognitive importance: A dual-process model of issue framing effects. *Political Psychology 29*(1), 1–28.

Tchetgen Tchetgen, E. J. and I. Shpitser (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics 40*(3), 1816.

VanderWeele, T. J. (2009a). Concerning the consistency assumption in causal inference. *Epidemiology 20*(6), 880–883.

VanderWeele, T. J. (2009b). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology 20*(1), 18–26.

VanderWeele, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology 21*(4), 540.

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

VanderWeele, T. J., S. Vansteelandt, and J. M. Robins (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology 25*, 300–306.

Vansteelandt, S., M. Bekaert, and T. Lange (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods 1*(1), 131–158.

Zaller, J. R. et al. (1992). *The Nature and Origins of Mass Opinion*. Cambridge university press.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pp. 459–474. New York, NY: Springer.

Zhou, X. and G. T. Wodtke (2019). A regression-with-residuals method for estimating controlled direct effects. *Political Analysis 27*(3), 360–369.

Zhou, X. and G. T. Wodtke (2020). Residual balancing: A method of constructing weights for marginal structural models. *Political Analysis 28*(4), 487–506.

Supporting Information

A Proof of Equation (2)

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

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

where the error terms ϵ_X, ϵ_A, ϵ_L, ϵ_M, and ϵ_Y are jointly independent but otherwise arbitrarily distributed. The potential outcomes described in Section 3.1 can thus be written as

$$
\begin{aligned}
L(a) &= f_L(X, a, \epsilon_L), \\
M(a, l) &= f_M(X, a, l, \epsilon_M), \\
Y(a, l, m) &= f_Y(X, a, l, m, \epsilon_Y).
\end{aligned}
$$

From the above equations and the joint independence of the error terms, we have the following conditional independence relationships: (a) $L(a^*) \perp\!\!\!\perp M(a^{**}, l)|X$; (b) $Y(a, l, m) \perp\!\!\!\perp (L(a^*), M(a^{**}, l))|X$; (c) $L(a) \perp\!\!\!\perp A|X$; (d) $M(a, l) \perp\!\!\!\perp (A, L)|X$; (e) $Y(a, l, m) \perp\!\!\!\perp (A, L, M)|X$.

Thus

$$
\begin{aligned}
&\mathbb{E}[Y(a, L(a^*), M(a^{**}, L(a^*)))|X = x] \\
&= \int \mathbb{E}[Y(a, l, M(a^{**}, L(a^*)))|X = x, L(a^*) = l] f_{L(a^*)|X=x}(l) dl \\
&= \int \mathbb{E}[Y(a, l, m)|X = x, L(a^*) = l, M(a^{**}, l) = m] f_{L(a^*)|X=x}(l) f_{M(a^{**},l)|X=x}(m) dl dm \quad \text{by (a)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x] f_{L(a^*)|X=x}(l) f_{M(a^{**},l)|X=x}(m) dl dm \quad \text{by (b)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x] f_{L(a^*)|X=x,A=a^*}(l) f_{M(a^{**},l)|X=x,A=a^{**},L=l}(m) dl dm \quad \text{by (c) and (d)} \\
&= \int \mathbb{E}[Y(a, l, m)|X = x, A = a, L = l, M = m] f(l|x, a^*) f(m|x, a^{**}, l) dl dm \quad \text{by (e)} \\
&= \int \mathbb{E}[Y|x, a, l, m] f(l|x, a^*) f(m|x, a^{**}, l) dl dm \qquad (15)
\end{aligned}
$$

Integrating 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

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

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

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

Here, the second line uses the fact that $\int \mathbb{E}[Y|x, A = 1, l, m]f(m|x, A = 1, l)dm = \mathbb{E}[Y|x, A = 1, l]$. Integrating the above expression over $f(x)$ yields equation (6). Now, consider equations (7) and (8). By the mediation formula (2), we have

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

Similarly,

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

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

Without loss of generality, let us consider the case of $(a, a^*) = (1, 0)$, i.e., Type I decomposition as defined in Appendix E. The pure imputation estimator proceeds as follows:

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

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

   (a) Fit an outcome model conditional on the treatment A, the mediators $\mathcal{M}_k$, and the pretreatment confounders X. For the 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 $\mathbb{P}[A = 0]$ using its sample analog and $\mathbb{P}[A = 0|X]$ using a propensity score model for the treatment.

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

(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 $\hat{\mathbb{P}}[A = 0]/\hat{\mathbb{P}}[A = 0|X]$.

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

In experimental studies, step (1) can be simplified as $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$ can be estimated using simple averages of the observed outcome within the control and treatment groups. In the meantime, 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.

D Proof of the Bias Formulas in the Presence of a Mediator-Outcome Confounder

Consider the DAG shown in Figure 2, 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 path-specific effects (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

$$\zeta_k(a^*) = \tau_{A \to Y}(a^*) + \sum_{l=k+1}^{K} \tau_{A \to M_l \rightsquigarrow Y}(a);$$
$$\delta_k(a) = \sum_{l=1}^{k} \tau_{A \to M_l \rightsquigarrow Y}(a).$$

Hence the PSEs can be written as

$$\tau_{A \to Y}(a^*) = \zeta_K(a^*);$$
$$\tau_{A \to M_k \rightsquigarrow Y}(a) = \delta_k(a) - \delta_{k-1}(a).$$

Under the three simplifying assumptions outlined in Section 6, 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

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

Thus the bias factors for the PSEs can be written as

$$\text{Bias}[\tau_{A \to Y}(a^*)] = \gamma_K \eta_K; \tag{16}$$

$$\text{Bias}[\tau_{A \to M_k \rightsquigarrow Y}(a)] = \gamma_{k-1}\eta_{k-1} - \gamma_k \eta_k. \tag{17}$$

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

$$\begin{aligned} \eta_k &= \mathbb{P}[U = 1 | X, A = 1, \mathcal{M}_k] - \mathbb{P}[U = 1 | X, A = 0, \mathcal{M}_k] \\ &= \mathbb{P}[U = 1 | X, \mathcal{M}_k] - \mathbb{P}[U = 1 | X, \mathcal{M}_k] \\ &= 0. \end{aligned} \tag{18}$$

It follows from equations (17-18) that

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

E Results from Alternative Decompositions and Models

Equation (1) holds both when $(a, a^*) = (1, 0)$ and when $(a, a^*) = (0, 1)$, representing two alternative decompositions of the average treatment effect. In the R package `paths`, we call them Type I decomposition and Type II decomposition, respectively. Figures E1 and E2 show results for our two empirical examples 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.

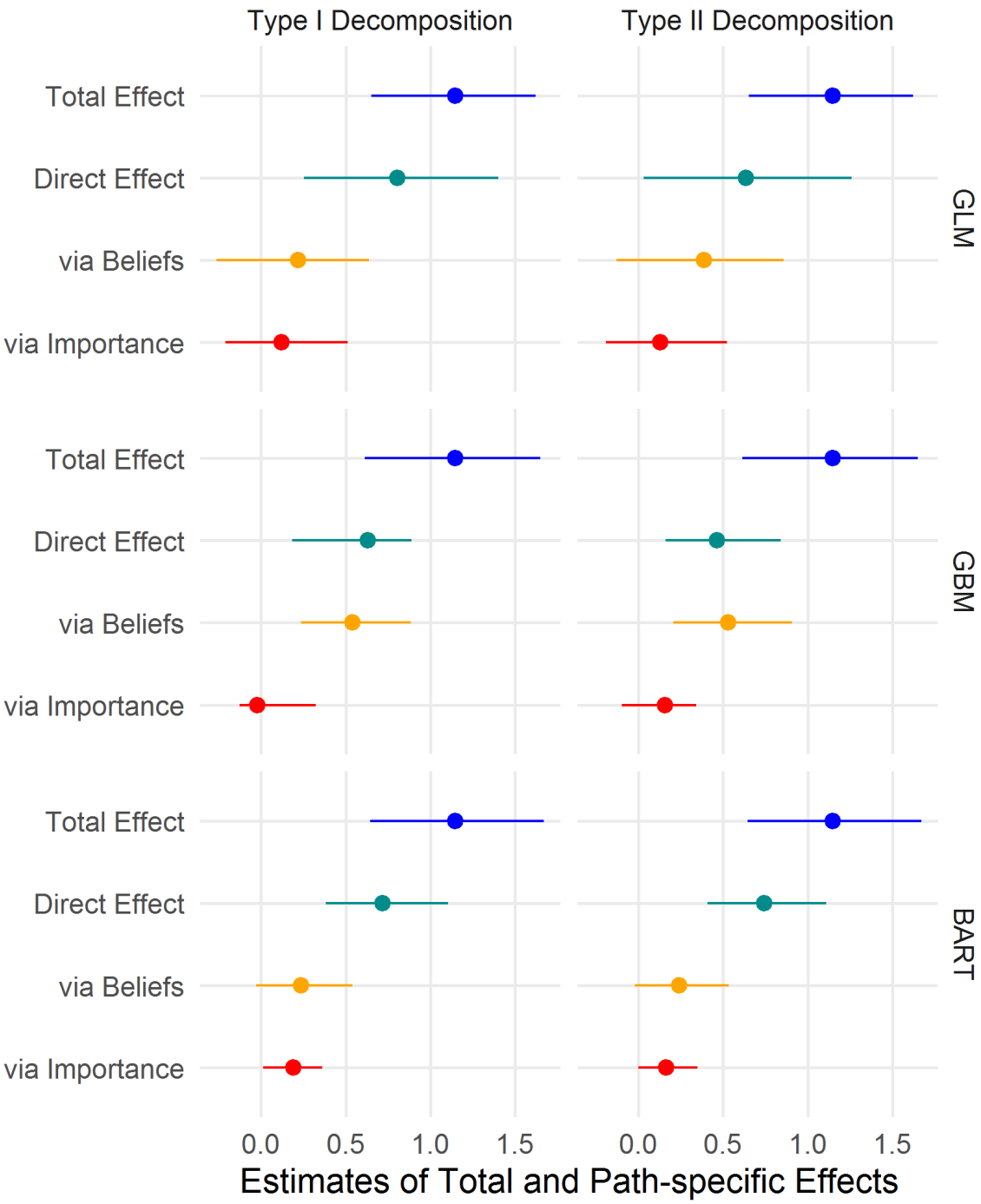

Figure E1: Alternative Estimates of Total and Path-Specific Effects of Issue Framing on Policy Support.

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).

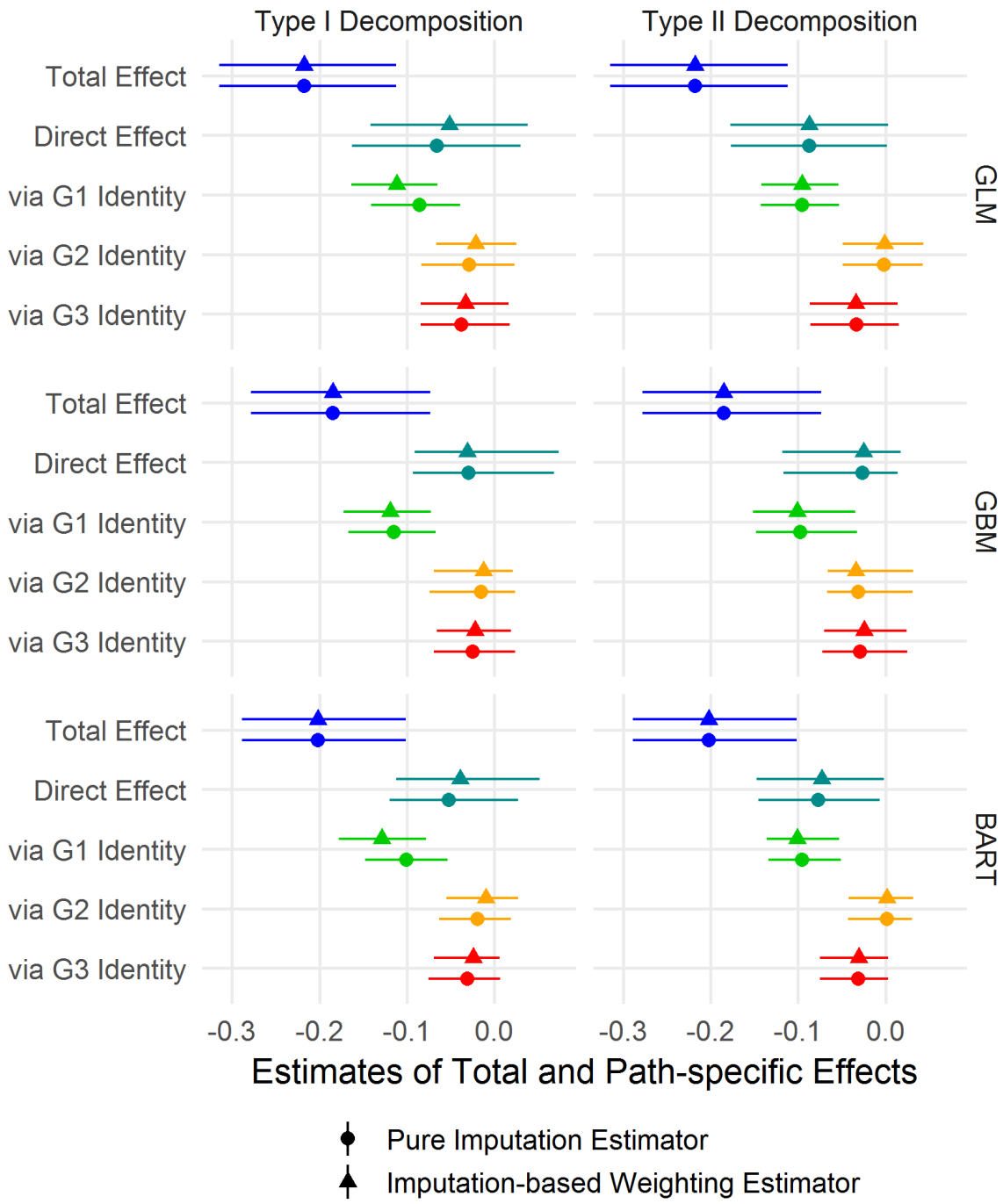

Figure E2: Alternative Estimates of Total and Path-Specific Effects of Ancestor Victimization on Support for Russia's Annexation of Crimea.

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).

F 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(BART)
#########################################################
# Example 1: Issue Framing Effects
#########################################################
welfare <- readRDS("welfare.rds")
# 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)
# baseline model for overall treatment effect
lm_m0 <- lm(Y ~ ttt, data = welfare)
# outcome models with varying sets of mediators
Y <- welfare[[y]]
M1 <- as.matrix(welfare[, c(x, a, m1)])
M2 <- as.matrix(welfare[, c(x, a, m1, m2)])
wbart_m1 <- wbart(x.train = M1, y.train = Y)
wbart_m2 <- wbart(x.train = M2, y.train = Y)
ymodels <- list(lm_m0, wbart_m1, wbart_m2)
```

```r
# causal paths analysis
welfare_paths <- paths(a, y, m, models = ymodels, data = welfare,
                       parallel = "multicore", ncpus = 4, nboot = 1000)
# causal paths plot (Figure 3)
plot(welfare_paths, mediator_names = c("Beliefs", "Importance"))
# sensitivity analysis
welfare_sens <- sens(welfare_paths, confounded = "M1", estimand = "direct",
                     gamma_values = seq(0, 2, 0.005),
                     eta_values = seq(-0.5, 0.5, 0.005))
# sensitivity plot (Figure 4)
plot(welfare_sens, outcome_name = "Support for Welfare Reform")
#########################################################
# Example 2: The Legacy of Political Violence
#########################################################
tatar <- readRDS("internal/tatar.rds")
# variable names
x <- c("kulak", "prosoviet_pre", "religiosity_pre", "land_pre",
       "orchard_pre", "animals_pre", "carriage_pre", "otherprop_pre")
a <- "violence"
y <- "annex"
m1 <- c("trust_g1", "victim_g1", "fear_g1")
m2 <- c("trust_g2", "victim_g2", "fear_g2")
m3 <- c("trust_g3", "victim_g3", "fear_g3")
m <- list(m1, m2, m3)
# design matrices for outcome models
Y <- tatar[[y]]
M0 <- as.matrix(tatar[, c(x, a)])
```

```r
M1 <- as.matrix(tatar[, c(x, a, m1)])
M2 <- as.matrix(tatar[, c(x, a, m1, m2)])
M3 <- as.matrix(tatar[, c(x, a, m1, m2, m3)])
# outcome models with varying sets of mediators
pbart_m0 <- pbart(x.train = M0, y.train = Y)
pbart_m1 <- pbart(x.train = M1, y.train = Y)
pbart_m2 <- pbart(x.train = M2, y.train = Y)
pbart_m3 <- pbart(x.train = M3, y.train = Y)
ymodels <- list(pbart_m0, pbart_m1, pbart_m2, pbart_m3)
# propensity score model for treatment
formula_ps <- violence ~ kulak + prosoviet_pre + religiosity_pre +
  land_pre + orchard_pre + animals_pre + carriage_pre + otherprop_pre
ps_model <- twang::ps(formula_ps, data = tatar, n.trees = 1000,
                      stop.method = "es.mean")
# causal paths analysis
tatar_paths <- paths(a, y, m, ymodels, ps_model = ps_model, data = tatar,
                     parallel = "multicore", ncpus = 4, nboot = 1000)
# causal paths plot (Figure 6)
plot(tatar_paths, mediator_names = paste0("G", 1:3, " Identity"),
     estimator = "both")
# sensitivity analysis
tatar_sens <- sens(tatar_paths, confounded = "M1", estimand = "via M1",
                   gamma_values = - seq(0, 0.5, 0.005),
                   eta_values = seq(-0.5, 0.5, 0.005))
# sensitivity analysis plot (Figure 7)
plot(tatar_sens, outcome_name = "Regime Support")
```