5-2018

Human Rights Treaty Commitment and Compliance: A Machine Learning-based Causal Inference Approach

Dan Sin Nguyen Vo
Purdue University

HUMAN RIGHTS TREATY COMMITMENT AND COMPLIANCE:

A MACHINE LEARNING-BASED CAUSAL INFERENCE APPROACH

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Dan Sinh Nguyen Vo

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2018

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ann Marie Clark, Chair
Department of Political Science
Dr. James A. McCann
Department of Political Science
Dr. Aaron M. Hoffman
Department of Political Science
Dr. Thomas Mustillo
Department of Political Science

Approved by:
Dr. Patricia A. Boling
Head of the Graduate Program

To my family and my best friend.

ACKNOWLEDGMENTS

Writing this dissertation has been a lengthy exercise in fun and frustration, but much more fun than frustration thanks to so many people around me.

Aaron Hoffman is the kind of professor who is more excited about his students' research than they are about their own research. Thomas Mustillo is so encouraging and supportive that whenever you doubt yourself, you should go see him for advice and a good pep talk. James McCann is largely responsible for setting me up on the route of quantitative empirical research. I took his quantitative methods course in my first year at Purdue and it completely changed me in so many ways.

Ann Marie Clark is the best advisor. She is meticulous and thoughtful, have high standards for her students, and is always incredibly supportive. Importantly, she gave me the space to explore my own academic identity and encouraged me to be a free thinker. During my time at Purdue, I also had the chance to attend Elias Bareinboim's two graduate courses on machine learning and Judea Pearl's causality framework. These two courses were eye-opening. When I told people that reading Pearl was like seeing the light, I was not exaggerating at all.

My parents, besides being responsible for my genetic makeup, also taught me to be fiercely independent, instilled in me a strong work ethic, and encouraged me to pursue my own happiness wherever that is. Finally, my best friend, Hao Duy Phan, is the best best friend one can hope for. He has been by my side and supporting me since the days we were two college kids living in a tiny dorm room with a bunch of other college kids.

Much of this dissertation was written during and after the 2016 presidential election campaign in the United States. The results of that election and what followed have caused me so much heartbreak and sadness. Hopefully better days will come.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Amnesty International
CAT	Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment
CDE	Controlled direct effect
CEDAW	Convention on the Elimination of All Forms of Discrimination against Women
CERD	Convention on the Elimination of All Forms of Racial Discrimination
CIRI	Human rights indicators by David Cingranelli and David Richards
DAG	Directed acyclic graph
FDI	Foreign direct investment
GAM	Generalized additive model
GDP	Gross domestic product
HRO	Human rights organizations
ICCPR	International Covenant for Civil and Political Rights
IMF	International Monetary Fund
INGOs	International non-governmental organizations
IPTW	Inverse probability of treatment weighting
MSE	Mean-squared error
NDE	Natural direct effect
NGOs	Non-governmental organizations
NIE	Natural indirect effect
ODA	Official development assistance

OHCHR	Office of the United Nations High Commissioner for Human Rights
OPCAT	Optional Protocol to the Convention against Torture
Polcon	Political constraints index
PTA	Preferential trade agreements
PTS	Political Terror Scale
RMSE	Root-mean-squared error
SCM	Structural causal model
SPT	Subcommittee on Prevention of Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment
TE	Total effect
TMLE	Targeted maximum likelihood estimation
UN	United Nations
WB	World Bank
WPE	Women's political empowerment index
XGBoost	Extreme gradient boosting

ABSTRACT

Author: Nguyen Vo, Dan Sinh. PhD
Institution: Purdue University
Degree Received: May 2018
Title: Human Rights Treaty Commitment and Compliance: A Machine Learning-based Causal Inference Approach
Committee Chair: Ann Marie Clark

Why do states ratify international human rights treaties? How much do human rights treaties influence state behaviors directly and indirectly? Why are some human rights treaty monitoring procedures more effective than others? What are the most predictively and causally important factors that can reduce and prevent state repression and human rights violations? This dissertation provide answers to these keys causal questions in political science research, using a novel approach that combines machine learning and the structural causal model framework.

The four research questions are arranged in a chronological order that reflects the causal process relating to international human rights treaties, going from (a) the causal determinants of treaty ratification to (b) the causal mechanisms of human rights treaties to (c) the causal effects of human rights treaty monitoring procedures to (d) other factors that causally influence human rights violations.

Chapter 1 identifies the research traditions within which this dissertation is located, offers an overview of the methodological advances that enable this research, specifies the research questions, and previews the findings. Chapters 2, 3, 4, and 5 present in chronological order four empirical studies that answer these four research questions. Finally, Chapter 6 summarizes the substantive findings, suggests some other research questions that could be similarly investigated, and recaps the methodological approach and the contributions of the dissertation.

1. INTRODUCTION

Reflecting on her seminal book *Mobilizing for Human Rights: International Law in Domestic Politics* [Simmons, 2009] a few years after it was published, the political scientist Beth Simmons issued a call for more "research on international law and human rights because the claim that the former has had important consequences for the latter is one of the more important claims of this century" [Simmons, 2012, 750]. Her book was also widely regarded as a milestone in the trend toward quantitative empirical research on human rights and human rights law [Hafner-Burton, 2010, Cingranelli, 2010]. This trend, which only dates back to the 1990s [Poe and Tate, 1994, Poe et al., 1999, Keith, 1999] and early 2000s [Hathaway, 2002, Landman, 2005], is a recent addition and complement to the much longer qualitative research tradition [Hafner-Burton and Ron, 2009, Clark and Sikkink, 2013]. As Simmons remembers, when she started working on her book in 2001, "there was almost no quantitative empirical research on human rights practices around the world" [Simmons, 2012, 731].

My dissertation, in a large sense, is a continuation of this research tradition of quantitatively analyzing "the relationship between human rights law and indicators of states' compliance," one that "did not begin in earnest until the turn of the twenty-first century" [Fariss and Dancy, 2017, 274]. Where my research diverges from, and contributes to, the existing literature is that it draws insights and employs methods from two other research areas that gained prominence roughly at the same time as quantitative human rights research but have remained largely separated from the fields of international relations and political science.

The first research area from which I primarily draw on is the causality framework that the computer scientist Judea Pearl first introduced in his monumental book *Causality: Models, Reasoning, and Inference* [Pearl, 2000]. This work was later

updated and expanded [Pearl, 2009a] and was also translated into a more accessible version [Pearl et al., 2016]. Even though scientists have wrestled with the problem of making causal claims from observational data for quite some time, most of the quantitative scholars of international human rights law, like most political scientists and many other empirical scientists [Hernán, 2018], have yet to openly embrace a causal language and a formal causal framework in which to express and conduct their research. For some studies in political science that are more deliberate and open in their efforts to draw causal inference, almost all of them operate within the potential outcomes framework, also known as the Neyman–Rubin model of causal inference [Holland, 1986, Rubin, 2005, Sekhon, 2008]. This causal inference framework is more well known in political science and its allied discipline of economics whereas Judea Pearl's graph-based structural causal model (SCM) framework proves more appealing in other fields such as epidemiology, cognitive science, and computer science. Despite strong opinions on both sides, Pearl has demonstrated that the two frameworks are logically equivalent [Pearl, 2009b, 126–132] although, at least according to proponents of the SCM framework, they are not equally transparent and efficient.

A combination of serendipitous exposure and considered personal preferences has led me to adopt the graph-based SCM framework to tackle the task of making causal inference front and center. The broad rationale of my dissertation is therefore to examine the relationship between international law and human rights as Simmons [2012] and others have called for, but from a new and transparent perspective of causal inference. Specifically, the methods and insights from the SCM framework enable me to revisit and investigate substantive questions about the causal determinants, causal mechanisms, and causal impacts of major United Nations (UN) human rights treaties as well as the causes of state repression and human rights violations. These questions either have not been answered sufficiently from a causal inference perspective or have not even been answered before. In the next four chapters, I seek to answer the following set of questions.

1. What are the most important factors that cause states to ratify three major UN human rights treaties, including the International Covenant on Civil and Political Rights (ICCPR), the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), and the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)?

2. What is the causal effect of each of these three treaties (ICCPR, CEDAW, and CAT) and how much of their causal effect is direct on human rights outcomes and how much is transmitted through intermediate causal mechanisms?

3. What is the causal effect of each of the four treaty monitoring procedures under the CAT and its Optional Protocol (OPCAT), including state reporting, inquiry, individual communication, and country visit?

4. What are the most important factors that predict and cause state repression and human rights violations?

Investigating these questions from a causal inference perspective enables my research findings to have an explicit causal interpretation. They also have a concrete substantive interpretation, including, for instance, quantifying the predicted number of percentage points that human rights protection will increase on average as a result of state membership in a human rights treaty. This is in contrast to, for example, finding a statistically significant relationship between a covariate and an outcome as often found in the literature.

It should be emphasized that employing a causality framework is not just for couching the research findings in a causal language. Some substantive questions such as how much of the causal impact of an international human rights treaty is transmitted through its causal mechanisms can only be formally defined and quantitatively estimated within a causality framework [Pearl, 2001, 2012] even though that same question has been raised for at least more than a decade [Goodman and Jinks, 2004a,b]. In other words, the SCM framework makes it possible to answer some substantive questions that were not even answerable quantitatively before.

My applications of the graph-based SCM framework in this dissertation also point to broader implications for the methodological directions of human rights research and international relations and political science research more generally. In the vast majority of current empirical quantitative human rights research, the traditional methodology is to establish "broad correlations [...] using statistical methods" and then evaluate the causal processes "through case studies" [Simmons, 2012, 734]. This methodological procedure has the effect of artificially separating the causal logic, which is usually expressed and supported qualitatively or sometimes derived through formal models, and the empirical evidence in quantitative, numerical format. As I demonstrate in Chapter 2, researchers may then run the risk of proposing a qualitative causal logic and making implicit causal assumptions that are disconnected from or even contradict the way their statistical models are constructed.

The SCM framework helps bridge this disconnect by transparently expressing the causal logic in the form of a causal directed acyclic graph (DAG). Causal DAGs are highly effective because they compactly and explicitly represent the underlying causal process as understood by the researcher as well as all the causal assumptions he or she makes. The same causal graph then facilitates causal reasoning about the resulting interventional or counterfactual outcomes, links the interventional/counterfactual distribution to the observational distribution (which is often known as causal identification), and informs the statistical models that estimate the causal effect of interest from the observational data. This graph-based causality framework thus provides both a principled method and powerful tools to seamlessly integrate (a) qualitative causal logics or, equivalently, mathematical formal models and (b) quantitative evidence and statistical models that summarize that evidence. The end goal of this methodological approach is to make credible causal claims about social and political reality.

The second research area from which my dissertation research draws from, albeit to a lesser extent, is the field of machine learning. Machine learning has a long,

decorated history [Hastie et al., 2009, James et al., 2013], but it seriously caught the attention of statisticians after the publication of the landmark article "Statistical Modeling: The Two Cultures" [Breiman, 2001a]. In this article, Breiman [2001a] contrasts two distinct approaches to statistical modeling, namely, data modeling (or statistical inference) and algorithmic modeling (or machine learning).

In the data modeling approach, the researcher collects data and assumes a stochastic model for the data-generating process. She then fits the model to the data and estimates the model parameters, some of which are then interpreted as indicating the true relationships between the variables of interest. Key to this data modeling approach is that the researcher assumes she has accurate knowledge about the data-generating process—knowledge that takes the form of correct probability distributions of the data and correct functional forms that link together variables in her statistical models. As an alternative, Breiman [2001a] advocates for a black-box approach that does not pretend to know how the data were generated other than that the sample data are reasonably representative of the target population the researcher is investigating. In this alternative modeling approach, the researcher instead uses learning algorithms to minimize a specified loss function that measures the discrepancy between the predicted values and the observed values of the outcome, that is, to minimize, for example, the classification errors or the sum of squares errors. The goal of this optimizing process is to arrive at a function that most closely approximates the true, unknown data-generating mechanism.

It is worth noting that the data modeling (statistical inference) approach still to this day has been dominant in quantitative political science research, including quantitative human rights research. A major reason could be that, except for a small number of cases [King and Zeng, 2001, Ward et al., 2010, Gleditsch and Ward, 2013, Hill and Jones, 2014, Bell, 2015], most political scientists and international relations scholars seek to understand and find empirical support for their explanatory theories about social and political reality rather than making good predictions. A major shortcoming of the statistical inference approach, however, is that

the assumed knowledge about the data-generating mechanism, especially in the context of complex social and political processes, could easily turn out to be wrong and, as a result, lead to biased inferences and invalid substantive findings.

This major limitation of the statistical inference approach is exactly where the machine learning approach shines because the latter does not depend as much on accurate assumptions about the true data-generating process as does the former. Instead, machine learning tries to imitate and approximate the data-generating mechanism through a trial-and-error learning process. Its goal is also narrower, focusing on making good predictions rather than understanding and explaining the underlying process that generates the data. One of the most important recent innovations of modern machine learning is the ensemble method that combines a large number of similar, comparatively weak prediction models to create an overall much more effective model [Friedman, 2001]. This innovation leads to the powerful prediction technique of extreme gradient boosting [Chen and He, 2015, Chen and Guestrin, 2016] that I apply in Chapter 5. Even more powerful still is an ensemble of different, highly diverse models, each of which is likely able to capture an important aspect of the true data-generating process, to create a hybrid model that performs as well as, and usually better than, even the best individual algorithm. This is the motivating idea and the underlying principle of the Super Learner prediction method [van der Laan et al., 2007] that I use in Chapters 2, 3, and 4.

Effective machine learning algorithms often have superior prediction power, but they tend to have limited interpretability. Furthermore, most do not quantify the uncertainty of their predictions, which can be essential if one wants to do effect estimation and inference. Finally, they are, for the most part, orthogonal to the problem of making causal inference. However, machine learning algorithms can be incorporated into the SCM framework to make functional form-robust causal effect estimation. A combination of machine learning and the SCM framework is the research methodology used in this dissertation to answer substantive research questions about human rights and human rights treaties. Specifically, a straightforward

procedure to employ machine learning for causal inference that I execute throughout this dissertation is to, first, set up a causal DAG to facilitate a translation from the interventional/counterfactual distribution to the observational distribution and identify the causal effect of interest; then employ machine learning algorithms to make predictions about the counterfactual outcome values; and finally, compute the point estimate of the causal effect. To quantify the uncertainty of causal effect estimation, I implement the bootstrap method [Efron and Tibshirani, 1994, Efron and Hastie, 2016, 181–198] throughout the dissertation.

To summarize, my dissertation, entitled *Human Rights Treaty Commitment and Compliance: A Machine Learning-based Causal Inference Approach*, presents a series of innovative applications of machine learning and the SCM framework to answer four sets of questions about human rights and human rights treaties. Each set of questions is motivated by a knowledge gap or an unresolved debate in the substantive literature and built upon what we already know in the research area of international human rights law. The answers to these questions advance our substantive understanding by making new inferences about the causes of human rights violations and the causes, causal mechanisms, and consequences of human rights treaties.

First, **Chapter 2** examines the unresolved question as to which factors cause governments to ratify international human rights treaties. The literature remains divided when it comes to explaining why states commit to human rights treaties even though they are well aware that these laws could potentially restrict their freedom of action. Multiple theories of treaty commitment in the literature can be categorized into three major approaches. The instrumental approach emphasizes the economic rationale for treaty commitment, according to which states ratify international human rights treaties in exchange for material benefits such as increased international investment, more foreign development aid, and membership in preferential trade agreements.

The sociological approach to treaty ratification, in contrast, tends to focus on international socialization and the pressure of normative conformity at both the regional and global levels in explaining why states commit and stay committed to international human rights regimes. Valid and reliable measures of international socialization as well as normative conformity are hard to obtain, but many studies in political science and sociology use the regional and global proportions of treaty members as proxy measurements of these normative factors [Cole, 2005, Goodliffe and Hawkins, 2006, Cole, 2009, Simmons, 2009].

Finally, some of the most popular explanations of human rights treaty commitment can be classified as taking the institutional approach because they often identify domestic institutions as the most salient explanatory variables. According to these theories, regime transitions, democratic institutions, *de facto* existence of multiple political parties, and judicial independence are some of the most commonly identified predictors of human rights treaty ratification.

Based on this ongoing debate in the literature, I recast all of these theoretically predictive variables as causal determinants of treaty ratification. I then estimate their causal effects on human rights treaty ratification. A causal effect in this case is defined as the average change in the probability of treaty ratification across the country–year population if one *intervenes* to alternate the values of the causal variable from its empirically lowest value to its empirically highest value. The estimated causal effect magnitudes are indications of which explanatory variables are truly causally important and thus suggest which theoretical approaches best explain why states commit to human rights treaties. My causal analysis of three major UN human rights treaties finds empirical support for the norms-based theories, deemphasizes the impact of some domestic institutional factors such as regime transitions and judicial independence, and casts doubt on the causal relevance of economic variables such as economic development, official development assistance, and international trade participation.

Second, **Chapter 3** revisits and investigates the question of how international human rights law influences state behavior. Theoretically, this is not an entirely new problem. For more than a decade, human rights scholars have researched and identified several primary causal pathways of three major UN human rights treaties, including the ICCPR, the CEDAW, and the CAT. What is still lacking in the substantive literature, however, is a concrete quantification of how much of the causal impact of a human rights treaty is transmitted through multiple causal pathways. This quantification is not a trivial puzzle. It rather has important implications for preserving or, for that matter, undermining the impact and efficacy of international law on the domestic behaviors of states parties. The absence of any concrete quantification in the empirical literature is not due to a lack of attention. Rather it is because empirical researchers are not familiar with or have not utilized the notation system, vocabulary, and tools of reasoning from the causal inference literature to represent their substantive knowledge and make inferences about causal mediation and causal mechanisms.

My causal analysis in this chapter builds upon the substantive literature that identifies four major causal pathways of human rights treaties, including legislative constraints, domestic judicial enforcement, political mobilization of civil society organizations, and international socialization. I then use causal reasoning tools from the causal inference literature and employ machine learning methods to estimate and decompose the causal impact of three major human rights treaties into the direct causal effect and the indirect causal effect that goes through multiple causal mechanisms.

The causal findings indicate that all three human rights treaties generally help reduce government abuses and increase human rights protection although the magnitudes of their causal effects vary from one treaty to another. Surprisingly, only CEDAW participation directly improves women's political empowerment whereas participating in the ICCPR and the CAT has a *negative* direct impact on human rights practices. However, the good news is that all three treaties exert a positive

indirect influence on state behaviors and, more importantly, their indirect causal impacts that are mediated through multiple mechanisms are disproportionately more substantial in size, ranging from three to 18 times larger than their corresponding direct effects. Taken together, these direct and indirect causal effect estimates provide the first concrete quantification in the literature of the mediated causal effects of human rights treaties.

Third, **Chapter 4** focuses on the causal impact of treaty monitoring procedures under the CAT and the OPCAT. Human rights treaties, like other international institutions, set the standards of behavior for their member states. They also engage in various forms of compliance monitoring. The existing quantitative literature on international human rights law, however, rarely focuses on these monitoring mechanisms and thus tends to overlook the differences in their institutional design and individual causal impact. I therefore unpack the monitoring practices under the UN human rights treaty on torture into multiple monitoring procedures and differentiate their causal effects on state behaviors. My research in this chapter thus provides the first empirical evaluation of the relative causal importance of existing monitoring procedures under a major human rights treaty.

The causal findings show that only the country visit procedure significantly and consistently reduces torture and improves government respect for physical integrity rights. Other monitoring procedures, including state reporting, inquiry, and individual communication, do not. These differing causal effects, I argue, are most likely the result of the variation in intrusiveness among the monitoring procedures. They also indicate that not all monitoring procedures are created equal or have similar causal impacts. Furthermore, the findings suggest that more intensive monitoring and extensive information-gathering by international bodies and panels of independent experts could prove more effective in protecting human rights. More broadly, the research in this chapter offers additional insights into an important topic in international relations regarding the relationship between institutional design and institutional impact. Most current research on this topic is usually conducted with

respect to international institutions in areas such as international trade and global environmental cooperation. My research brings to bear the evidence from an international institution in the area of human rights.

Finally, **Chapter 5** presents a predictive analysis and a causal analysis of the factors that may influence state repression and human rights violations. Both the predictive approach and the causal approach adopted in this chapter diverge from the association-based statistical inference approach often found in the literature. Existing studies mostly identify the covariates that significantly account for the variation in the outcome measures by fitting parametric models of state repression and estimating the regression coefficient of each covariate on the outcome. Regression coefficients that cross a certain threshold of statistical significance are then interpreted as indications of the significance of the corresponding covariates in impacting state repression.

Over time the literature has accumulated a collection of covariates believed to have a significant effect on human rights violations, ranging from demographics to macroeconomic factors, from domestic political institutions to international law, and from international economic variables to the robust presence of the civil society. However, in the absence of additional causal information outside the observed data, the way these covariates are selected—via prior theoretical justification and estimated regression coefficients—does not guarantee that they are causally important in preventing state repression. Nor are they necessarily strongly predictive of human rights violations. The predictive analysis and causal analysis in this chapter reevaluate these variables to identify those that are truly predictive of and causally important to state repression and human rights violations.

For that purpose, this chapter explores and estimates both the predictive power and the causal effect of the same covariates that have been accumulatively identified in the substantive literature. It does that by embedding these covariates in various machine learning prediction models and use the same covariates to construct a causal model for causal effect estimation. The results of both the predictive

analysis and the causal analysis overlap somewhat in underscoring the role of economic development and trade participation in reducing state repression and human rights violations. The causal analysis, however, depicts a more challenging situation for human rights defenders and anyone who wants to prevent and mitigate state repression. It shows that preferential trade agreements or foreign direct investment or domestic democratic institutions on their own have little substantial impact to improve human rights protection. It also highlights, though, the importance of an independent domestic court system as the most consistently impactful factor to improve human rights protection across multiple causal analyses.

In terms of the big picture, this dissertation demonstrates that a combination of the graph-based SCM framework and advanced machine learning methods could help answer important substantive questions that have not been addressed sufficiently or have not even been answered before. More broadly, this combination has a tremendous potential to improve and even transform empirical political science research. It is useful, however, to clarify and reiterate the benefits, tradeoffs, and implications for empirical research of two separate and relatively orthogonal components in this approach: machine learning and the SCM framework.

First, the use of machine learning in this dissertation, in combination with the standard nonparametric bootstrap method for inference, is solely for the purpose of conducting robust estimation. Machine learning is not the only estimation method for an associational or causal analysis, but its significant utility is its robustness and its ability to detect complex, interactive, and non-linear relationships between variables in the observed data. Parametric regression models, on the other hand, may not be able to handle well many non-linear relationships in the data. As a result, the key benefit of using machine learning is to take advantage of a more flexible, nonparametric estimation method that does not depend as much on the assumption of correct model specification.

Without lots of replication studies, it is hard to tell how much of a difference this flexible machine learning-based estimation method would make to the substantive

findings in the human rights literature. However, it is not at all inconceivable that many previous findings in the literature may have to be revised if their validity depends on some functional form assumptions that turn out to be incorrect. More importantly, since machine learning methods are more robust and less assumption-dependent, that is, they allow us to make inferences from observational data even when we do not know what the correct data model is, it is only reasonable and even recommended that researchers should consider adopting a machine learning-based approach as their default estimation method. If, for some reasons, the researchers have justifiable, concrete knowledge about the correct functional forms, they then can use that knowledge to construct parametric regression models to estimate the treatment effects in a simpler and possibly more efficient way.

The flexibility of machine learning methods can prove even more beneficial going forward when more powerful, more flexible, and less computationally expensive machine learning techniques are developed. Still more advanced methods are being developed to apply machine learning techniques for the purpose of robust effect estimation and making statistical and causal inference [Chernozhukov et al., 2017]. Coupled with the likelihood that political scientists will examine more variables in their research, use bigger data, and investigate more complex relationships among their variables of interest, machine learning should and will likely be adopted more widely, if for no other reasons than to be able to discover more complex patterns in high-dimensional data.

This machine learning-based estimation approach, however, does have certain tradeoffs. The benefits of this machine learning-based approach should nonetheless be put in the proper context of serving the sole purpose of flexible and robust effect estimation. It is obviously not a substitute for substantive knowledge and for understanding the research problem. It is also orthogonal to the task of endowing any effect estimates from observational data with a causal interpretation. That task is accomplished using the SCM framework, which is the other component in my methodological approach.

Second, perhaps the most important part in my approach is the graph-based SCM framework that I adopt to conduct causal analysis. Given that much empirical research, including quantitative human rights research, can be reasonably classified as aiming to make causal claims from observational data, researchers should be "explicit about the causal objective" of their studies so as to "reduce ambiguity in the scientific question, errors in the data analysis, and excesses in the interpretation of the results" [Hernán, 2018, e1].

More importantly, researchers should openly embrace a coherent framework to articulate and reason about cause and effect. The graph-based SCM framework, combining the graphical language and counterfactual language, precisely fills this need. It gives researchers the mathematical notations to concisely represent their causal queries. It offers the vocabulary to define the causal questions and the causal quantities of interest in the interventional and/or counterfactual language rather than in terms of probabilistic distribution and a statistical relationship between an independent variable and an outcome. Crucially, it provides the graphical tool of causal DAGs to efficiently represent our background knowledge and prior causal information. Finally, it supplies the tools of inference that help researchers to determine whether and under what conditions they can establish the estimability of the causal effect of interest from observational data. These are the key benefits of this causality framework that traditional approaches and more familiar methods either overlook or fail to provide.

Of particular significance, representing the background knowledge is critical in any causal analysis because no causal questions can be answered without prior assumptions about the causal structure that generates the data. This is most apparent in the practice of selecting "control variables" for inclusion in a regression model to adjust for potential confounding. However, only confounding variables (also known as confounders, which *cause* both the independent variable and the outcome) should be "controlled for." If a covariate is a mediator (an intermediate variable that both causally follows the independent variable and has a causal influ-

ence on the outcome) or a collider (a variable that is directly or indirectly caused by both the independent variable and the outcome), its inclusion in a regression model would introduce bias in the effect estimate. The tricky thing is that all three types of covariates—confounders, mediators, and colliders—are correlated with the outcome and the independent variable. As a result, the conventional practice of including any variables that are correlated with both the outcome and the independent variable of interest [Hill and Jones, 2014, footnote 2] is actually a bad practice. Instead, confounding adjustment has to rely on sufficient background knowledge that is not available from the observed data. Only the subject matter knowledge of the causal structure (that is, whether and how each variable in the causal model is causally related to every other variable) can form the basis on which to determine whether a covariate is a confounder or a mediator or a collider on which specific causal pathways and, as a result, which variable should be "controlled for" and which variable should not be adjusted for. The key utility of a causal DAG is to make explicit and transparent in a graphical form the subject matter knowledge that leads to this assumed causal structure .

When researchers employ the graph-based SCM framework they become highly aware of and sensitive to the fact that if the background knowledge is tenuous, permitting different hypothetical causal structures, then different and even contradictory findings will ensue. It is very likely that the current literature on human rights and human rights treaties is in this kind of situation where divergent and contradictory findings abound partly because different researchers, and even a single researcher over different research projects, would implicitly assume different underlying causal structures. As a result, another implication of the graph-based SCM framework for any substantive debates in the literature is that it would focus the attention of the scientific community more on the subject matter and rightly so than on specific statistical debates and estimation techniques.

Using the graph-based SCM framework could help move scientific research forward by highlighting, clarifying, and contributing to reconciling different assump-

tions about the true causal structure. In addition to providing the same causal vocabulary and language, it requires researchers to explicitly represent their causal assumptions in a graphical form and makes any differences in causal assumptions easily recognizable. Researchers can then communicate much more easily about their different assumptions and have a more productive debate about the underlying causal structure, which contributes to the overall scientific progress. At a minimum, this graph-based SCM framework will focus the attention of the researchers on the underlying causal process and, at least in the context of social scientific research using observational data, it underscores the tentative nature of individual research findings and the cumulative and collective nature of scientific research.

It should be noted that while the graph-based SCM framework is not the only causality framework out there, it is a very efficient and intuitive framework. It also enables researchers to be more rigorous in executing the task of making causal inference. This is because the key methods and methodological findings from this framework, including, for example, the backdoor criterion for causal identification and the mediation formula for estimating natural direct causal effect and natural indirect causal effect, have been proven sound and complete [Pearl, 2014a, 2017]. These methodological results are sound in the sense that if we apply them and have the necessary causal assumptions, we are guaranteed to get valid causal inference. They are complete in the sense that if these methods require certain conditions or assumptions to make valid causal inference, there are no other methods in any framework that can do better without additional information or assumptions. As a result, while there is no guarantee that every causal inference task can be completed using this framework or that there is no other competing framework that can accomplish the same task, the graph-based SCM framework is perhaps the most efficient framework at this moment.

All of these benefits notwithstanding, there are certainly some barriers. It takes some cognitive flexibility, methodological pluralism, different allocation of research and training resources by individual researchers and the scientific community as a

whole, and a considerable amount of time for a novel methodological framework to be embraced and widely adopted. Different academic disciplines will likely proceed at different paces in terms of adopting this graph-based causality framework with the field of epidemiology probably leading the pack [Leeder, 2016], but hopefully other fields will be able to catch up quickly so as to facilitate more scientific progress.

Finally, there are two methodological and substantive issues that this dissertation has not actually dealt with to any significant extent, including measurements and missing data. The issue of measurements is particularly vexing in human rights research. Essentially, the questions that every quantitative human rights researcher has to keep in mind are whether the measurements of human rights practices such as specific human rights scores could capture the underlying theoretical constructs and whether there are any significant measurement errors. These issues are even more challenging because most human rights scores are actually complex indices that aggregate many sources of raw information. Recently, international relations scholars have started to focus more attention on the issues of measurements in quantitative human rights data [Clark and Sikkink, 2013, Fariss, 2014, Fariss and Dancy, 2017], but these issues remain important challenges.

In this dissertation, I also implicitly assume that all missing data are either missing completely at random or missing at random. I therefore use a standard multiple imputation procedure to deal with the issue of missing data. To be fair, this is in line with most other research not just in political science but in other disciplines as well. In fact, as Little and Rubin [2002, 22] observe, "[e]ssentially all the literature on multivariate incomplete data assumes that the data are MAR [missing at random], and much of it also assumes that the data are MCAR [missing completely at random]." Interrogating these assumptions about missing data, investigating the implications of these assumptions and their validity for data analyses, and incorporating plausible causal information about the missingness mechanisms into the causal analyses are quite beyond the scope of this dissertation. It does not mean

that they are not important. More recent advances in the methodology of dealing with the issue of missing data from a causal inference perspective [Daniel et al., 2012, Thoemmes and Mohan, 2015, Mohan and Pearl, 2018] are worth the efforts to examine and apply in future research.

2. CAUSAL INFERENCE USING MACHINE LEARNING: AN APPLICATION TO HUMAN RIGHTS TREATY RATIFICATION

2.1 Theories of Treaty Ratification

International human rights law is created to protect and promote universal human rights. It does that by establishing substantive obligations for states parties and designing procedural mechanisms to monitor the implementation of those obligations [De Schutter, 2010, Alfredsson et al., 2009, Buergenthal, 2006]. A major global regime is the UN human rights treaty system, which includes many treaties and their associated monitoring bodies [Keller and Ulfstein, 2012, Rodley, 2013]. A natural question arises in the literature as to why more and more countries have ratified and remained committed to human rights treaties that are designed precisely to limit their freedom in how to treat their own citizens. Figure 2.1 shows the increasing number of states parties to three major human rights treaties from 1966 when the ICCPR was opened for ratification until 2013.

The question of treaty ratification is a simple, yet vexing, puzzle that scholars have wrestled with for a long time. Many theories have been proposed, identifying various explanatory variables, but any consensus and agreements remain elusive. First, some scholars believe that international socialization and the pressure of normative conformity make cause state leaders to realize that treaty ratification is the expected and appropriate thing to do [Finnemore and Sikkink, 1998]. Two studies by Goodliffe and Hawkins [2006] and Hathaway [2007] find correlative evidence to support this argument when they use global and regional ratification rates as proxies for international socialization. A prominent study that follows, however, casts doubt on the role of socialization as the driving force behind treaty ratification. Simmons [2009, 90–96] creates a series of variables (measuring regional

normative convergence, socialization opportunities, an index for two different time periods, and information environments) that interact with density of regional ratification and argues that regional ratification rates do not necessarily reflect a normative force as much as a strategic calculation. It is not immediately clear what causal models that Simmons [2009] assumes would generate the data and whether and how the effect estimates of those interactive variables could be causally interpreted.

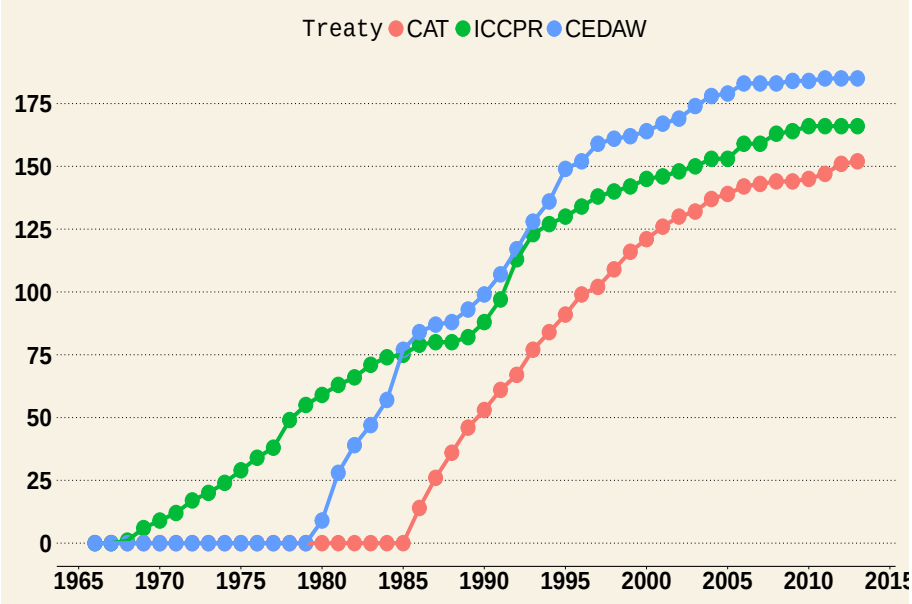


Fig. 2.1.: Numbers of states parties to the ICCPR, the CEDAW, and the CAT from 1966 to 2013. The three treaties were opened for ratification in 1966, 1979, and 1984, respectively.

The second group of explanations focuses on the economic reasons that states voluntarily commit to universal human rights standards and subject themselves to international monitoring. According to these explanations, states use ratification as a signaling device to improve their social standing, expecting to gain material benefits in return, even if they are disingenuous about treaty compliance. The need for social signaling could be significant given the pressures on lending institutions, foreign investors, and developed countries to link foreign aid [Lebovic and Voeten, 2006, Spence, 2014], international investment [Blanton and Blanton,

2007], and preferential trade agreements [Hafner-Burton, 2005] to human rights issues in recipient countries. Participation in international trade in particular has been shown to be a significant predictor of treaty commitment [Lupu, 2014]. The transactional rationale of treaty ratification could be even more pressing for transitional and newly independent countries since they often need external economic assistance and financial support [Smith-Cannoy, 2012, 64–91]. This instrumental argument, however, turns out to have virtually no empirical support according to a critical study by Nielsen and Simmons [2015]. The two authors find no correlation between ratifications of four major human rights regimes (under the ICCPR and the CAT) and either the amounts of foreign aid from OECD countries or other measures of tangible and intangible benefits.

Third, the most popular explanations of treaty ratification often identify domestic institutions as the key predictors. An early theory advances what is often referred to as the "lock in" argument, according to which transitional countries or those facing potential democratic instability tend to join human rights regimes to lock in and consolidate their democratic institutions [Moravcsik, 2000]. Although this argument finds some empirical support in another study [Neumayer, 2007], there are some dissenting findings as well, indicating that neither new democracies nor unstable, volatile regimes are significant predictors of CAT ratification [Goodliffe and Hawkins, 2006].

Researchers also focus on the interaction of domestic institutions and human rights practices to explain ratifications [Hathaway, 2007]. Post-ratification, they argue, states that have sub-standard human rights protection will likely incur a higher cost of policy adjustment. This cost, in turn, is more likely to actually materialize if democratic institutions are in place to constrain state leaders. As a result, a poor human rights record predicts a low probability of ratification, but only among democracies. Ratification cost may rise as well, depending on the types of domestic institutions, including constitutional ratification rules, political regimes, and an independent court system [Simmons, 2009, 67–77]. Hill [2016a] applies the

same logic to explain how governments selectively make reservations when they ratify human rights treaties based on their domestic standards and legal institutions. Conversely, autocracies are just as likely to ratify human rights treaties since their ratifications are usually empty promises that do not bring any real cost of behavioral change [von Stein, 2016]. The theoretical expectation is that, among autocracies, prior human rights practices have little impact on the probability of treaty ratification.

Generally, it should be noted, states are believed to be less likely to commit to international treaties if their prior level of compliance is low. This is often known as the selection effect argument [Downs et al., 1996, von Stein, 2005, Simmons and Hopkins, 2005]. In the literature on international human rights law, however, this selection effect is often treated as source of potential bias where prior measures of human rights outcome may confound the causal relationship between human rights treaties and contemporaneous measure of the outcome. The causal impact of prior human rights practices on treaty ratification is rarely a quantity of interest to investigate.

For the most part, democracies are also believed to be more likely than autocracies to ratify human rights treaties [Landman, 2005] because of their domestic pressures or an incentive to export rights-respecting norms. Hafner-Burton et al. [2015a] similarly argue that autocracies are less likely to join human rights regimes that may expose them to a high cost of compliance. Vreeland [2008] adds an important caveat, however. He agrees that because dictators are more inclined to use torture to retain power, they are indeed less likely to ratify the CAT so as to avoid the cost associated with treaty violations. Yet, for dictators that co-exist with multiple political parties, they have to bear the cost of non-ratification in the form of pressures from the opposition parties. It turns out, according to Vreeland [2008], dictatorships with multiple parties are actually *more* likely to ratify the treaty.

Hollyer and Rosendorff [2011] concur with Vreeland [2008], but they differ with respect to his reasoning. For repressive leaders, the two authors claim, rat-

ifying the CAT can actually bring some significant signaling benefits with respect to a particular audience: the domestic opposition. Opposition groups perceive an authoritarian leader's act of committing to the CAT (and then flaunting treaty violations) as a credible signal of her strength. As a result, the opposition is less likely to mount a challenge, in effect prolonging the survival of the authoritarian leader. The implication is that autocracies are *more* likely to ratify costly human rights treaties not because they concede to pressures from the opposition parties as Vreeland [2008] argues, but rather because they actively seek ratification to reap its domestic signaling benefits. For many human rights scholars, this credible commitment argument to explain treaty ratification among autocratic regimes "has some plausibility problems on its face" [Simmons, 2012, 743], but it has not been disputed empirically. Even Hollyer and Rosendorff [2011] have conducted no causal tests, pointing instead to the statistical association between CAT ratification and several different outcomes such as leadership survival, level of government repression, and the extent of opposition efforts.

To summarize, exactly why states ratify human rights treaties is still unclear. There could be many reasons and multiple theories, but findings are all over the map and often contradict each other or go untested from a causal inference perspective. Whether they are ideational, instrumental, or institutional, theories of treaty ratification remain contested and the issue of treaty ratification "has not yet been fully explored" [Hafner-Burton, 2012, 271]. As Simmons [2012, 737–744] similarly observes, the question of why states ratify international human rights law remains "an enduring puzzle." This unresolved puzzle is both the substantive premise and motivation to develop a different test of major theories of treaty ratification that is based on an explicit causal inference perspective.

For that purpose, in this chapter I take a novel approach to testing theories of treaty ratification and addressing the question of why countries ratify human rights treaties. The basic idea underlying my test strategy is that, since different theoretical approaches propose different explanatory variables, one can adjudicate these

theories by directly estimating the causal effects of these variables and compare their causal effect magnitudes as a direct measure of their causal importance for treaty ratification. Theories that propose more causally important variables will be not only more empirically supported but also more substantively relevant.

In terms of implementation, this chapter builds upon the substantive knowledge in the literature to set up causal models of treaty ratification. These causal models would enable the identification of the causal effects of various factors that have been theoretically hypothesized to cause states to ratify human rights treaties, including the International Covenant on Civil and Political Rights (ICCPR), the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), and the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT). Once causal identification is established, I then apply two machine learning-based estimation methods, the targeted maximum likelihood estimator and the substitution estimator, to actually compute the causal effects from observational data. The application of machine learning is aimed to relieve us of our dependence on the assumption of a correct functional form specification for unbiased effect estimation and make our estimates more robust.

While this chapter does not propose a new theory as to why states ratify human rights treaties, it nonetheless makes both substantive and methodological contributions to the human rights literature. First, it subjects multiple theories of treaty ratification to a different kind of empirical testing that does not rely on detecting a statistically significant relationship between treaty ratification and other covariates. Rather, the strength and substantive relevance of a theory will be based on how much the explanatory variables that the theory proposes actually cause states to ratify human rights treaties. My analysis thus provides insights into the causal determinants of treaty ratification by identifying the variables that are most causally relevant to be intervened upon if the goal is to promote universal ratification of human rights treaties.

Second, methodologically my causal analysis constructs causal models that are more transparent in their causal assumptions and uses machine learning-based estimation methods that are less dependent on correct functional forms assumption. These two features of identification transparency and estimation robustness are missing in many current empirical inquiries. Previous research has analyzed predictors of state commitment to universal treaties [Lupu, 2014]. Others have applied the machine learning technique of random forest to examine the predictive associations between various covariates and state repression [Hill and Jones, 2014]. My investigation improves upon the former by using machine learning in lieu of parametric linear regression models and upon the latter by endowing the findings with a causal interpretation.

Fundamentally, my causal analysis follows Judea Pearl's philosophy of "define first, identify second, estimate last" [van der Laan and Rose, 2011]. I start by examining the literature, describing the research gaps, and formulating a new form of direct theory-testing from a causal inference perspective. I then employ Pearl's causal inference method [Pearl, 2009a] to identify the causal effects of interest and use an ensemble machine learning technique called Super Learner [Polley and van der Laan, 2010] to produce more robust effect estimates. Finally, I interpret the causal findings in the substantive context of adjudicating competing theories of treaty ratification.

2.2 Empirics of Treaty Ratification

My causal analysis offers a solution to the puzzle of treaty ratification by evaluating and comparing the causal effects of many theoretically identified predictors of treaty ratification across three major human rights treaties. The novelty of this test strategy is to apply a machine learning-based causal inference approach to address two major limitations in existing empirical inquiries. First, existing studies almost always use parametric regression models that rely on the statistical significance of

ratification predictors. These models have to make restrictive assumptions such as linearity, normality, and additivity in order to characterize the shape of the relationships between treaty ratification and its predictors. Usually no justifications are provided as to why a linear functional form, for example, or additivity of covariate effects is appropriate or accurate instead of exponential, U-shaped, higher-order, threshold effects or any of an infinite number of other forms. Since we usually do not know *a priori* the underlying data-generating process and it is often virtually impossible to know the correct functional form when it comes to modeling complex political phenomena, a conveniently specified statistical model is likely a misspecified one, which will then produce unreliable and biased effect estimates.

By using flexible machine learning methods, we are essentially relieving ourselves of the burden of having to correctly specifying our parametric models. In other words, machine learning helps make up for the lack of accurate prior information about the functional form of the data-generating process. The trade-off, however, is that machine learning methods often add a certain amount of complexity to our estimation while also accruing higher computational costs. Depending on the background knowledge and the specificities of a research analysis, different trade-offs can be made. If it is reasonable to assume a linear regression model happens to accurately reflect the underlying data-generating process, it is probably more efficient to use a parametric model. That kind of assumption, however, is typically untenable outside randomized controlled trials and especially in the context of high-dimensional joint probability distribution, that is, when we have a large number of covariates. The more covariates we have, the more likely there will be complex interactions and relationships among them and the less likely our parametric models can capture these relationships.

The second limitation is that virtually every study implies a causal query about the determinants of treaty ratification. Yet, none has openly embraced a causal language and framework within which to formulate the causal quantities of interest that correspond to the research questions and link these quantities to the observa-

tional data that are sampled from an observational population. This is the essence of what is often called causal identification. Causal identification is difficult because of one key problem, which is that the same probability distribution, from which our observational data are sampled from, can be generated by different underlying causal processes [Peters et al., 2017, 10] or, one may say, different causal stories. The task of identification is therefore completely separate from and prior to the task of estimation. Estimation is computing the numerical values of our quantities of interest from the observational data. Identification is establishing that there is a unique, one-to-one mapping between the observational data and the underlying causal story in the form of causal assumptions. Estimation thus provides an answer to our causal question. Identification determines whether our causal question is even answerable in the first place.

It should be noted that one uses different methods to complete different tasks. We use statistical methods for estimation and "*extra*-statistical methods [...] to express and interpret causal assumptions" [Pearl et al., 2016, 5]. The causal framework that Pearl [2009a] and others have developed provide these "*extra*-statistical methods". Not using these causal framework and methods leads to unfortunate implications for empirical research. For example, endogeneity, an identification issue, is often viewed as a statistical problem because "there is no agreement on the most appropriate statistical approach" [von Stein, 2016, 661]. However, the reason there is no agreement on the most appropriate statistical approach is because "there is no statistical method that can determine the causal story from the data alone" [Pearl et al., 2016, 5]. Without clearly separating and distinguishing between the task of identification and that of estimation, researchers often mistake estimation techniques such as propensity score matching for an identification strategy [Pearl, 2009a, 349]. Similarly, they fail to employ highly useful causal identification tools such as the backdoor criterion [Pearl et al., 2016, 61–64], a simple and intuitive test to see if our causal story is sufficient to allow a computation of causal effects from observational data, to subsequently guide their covariate selection and inform their

statistical modeling. Instead, researchers resort to statistical "fixes" such as country fixed effects and time trends that could prove arbitrary or even counterproductive [Chaudoin et al., 2016] and, in any case, skirt around the crux of the problem, which is to explicitly link the causal story to the observed data.

The following two examples underscore the benefits of embracing a transparent causal inference framework. In a prominent study of treaty commitment, the researcher fits multiple regression models and successively regresses ratifications of human rights treaties and optional protocols and provisions on several predictors that are measured contemporaneously, including democracy, human rights violations, and their interaction term. The regression coefficient for democracy is then interpreted as indication that "for each point increase in the measure of Democracy, states with no human rights violations have between 10 and 54 percent increased chance of ratifying human rights treaties than nondemocratic ones" [Hathaway, 2007, 609].

This modeling procedure and interpretation are appropriate for a causal model represented in Figure 2.2a where X denotes democracy, Y stands for human rights violations, and A is ratification. The majority of the literature, however, suggests that it is at least as likely that democracy contemporaneously influences the extent of human rights violations rather than the other way around even if it is possible that state repression may impede democratization or undermine democracy in the next time period. A different causal model in Figure 2.2b could be deemed just as, if not more, plausible, in which conditioning on human rights violations Y would induce a post-treatment bias in estimating the causal effect of democracy X on ratification A. The broader point is that whether the causal effect of interest can be identified and estimated without bias depends intimately on the topology of the causal model and it is unnecessarily difficult, if not impossible, to fairly evaluate the causal model's substantive plausibility in the absence of an explicit, preferably graphical, representation of the causal model.

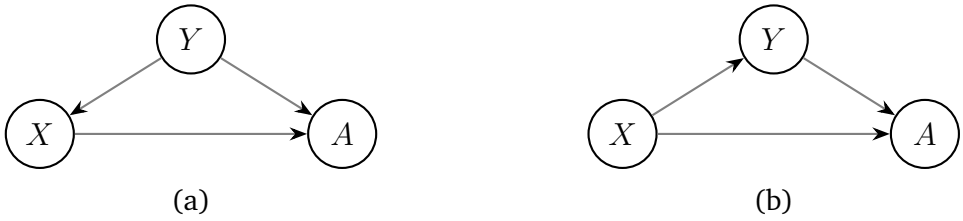


Fig. 2.2.: (a) Simplified causal model inferred from Hathaway [2007] of the effect of X (democracy) on A (treaty ratification), which is confounded by Y (torture practice); (b) Modified causal model adapted from Hathaway [2007] of the effect of X (democracy) on A (treaty ratification) both directly and indirectly through Y, suggesting a potential post-treatment bias in the simplified model.

For a more complicated example, the study by Vreeland [2008] raises the possibility of omitted variable bias in explaining the *positive* correlation between CAT ratification and torture practices in dictatorships. The situation is represented in Figure 2.3 where the vertices X, Y, and A respectively denote multiple parties, torture, and CAT ratification. Failing to condition on X in this case would confound the potential (non)relationship between Y and A and explain why "the more a dictatorship practices torture, the more likely it is to sign and ratify the CAT" [Vreeland, 2008, 68].

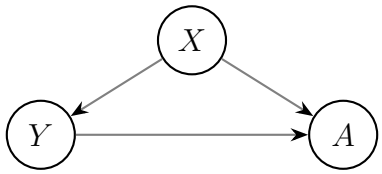


Fig. 2.3.: Simplified causal model inferred from Vreeland [2008] of the effect of X (multiple parties) on both Y (torture) and A (CAT ratification).

Assuming the goal of Vreeland [2008] is to make causal inference, we can infer from his statistical models various causal models that the author implicitly assumes. Table 1 in Vreeland [2008, 83] presents multiple regression models that estimate the instantaneous effect of multiple political parties on torture among dictatorships. These models are represented in Figure 2.4a where X denotes multiple parties, Y

denotes torture, and W_1 is a set of control variables (gross domestic product per capita, population, trade/GDP, civil war, and communist regime). I add the node S in double circle to indicate the sample selection of only dictatorships.[1]

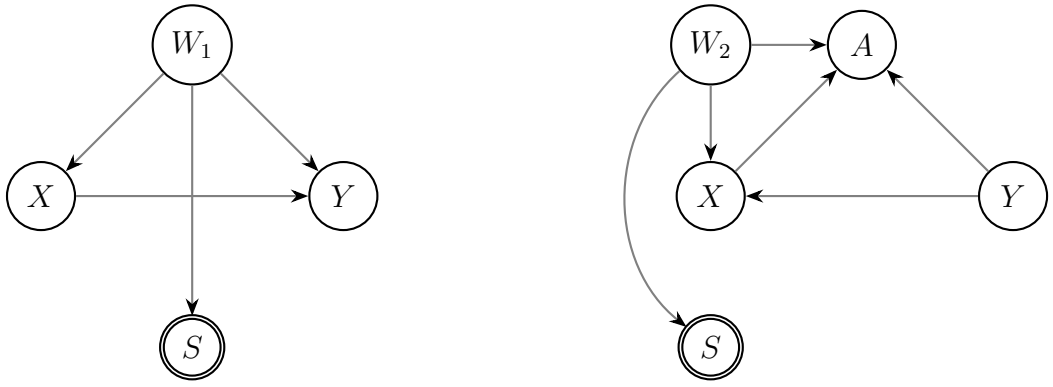


Fig. 2.4.: (a) Causal model inferred from Vreeland [2008, 83] of the effect of X (multiple parties) on Y (torture) among S (dictatorships) with control variables W_1; (b) Causal model inferred from Vreeland [2008, 90] of the effect of X (multiple parties) on A (CAT ratification) among S (dictatorships) with control variables W_2. Arrows of opposite directions between X and Y across the two causal models suggest incoherent assumptions about the causal process.

Vreeland [2008] then proceeds to estimate the instantaneous effect of multiple parties (X) on CAT ratification (A) among dictatorships (S). His regression models in Table 3 [Vreeland, 2008, 90] assume the causal model in Figure 2.4b where W_2 is a different set of control variables (communist regime, lagged regional score of CAT ratification, the number of countries that have ratified the CAT, the percentage

[1]The original study does not discuss sample selection and its consequences for identification. Here I assume that sample selection S, which is based on regime type, is dependent on the control variables W_1 and W_2. This is not unreasonable since democracy arguably depends on economic development, the presence or absence of civil war, trade, among others. This assumption is also convenient because we can then remove from consideration the consequences of sample selection in order to focus on the causal relationships between multiple parties and, respectively, torture and treaty ratification. In other cases, though, as Bareinboim et al. [2014] demonstrate, sample selection could potentially render the causal effect of X on Y in Figure 2.4a non-identifiable from the sample data. For example, insofar as legally organized political parties (treatment X) and torture (outcome Y) both influence sample selection S, that is, the use of torture may suppress and undermine democracy ($Y \rightarrow S$) while mobilization by opposition parties promotes democratization ($X \rightarrow S$), we will end up with a collider bias $X \rightarrow \boxed{S} \leftarrow Y$ and the causal effect of X on Y will not be recoverable from the sample data.

of the population that are Muslims, GDP per capita, population, and the trade/GDP proportion). Vreeland [2008, 89] also controls for "the log of the Hathaway torture scale." This is a curious modeling decision, however, since it implies that Y is a confounding variable that affects both X and A. Thus, it can be seen that between the causal model in Figure 2.4a (where $X \rightarrow Y$) and the causal model in Figure 2.4b (where $Y \rightarrow X$), some incoherent assumptions are made with respect to the contemporaneous causal relationship between multiple parties and torture. If multiple parties only affect torture as assumed in Figure 2.4a but not the other way around, then controlling for torture as Vreeland [2008, 90] does would introduce a post-treatment bias. It might be that X and Y mutually cause each other instantaneously, but then it would not be possible to identify the causal effect of X (multiple parties) on either A (CAT ratification) or Y (torture).

It should be emphasized that I remain agnostic at this point as to whether these causal models accurately depict the true underlying causal process or which specific statistical methods are used to estimate the causal quantities of interest from observational data. Nevertheless, the two examples illustrate the critical importance of graphically representing our causal models. A graphical model would make explicit our assumptions, consistent or otherwise, about the underlying data-generating process and reveal potential identification problems that may arise.

2.3 Causal Variable Importance Analysis of Treaty Ratification

2.3.1 Notation and causal model formulation

Traditional variable importance analyses use parametric models to estimate the association between input variables and an outcome, using a variety of metrics such as regression coefficients and p-values, model fit, or predictive accuracy. Taking a causal inference approach, rather than an associational one, I instead formulate variable importance in terms of their average causal effects. Informally, the causal effect of a variable is defined as the effect of an intervention to artificially fix, as

opposed to just naturally observe, the values of that variable. For a binary variable, the treatment and control values are intuitively clear. For a continuous variable, I use its observed maximum and minimum values.

In an observational setting, the first step in identifying and estimating causal effects is to build a non-parametric structural causal model as a set of equations to describe, to the best of our knowledge, the underlying data-generating process. In my following model, W is a set of time-invariant covariates; $X1$ and $X2$ are either binary or continuous time-varying predictors; Y is human rights outcome; and A is treaty ratification.[2] The subscript t indicates the time periods during which the variables are measured. Together these equations form a generative system from which n country–year observations O_n are sampled and the joint probability distribution of the observed data is $O_n = (W, X1_t, X2_t, A_t, Y_t) \sim P_O$.

$$
\begin{aligned}
W &= f_W(U_W) \\
X1_t &= f_{X1}(W, A_{t-1}, Y_{t-1}, X1_{t-1}, X2_{t-1}, U_{X1}) \\
X2_t &= f_{X2}(W, A_{t-1}, Y_{t-1}, X1_{t-1}, X2_{t-1}, U_{X2}) \\
A_t &= f_A(W, A_{t-1}, Y_{t-1}, X1_t, X2_t, U_A) \\
Y_t &= f_Y(W, Y_{t-1}, A_t, X1_t, X2_t, U_Y)
\end{aligned}
\tag{2.1}
$$

A structural causal model is best represented in the form of an acyclic directed graph (DAG). A causal DAG [Darwiche, 2009, Elwert, 2013, Pearl, 2009a] comprises a set of nodes/vertices denoting random variables. An edge/arrow denotes one variable's (the parent node) direct causal influence on another node (the child node). A path in a causal DAG is an arrow or a sequence of arrows, regardless

[2]Quantitative research on international human rights law mostly focuses on the influence of human rights treaties on state practices. It therefore often considers treaty ratification as the treatment, the impact of which is to be evaluated. In the epidemiology and biomedical literature, from which I derive a lot of methodological insights, the treatment is usually denoted A and the outcome Y. To be consistent with the larger research program on international human rights law, throughout the chapter I use A to denote treaty ratification, which is the *outcome* in this study. The *treatments* in my causal variable importance analysis are ratification predictors denoted X such as $\{X1, X2\}$. As annotated and explained later in my graphical causal model, human rights practice, denoted Y, is actually a potential *confounder*.

of their directions, that connects one node to another. A causal (or directed) path have all arrows on its path point to the same direction. Otherwise, it is a non-causal path.

My causal DAG in Figure 2.5 has a dynamic structure that reflects a temporal order with past nodes in the left shaded block and future nodes in the right shaded block. Each block represents a single time period. There are no arrows or sequence of arrows going from the block on the right to the block on the left, meaning that no variable in the future should have a causal influence on any variable in the past. The DAG is also acyclic in the sense that, within the same temporal block, there are no loops or directed paths going from a node to itself. I make no assumptions about any of the functional forms $f = \{f_W, f_{X1}, f_{X2}, f_A, f_Y\}$, which is consistent with the recognition that usually we do not have enough knowledge to specify the exact functional forms that characterize the relationships between variables. For the sake of simplicity and without loss of generality, I construct a causal model with only two time-varying predictors $X1$ and $X2$ over two time periods from $t-1$ to t. A larger number of predictors over a longer time span can be represented in a similar fashion.

As in any causal analyses, we have to make a few assumptions about the underlying causal process. Similar to Díaz et al. [2015, 6], I assume ratification predictors do not instantaneously affect each other although they may influence every other predictor of the next time period. That means, for example, the amount of official development assistance (ODA) and economic development are conditionally independent from each other in the same time period. ODA at time $t-1$, however, could certainly affect economic development at time t (notationally, $X1_{t-1} \rightarrow X2_t$). From an identification standpoint, this assumption is necessary because if the predictors are allowed to mutually cause each other instantaneously, it would render the causal model cyclical and make it impossible to identify their causal effects.

I further assume the exogenous variables $U = \{U_W, U_{X1}, U_{X2}, U_A, U_Y\}$ are jointly independent. As a result, the values of any node is strictly a function of its parent

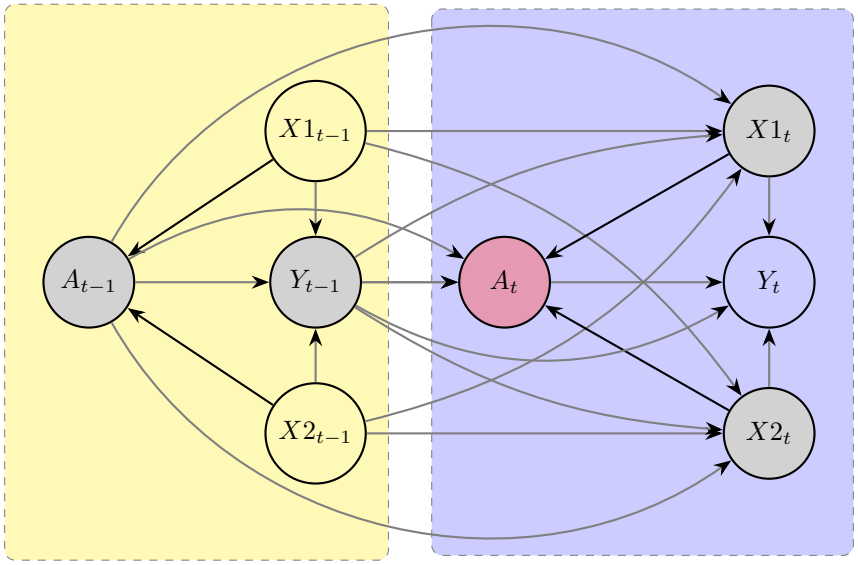


Fig. 2.5.: A dynamic graphical causal model with shaded blocks indicating two temporal periods. Time-invariant covariates W, which precede and potentially affect all other variables, are not represented. The sufficient adjustment sets to identify the causal effects of $X1_t \rightarrow A_t$ and $X2_t \rightarrow A_t$ are $\{W, A_{t-1}, Y_{t-1}, X2_t\}$ and $\{W, A_{t-1}, Y_{t-1}, X1_t\}$, respectively.

nodes and some exogenous factors. This implies that observing a variable's parent nodes will render that variable independent from other covariates except for its descendants. For example, treaty ratification A_t has as its parent nodes time-invariant covariates W, predictors X_t, human rights practice in the immediate past Y_{t-1}, and prior ratification status A_{t-1}. If we observe the set $\{W, Y_{t-1}, A_{t-1}, X_t\}$, then A_t is conditionally independent from other nodes, including all X_{t-1}, except for the descendants of A_t such as Y_t and A_{t+1}.

It should be emphasized that, short of a randomization of the treatment as in an experimental design, any observational studies that aim to make causal inference have to make this exogeneity assumption and the only way to justify it is to rely on the domain knowledge in the literature (Table 2.1). In other words, since one cannot know if a model accurately represents the causal process based on a scrutiny

of the observed data alone, it is important that the body of knowledge in the literature should guide and justify the construction of my causal model as follows. First, the causal dependence $Y_{t-1} \rightarrow A_t$ is informed by the selection effect argument that a state may make a commitment decision based in part on its prior level of compliance because they will significantly determine its ratification cost [Downs et al., 1996, von Stein, 2005].

Second, I allow for the causal dependencies $X_{t-1} \rightarrow X_t$ and $Y_{t-1} \rightarrow Y_t$. This is a routine assumption in the context of time-series cross-section data structure. Substantively, this assumption also permits the possibility that human rights violations may have some inherent dynamic that goes beyond contextual factors such as poverty, dictatorship, involvement in conflicts, and so forth. As Hill and Jones [2014, 674] observe, this argument means that "the governments can become habituated to the use of violence to resolve political conflict." I include this causal relationship, bearing in mind that, in a graphical causal model, an arrow between variables indicates a possible, but not necessarily an actual causal link. A missing arrow, on the other hand, is equivalent to ruling out any direct causality.

Third, an argument can also be made that human rights practices affect some ratification predictors in the next time period. An obvious example is that the use of torture and other extrajudicial measures by the government could intimidate its critics, suppress movements for democratization, and undermine democracy. The inclusion of the directed arrows $Y_{t-1} \rightarrow X1_t$ and $Y_{t-1} \rightarrow X2_t$ in my causal model is informed by this argument.

Fourth, I similarly speculate a direct causal dependence $A_{t-1} \rightarrow A_t$ based on the observation that once governments ratify an international human rights treaty, they are unlikely to withdraw from that treaty. It should be noted that in many cases withdrawal is entirely legally possible. Many human rights treaties and their optional protocols have denunciation provisions that allow states to exit from these institutions, including Article 31 of the CAT, Article 12 of the First Optional Protocol to the ICCPR, and Article 19 of the Optional Protocol to the CEDAW. This is not the

case with the ICCPR and the CEDAW, which do not have a denunciation clause or provision. That, however, has not prevented some states from denouncing and attempting to withdraw from the ICCPR [Tyagi, 2009]. I therefore code treaty membership as an implicit annual ratification as opposed to a terminal event. This is similar to many other studies in the literature [Cole, 2005, Lupu, 2013a, Hafner-Burton et al., 2015a]. Importantly, it is also consistent with the prevailing modeling practices in almost every single study that estimates the impact of human rights treaty ratification as a time-varying treatment.

Finally, the causal dependencies $A_{t-1} \rightarrow X1_t$ and $A_{t-1} \rightarrow X2_t$ suggest that we leave open the possibility that a human rights treaty, once ratified, could influence state behavior in the next time period through a variety of mediators such as public opinion and electoral accountability in democracies [Dai, 2005, Wallace, 2013], legislative constraints of the executive by the opposition parties [Lupu, 2015], and judicial effectiveness of the domestic court system [Crabtree and Fariss, 2015, Powell and Staton, 2009].

Table 2.1 lists the model variables and data sources for their measurements. It also refers to studies in the literature that similarly classify or assume these variables as time-invariant covariates, confounders, and ratification predictors. For example, if a study that investigates the impact of a human rights treaty on state practice includes democracy and independent judiciary as time-varying control variables in its statistical models, we can infer that study views these two covariates as ratification predictors. Appendix A.1 provides more detailed variable descriptions, coding, and data sources.

Given the causal model and its encoded assumptions, I formulate the causal importance of a predictor in terms of its contemporaneous average causal effect, that is, the difference in the average probability of ratifying a treaty if that predictor has one value as opposed to a different value across all country–year observations. It is denoted by $\tau = E\big[A_t|do(X1_t = 1)\big] - E\big[A_t|do(X1_t = 0)\big]$ where the do-operator is notation for an active intervention to fix the value of $X1$. In

Table 2.1.: Ratification model variables

Sets	*Variables and references*
W	Ratification rules [Simmons, 2009] measured by Simmons [2009]. Domestic legal traditions [Mitchell et al., 2013] measured by La Porta et al. [2008].
X	ICCPR proportion of ratification globally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by Office of High Commissioner for Human Rights (OHCHR). CEDAW proportion of ratification globally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by OHCHR. CAT proportion of ratification globally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by OHCHR. ICCPR proportion of ratification regionally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by OHCHR. CEDAW proportion of ratification regionally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by OHCHR. CAT proportion of ratification regionally [Goodliffe and Hawkins, 2006, Hathaway, 2007] measured by OHCHR. Democracy/dictatorship classification [Hathaway, 2007, Chapman and Chaudoin, 2013, Neumayer, 2007] measured by Cheibub et al. [2010]. Multiple parties [Vreeland, 2008, Hollyer and Rosendorff, 2011] measured by Cheibub et al. [2010]. Transition to/from democracy [Goodliffe and Hawkins, 2006, Moravcsik, 2000] measured by Cheibub et al. [2010]. Involvement in militarized interstate dispute [Chapman and Chaudoin, 2013] measured by Melander et al. [2016] and Gleditsch et al. [2002]. Judicial independence [Powell and Staton, 2009] measured by Linzer and Staton [2015]. Population size [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators. Gross domestic product (GDP) per capita [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators. Participation in international trade [Hafner-Burton, 2013] measured as trade volume/GDP by the World Bank Indicators. Net official development assistance [Nielsen and Simmons, 2015] measured by the World Bank Indicators.
Y	CIRI torture index [Cingranelli et al., 2013]. CIRI women's political rights index [Cingranelli et al., 2013]. Human rights dynamic latent score [Fariss, 2014].
A	ICCPR ratification measured by OHCHR. CEDAW ratification measured by OHCHR. CAT ratification measured by OHCHR.

the interventional framework of causal inference [Pearl, 2009a], that means we would intervene on the generative system (Equation set 2.1) to fix the equation

$X1_t = f_{X1}(W, A_{t-1}, X1_{t-1}, X2_{t-1}, U_{X1})$ reiteratively at $X1_t = \{0, 1\}$. From the two resulting modified generative systems $A_t = f_A(W, A_{t-1}, Y_{t-1}, x, X2_t, U_A)$ for $x = \{0, 1\}$, we then compute the difference between the two mean values of treaty ratification, which will be a consistent estimate of the causal effect of $X1$ as long as causal identification is established.

2.3.2 Causal identification

Causal identification involves establishing the conditions under which a property of an interventional distribution such as the expectation $E\big[A|do(X = x)\big]$ can be computed without bias from an observational probability distribution. My causal identification strategy is to identify a valid adjustment set of observed variables that makes the interventional distribution of the outcome A (treaty ratification) essentially equivalent to its observed conditional distribution.

Any causal identification in the setting of observational data ultimately depends on the underlying causal structure, which is best represented by a causal DAG. DAGs are thus an effective tool to make all causal assumptions transparent and facilitate a clear and easy determination of sufficient adjustment sets using the backdoor criterion. To illustrate identification of the causal effect of $X1_t$ on A_t, for example, I apply the following backdoor criterion [Pearl et al., 2016, 61–66] to find an adjustment set of variables such that conditioning on that set will:

(a) block any (non-causal) paths from $X1_t$ to A_t that have an arrow coming into $X1_t$;

(b) leave open all causal paths from $X1_t$ to A_t; and

(c) not condition on a collider (a node that lies on any paths between $X1_t$ and A_t and has two arrows coming into it) or a descendant of a collider (a node connected to a collider through a directed path emanating from the collider).

When we condition on an adjustment set that satisfies the backdoor criterion, we essentially remove all non-causal pathways from $X1_t$ to A_t and render these two variables conditionally independent or d-separated and, as a result, the interventional distribution of the outcome A when $X1$ is intervened upon is essentially equivalent to its observational distribution. More generally, when all non-causal paths between a predictor and the outcome are closed off, any remaining significant correlation between them is evidence of a causal relationship.

From the graphical causal model in Figure 2.5, I derive a sufficient set of covariates for adjustment $Z_1 = \{W, A_{t-1}, Y_{t-1}, X2_t\}$ that satisfies the backdoor requirement to identify the causal effect of $X1_t$ on A_t. Specifically, conditioning on Y_{t-1} will, according to rule (a), block five non-causal paths from $X1_t$ to A_t, including (i) $X1_t \leftarrow A_{t-1} \rightarrow \boxed{Y_{t-1}} \rightarrow A_t$; (ii) $X1_t \leftarrow X1_{t-1} \rightarrow A_{t-1} \rightarrow \boxed{Y_{t-1}} \rightarrow A_t$; (iii) $X1_t \leftarrow \boxed{Y_{t-1}} \rightarrow A_t$; (iv) $X1_t \leftarrow \boxed{Y_{t-1}} \rightarrow X2_t \rightarrow A_t$; and (v) $X1_t \leftarrow X2_{t-1} \rightarrow \boxed{Y_{t-1}} \rightarrow A_t$. Similarly, conditioning on A_{t-1} will, by the same rule, block two other non-causal paths from $X1_t$ to A_t, including (i) $X1_t \leftarrow \boxed{A_{t-1}} \rightarrow A_t$ and (ii) $X1_t \leftarrow \boxed{A_{t-1}} \rightarrow X2_t \rightarrow A_t$.

However, Y_{t-1} is also a collider on the path $X1_t \leftarrow X1_{t-1} \rightarrow \boxed{Y_{t-1}} \leftarrow X2_{t-1} \rightarrow X2_t \rightarrow A_t$. Conditioning on Y_{t-1} will therefore open that non-causal path and violate rule (b) of the backdoor requirement. I therefore further condition on $X2_t$ to block this non-causal path. For the same reason that I have accidentally opened the non-causal path $X1_t \leftarrow X1_{t-1} \rightarrow \boxed{A_{t-1}} \leftarrow X2_{t-1} \rightarrow X2_t \rightarrow A_t$ when conditioning on the collider A_{t-1}, I block this path by conditioning on $X2_t$. Conditioning on $X2_t$ also happens to block three other non-causal paths that traverse through $X2_t$, including (i) $X1_t \leftarrow X2_{t-1} \rightarrow \boxed{X2_t} \rightarrow A_t$; (ii) $X1_t \leftarrow A_{t-1} \rightarrow \boxed{X2_t} \rightarrow A_t$; and (iii) $X1_t \leftarrow X2_{t-1} \rightarrow A_{t-1} \rightarrow \boxed{X2_t} \rightarrow A_t$. The latter two of these three non-causal paths run through A_{t-1} as well and therefore are already blocked when we condition on A_{t-1}.

We should not condition on contemporaneous measure of human rights practice Y_t when estimating the causal effect of $X1_t$, however. Since it is a collider on the

path $X1_t \rightarrow Y_t \leftarrow A_t$, conditioning on Y_t would violate rule (c) of the backdoor criterion, introducing a non-causal association between $X1_t$ and A_t and biasing the causal effect estimate of $X1_t$. For identification of the causal effect of $X2_t$ on A_t, I apply the same rules and similarly derive a sufficient adjustment set $Z_2 = \{W, A_{t-1}, Y_{t-1}, X1_t\}$. In summary, to identify the contemporaneous causal effect of a ratification predictor, I condition on time-invariant covariates, immediately prior ratification status and level of compliance, and other contemporary time-varying covariates.

In addition to a causal variable importance analysis, I use the same graphical causal model to develop a causal test of many theories of CAT ratification. First, I test the argument by Hathaway [2007] that democracy ($X1_t$) and torture practices (Y_t) interact to lower the probability of CAT ratification (A_t). Based on the causal DAG in Figure 2.5, one should not condition on Y_t or, for that matter, use an interaction term of Y_t and $X1_t$ while estimating the effect of $X1_t$ on A_t. Since Y_t is a collider on two different paths $X1_t \rightarrow \boxed{Y_t} \leftarrow A_t$ and $X1_t \rightarrow \boxed{Y_t} \leftarrow Y_{t-1} \rightarrow A_t$, conditioning on Y_t will induce a collider bias. I instead causally test this interactive effect argument by estimating the Y_{t-1}-specific effect of $X1_t$ on A_t, using the adjustment set $Z = \{W, A_{t-1}, X2_t\}$ that satisfies the backdoor requirement within each subset of observations based on the values of Y_{t-1} [Pearl et al., 2016, 71–72]. The test results will provide evidence as to whether there is any effect modification by past torture practice, that is, whether the effect of democracy on treaty ratification varies across levels of compliance in the previous year. The conventional expectation is that the positive causal effect of democracy on treaty ratification will diminish and eventually reverse its direction as the level of torture in the prior year increases. Note that we cannot identify the $X1_t$-specific causal effect of Y_{t-1} on A_t because of potential post-treatment bias since $X1_t$ could be a descendant of Y_{t-1} along the path $Y_{t-1} \rightarrow X1_t \rightarrow A_t$ if the use of torture possibly undermines democratic institutions.

Second, I test Vreeland's omitted variable bias argument by directly estimating the causal effect of multiple political parties ($X2$) on CAT ratification (A) among dictatorships ($X1 = 0$). The quantity of interest corresponding to the test is formulated as the $X1_t$-specific causal effect of $X2_t$ on A_t, that is, the causal effect of multiple parties on treaty ratification among observations with the value $X1_t = 0$. The sufficient adjustment set for identification is $Z = \{W, A_{t-1}, Y_{t-1}, X1_t\}$. As Vreeland [2008, 79] predicts, "the effect of the multiparty institution is to make a dictatorship more likely to enter into the CAT," implying a positive causal effect of multiple parties.

Third, I estimate the average causal effect of prior torture practice on CAT ratification ($Y_{t-1} \rightarrow A_t$) in a causal test of the selection effect argument. This argument is often made but has rarely been empirically quantified within a causal inference framework. The theoretical expectation is a negative causal effect of Y_{t-1}, suggesting that higher level of torture in the previous year is expected to cause state leaders to be less likely to ratify the CAT in the following year. A sufficient adjustment set I derive for identification is $Z = \{W, A_{t-1}, X1_{t-1}, X2_{t-1}\}$.

Finally, I also test the argument with respect to the signaling benefits of CAT ratification for dictators [Hollyer and Rosendorff, 2011] by estimating the causal effect of torture on CAT ratification among autocracies, that is, the $X1_{t-1}$-specific causal effect of Y_{t-1} on A_t. The theoretical expectation is that "authoritarian governments that torture heavily are more likely to sign the treaty than those that torture less" [Hollyer and Rosendorff, 2011, 276], which implies a positive effect of Y_{t-1} among observations that have the value $X1_{t-1} = 0$. A sufficient set that satisfies the backdoor criterion for causal effect identification is $Z = \{W, A_{t-1}, X2_{t-1}\}$.

2.3.3 Machine learning-based estimation

Once we have determined the sufficient adjustment sets Z that satisfy the backdoor requirement for identification of various causal effects, I adopt two machine

learning-based methods for causal effect estimation: substitution estimation and targeted maximum likelihood estimation (TMLE). My estimation methods are analogous to the OLS estimator if the underlying causal system in Equation set 2.1 is assumed to be linear and all covariate effects are additive and all the noise terms U are Gaussian. The use of machine learning is aimed to relax this assumption.

For each of the continuous predictors of treaty ratification (including, global proportion of ratification, regional proportions of ratification, population size, GDP per capita, trade/GDP proportion, net amount of ODA, and judicial independence) the substitution estimator [Robins, 1986, Robins et al., 1999] computes the average causal effect of the predictor $\tau = E\big[A|do(X = 1)\big] - E\big[A|do(X = 0)\big]$, using the estimate $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\big[Q_n(1, Z) - Q_n(0, Z)\big)\big]$. Specifically, I fit a prediction model $\bar{Q}_n(X, Z) = E\big[A|X, Z\big]$ of treaty ratification A using X and the corresponding sufficient adjustment set Z. I then reiteratively substitute the predictor values with $X = 1$ (empirically maximum value) and $X = 0$ (empirically minimum value) for each observation, use the fitted prediction model to generate the counterfactual outcomes, and compute the mean difference.

For variance estimation, I use the nonparametric bootstrap method. In the presence of missing data, my procedure is similar to Daniel et al. [2011, 491] and suggested by Tsiatis [2007, 362–371]. I combine bootstrap with single stochastic imputation rather than multiple imputation in order to make efficient and still valid inference. In addition to its greater efficiency, another benefit of combining nonparametric bootstrap and single (improper) imputation is that we do not have to rely on the normality assumption as required by the Rubin's approach [Little and Rubin, 2014] when pooling variances across imputed datasets. Instead, I create distribution-free confidence intervals, using the 2.5% and 97.5% quantiles of the bootstrap distribution to obtain the desired coverage.

The key to obtaining consistent effect estimates with a substitution estimator is to fit a correctly specified outcome model $\bar{Q}_n$ that approximates the (unknown) data generating mechanism. The standard practice is to assume a binomial dis-

tribution for the binary outcome of treaty ratification and then model a property of the outcome distribution as a linear, additive function of a set of covariates, sometimes with an interaction term included. If these distributional and functional form assumptions are wrong, which they likely are for probably non-linear, highly complex political phenomena, the results will be misspecified models, biased effect estimates, invalid inference, and misleading conclusions. The ensemble machine learning technique Super Learner [van der Laan et al., 2007, Sinisi et al., 2007] offers a powerful solution to this problem of correct functional forms.

Super Learner has been used in economic research [Kreif et al., 2015], political science [Samii et al., 2016], and epidemiology [Neugebauer et al., 2013, Pirracchio et al., 2015]. It stacks a user-selected library of predictive algorithms and uses cross-validation to evaluate the performance of each algorithm in minimizing a specified loss function. For the binary outcome of treaty ratification, an appropriate loss function is the negative log-likelihood $-log\Big[Q(X,Z)^A\big(1-Q(X,Z)\big)^{1-A}\Big]$, which measures the degree of misfit with the observed data. User-selected predictive algorithms can include simple main-term linear regression model, semi-parametric generalized additive model [Hastie and Tibshirani, 1990], regularized regression models [Tibshirani, 1996], and non-parametric tree-based ensemble methods such as boosting [Friedman, 2001] and random forest [Breiman, 2001b]. Table 2.2 lists the algorithms I use for my machine learning-based substitution estimation given the constraints in terms of computational resources.

Table 2.2.: Algorithms used in Super Learning-based Substitution Estimation

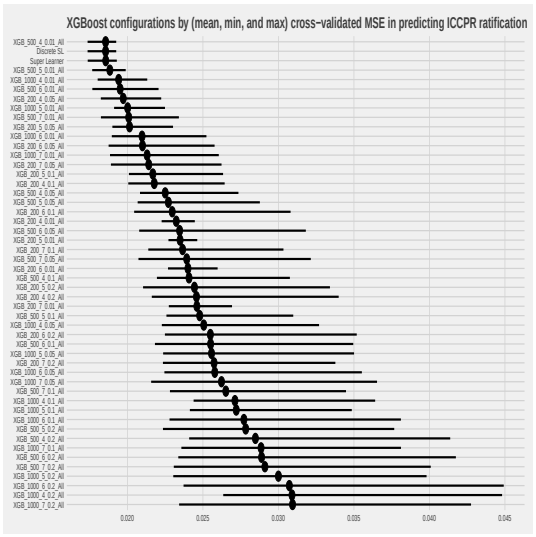
Algorithm	*Description*
GLMnet	Regularized logistic regression with lasso penalty $\sum_{j=1}^{p}\|\beta_j\|$.
GAM	Generalized additive model.
(Tuned) XGBoost	Extreme gradient boosting (eta = 0.01, depth = 4, ntree = 500).

The use of cross-validation is crucial for the algorithms to generalize well in terms of predicting unknown outcome values and avoiding overfitting. Super Learner

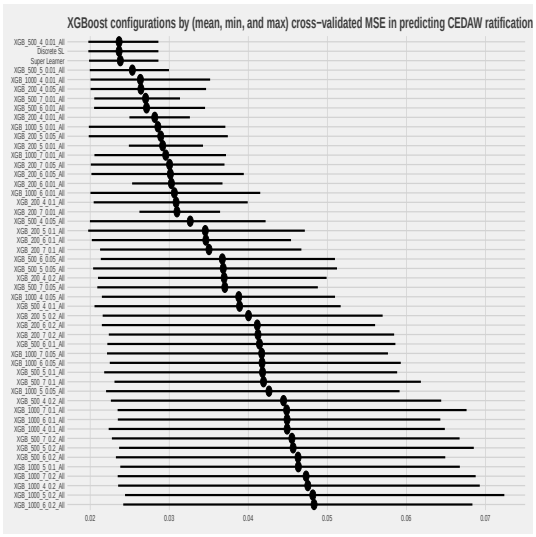
then creates a linear combination of these algorithms, each of which is weighted by its average predictive accuracy, to build a hybrid prediction function that performs approximately as well as and usually better than the best algorithm in the library. The ability of Super Learner to assemble a rich, diverse set of algorithms makes it particularly effective and much more likely to approximate the underlying data generating process [Polley and van der Laan, 2010].

One state-of-the-art algorithm is extreme gradient boosting [Chen and He, 2015, Chen and Guestrin, 2016], a faster implementation of the popular and effective machine learning technique of gradient boosting machine [Friedman, 2001, Schapire and Freund, 2012, Natekin and Knoll, 2013]. Extreme gradient boosting (XGBoost) is non-parametric and able to capture non-linear, interactive dynamics among a large number of predictors. Furthermore, unlike other tree-based methods such as random forest and gradient boosting machine, XGBoost has greater computational efficiency, which makes it particularly suitable to use in the context of nonparametric bootstrap for inference.

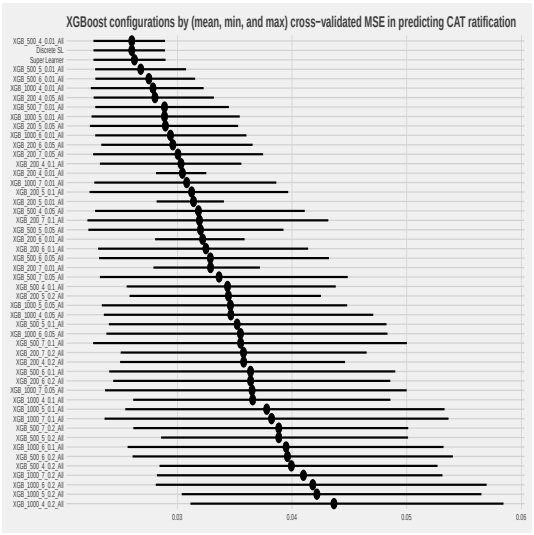
The performance of XGBoost could be sensitive to different hyper-parameter settings. I employ a combination of 5-fold cross-validation and grid search in Figure 2.6 to select the best among a large number of configurations (comprising varying learning rates, tree depths, and numbers of trees) that are tuned specifically to each of the three singly imputed ICCPR, CEDAW, and CAT datasets. Each configuration of XGBoost hyper-parameters is iteratively trained on a random sample of four-fifths of the country–year observations and then used to predict the probability of treaty ratification, using the last fifth of the data. For each configuration, its minimum, maximum, and average mean-squared-prediction-errors across five folds are plotted on the graph in descending order. The most effective configuration is the one on the top with the smallest average mean-squared-prediction-error. This data-driven selection process will help us decide which configuration of the XGBoost algorithm is most effective in predicting treaty ratification for each human rights treaty and thus, presumably, most accurately captures the underlying data-generating process.



(a)



(b)



(c)

Fig. 2.6.: Cross-validated risk of XGBoost algorithms in predicting (a) ICCPR ratification, (b) CEDAW ratification, and (c) CAT ratification.

To estimate the causal effect of the binary predictors (democracy, multiple political parties, democratic transition, and involvement in militarized interstate disputes), I use the targeted maximum likelihood estimation (TMLE) method [van der Laan and Rose, 2011]. Similar to the substitution estimator, TMLE also starts by fitting an initial predictive outcome model of treaty ratification $Q_n^0 = E(A|X, Z)$.

It then modifies the initial model $Q_n^0(X, Z)$ into an updated model $Q_n^1(X, Z)$, using the modifying equation $\text{logit}(Q_n^1) = \text{logit}(Q_n^0) + \epsilon_n H_n$ where the "clever covariate" $H_n(X, Z) = \left[\frac{I(X=1)}{g_n(X=1|Z)} - \frac{I(X=0)}{g_n(X=0|Z)}\right]$ is a function of the treatment mechanism $g_n = E(X|Z)$ and the coefficient ϵ_n is obtained via a separate regression model $\text{logit}(A) = \text{logit}(Q_n^0) + \epsilon_n H_n$. In the third and final step, TMLE similarly substitutes two distinct values of a binary predictor, plugs them into the updated outcome model $Q_n^1(X, Z)$ to generate the counterfactual outcomes for each observation, and computes the average causal effect as the mean difference of the counterfactual outcome values.

TMLE is essentially the substitution estimator but with an additional updating step in between to incorporate information about treatment assignment. This updating step is at the heart of the TMLE methodology. It makes the estimator doubly robust by reducing any remaining bias in the initial outcome model, producing unbiased estimates if either the initial outcome model Q_n^0 or the treatment assignment model g_n is consistent. It is maximally efficient asymptotically if both Q_n^0 and g_n are consistent. Note that both Q_n^0 and g_n are already more robust to misspecification, and thus more likely to be consistent than standard parametric statistical models, because I have incorporated machine learning in my estimation.

In short, the TMLE methodology computes causal effect estimates of binary treatment variables that are more robust than both parametric regression models and propensity score-based estimators. Machine learning-based TMLE is even more robust and less computationally expensive than the machine learning-based substitution estimator with bootstrapped samples thanks to its efficient influence function-based approach to variance estimation [van der Laan and Rose, 2011, 94–97]. Because of TMLE's greater computational efficiency, I am able to employ a more diverse and richer set of learning algorithms in Table 2.3.

Table 2.3.: Algorithms used in Super Learner-based Targeted Maximum Likelihood Estimation

Algorithm	*Description*
GLMnet	Regularized logistic regression with lasso penalty $\sum_{j=1}^{p}\|\beta_j\|$.
GAM	Generalized additive model (degree of polynomials = 2).
polymars	Polynomial multivariate adaptive regression with splines.
randomForest	Random Forest (ntree = 1,000).
XGBoost	Extreme gradient boosting (eta = 0.01, depth = 4, ntree = 500).

To handle the missing data, I conduct multiple imputation, using the Amelia II program [Honaker et al., 2011], and combine estimates across $m = 5$ imputed data sets. Appendix A.1 provides the summary statistics of the observed data and Appendix A.3 summarizes the imputation process. The ICCPR, the CEDAW, and the CAT were opened for ratification at different times. I thus create three separate datasets (and, correspondingly, 15 imputed datasets) that have different temporal coverage periods, including 1967–2013 for the ICCPR (opened for ratification in 16 December 1966), 1982–2013 for the CEDAW (adopted and opened for ratification in 18 December 1979, but the CIRI measure of women's political rights only begin in 1981), and 1985–2013 for the CAT (opened for ratification in 10 December 1984). For algorithmic learning stability and ease of interpretation, I standardize all continuous covariates into a bounded range between zero and one.

2.3.4 Results and interpretation

Table 2.4 reports the estimates of the contemporaneous average causal effects of the ratification predictors. Despite some differences, their causal effect estimates are relatively consistent across three human rights treaties. First, the results underscore the importance of regional socialization and norm diffusion in causing states to ratify human rights treaties. Going from the observed lowest proportion to the observed highest proportion of regional ratifications will increase a country's probability of becoming and remaining a state party by somewhere between 7.2 and

9.5 percentage points, depending on the treaties. Similar to a finding by Simmons [2009], density of regional ratification is, in fact, the single most causally consistent and the second most causally important predictor of treaty ratification across all three human rights treaties.

Table 2.4.: Causal effect point estimates and 95% CI of predictors on treaty ratification

Predictors	ICCPR	CEDAW	CAT
Super Learner-based Targeted Maximum Likelihood Estimator Influence function-based CI with multiple imputation			
Democracy	**0.237** [**0.121, 0.353**]	**0.116** [**0.064, 0.168**]	0.093 [−0.065, 0.251]
Multiple parties	0.153 [−0.063, 0.370]	0.197 [−0.114, 0.508]	**0.192** [**0.040, 0.344**]
Democratic transition	0.186 [−0.080, 0.451]	0.091 [−0.046, 0.227]	−0.013 [−0.144, 0.118]
Involvement in militarized interstate disputes	−0.004 [−0.015, 0.007]	−0.002 [−0.017, 0.013]	−0.010 [−0.023, 0.004]
Super Learner-based Substitution Estimator Bootstrap ($B = 500$) quantile-based CI with single stochastic imputation			
Global proportion of ratification	−0.011 [−0.032, 0.000]	−0.011 [−0.025, 0.000]	−0.019 [−0.042, 0.002]
Regional proportions of ratification	**0.095** [**0.039, 0.190**]	**0.072** [**0.034, 0.155**]	**0.094** [**0.033, 0.241**]
Population size	0.009 [−0.004, 0.027]	**0.025** [**0.001, 0.087**]	**0.028** [**0.005, 0.056**]
GDP per capita	−0.003 [−0.020, 0.011]	−**0.017** [−**0.043**, −**0.001**]	0.037 [−0.007, 0.121]
Trade/GDP	−0.002 [−0.015 , 0.011]	0.007 [−0.010, 0.032]	0.003 [−0.014, 0.016]
Net official development assistance	0.014 [−0.010, 0.043]	0.003 [−0.025, 0.019]	0.004 [−0.027, 0.025]
Judicial independence	−0.005 [−0.031, 0.014]	**0.029** [**0.004, 0.094**]	0.024 [−0.008, 0.108]
Number of countries	192	192	192
Number of years	47	32	29
Number of observations	7,870	5,823	5,354

Second, also similar to other studies in the literature [Landman, 2005], my findings further confirm that democracy is a significant predictor of treaty ratification. In fact, I find that democracy is the most causally important variable for the ratification of the ICCPR and the CEDAW. Being a democracy causes the probability of being a state party to these two treaties to go up by 23.7 and 11.6 percentage points, respectively. Democracy is being defined here as having direct election of the executive, election of the legislature, and an alternation of power, among other criteria [Cheibub et al., 2010]. The coding criteria for democracy, in other words, are unlikely to overlap conceptually with various measures of human rights outcomes [Hill, 2016b, von Stein, 2016]. By implications, my findings suggest that the best way to push a state to ratify and remain committed to human rights treaties is to support its domestic democratic institutions and promote ratifications by its regional neighbors. In the case of CAT ratification, it should be cautioned, it is not democracy per se that has a significant causal impact. Rather, it is the existence of de facto multiple political parties that increases the probability of ratification by 19.2 percentage points.

Third, as to other predictors, their causal importance is either very limited or inconsistent. Like Goodliffe and Hawkins [2006], I find that democratic transition does not significantly affect ratification of any treaties, indicating a lack of empirical support for the "lock in" argument. Involving in militarized interstate disputes is not causally important, either. My findings also corroborate the skepticism by Nielsen and Simmons [2015] with respect to many economic variables such as economic development, the amount of ODA received, and participation in international trade. These variables do not seem to matter causally for human rights treaty ratification. Population size tends to have a significantly positive, but substantively very small, causal impact, averaging about two percentage points across all three treaties. Independence of the judiciary makes states slightly more likely to ratify the CEDAW, but otherwise has no impact on the ratification of the ICCPR and the CAT.

I employ the same template of causal analysis, including graphical identification and machine learning-based TMLE estimation, to test many theories of CAT ratification. The results reported in Table 2.5 offer several interesting findings. First, I find scant evidence to support the commonly accepted argument regarding the interactive effect of democratic institutions and human rights practice on CAT ratification [Hathaway, 2007]. Instead, my findings suggest that, irrespective of a state's torture practice in the year prior, changing the regime type from a dictatorship to a democracy does not lower the probability of its CAT ratification status. If anything, being a democracy causes an increase, not a decrease, by 8.2 percentage points in the chance of becoming and remaining a state party to the CAT even at the highest level of torture practice during the previous year, although this estimate is certainly not statistically significant.

Table 2.5.: CAT ratification theories and causal effect point estimates and 95% CI

Theory tested	Notation	Mean	SE	Lower	Upper
Interactive effect argument					
Democracy w/ No Torture	$X1_t \rightarrow A_t$ at $Y_{t-1} = 2$	0.140	0.075	−0.007	0.287
Democracy w/ Occasion Torture	$X1_t \rightarrow A_t$ at $Y_{t-1} = 1$	0.056	0.047	−0.037	0.148
Democracy w/ Freq. Torture	$X1_t \rightarrow A_t$ at $Y_{t-1} = 0$	0.082	0.071	−0.056	0.221
Omitted variable bias argument					
Multiple parties in Dictatorships	$X2_t \rightarrow A_t$ at $X1_t = 0$	0.050	0.043	−0.034	0.134
Selection effect argument					
Torture in All	$Y_{t-1} \rightarrow A_t$	0.116	0.044	0.029	0.202
Torture in Democracies	$Y_{t-1} \rightarrow A_t$ at $X1_{t-1} = 1$	−0.018	0.012	−0.042	0.005
Credible commitment argument					
Torture in Dictatorships	$Y_{t-1} \rightarrow A_t$ at $X1_{t-1} = 0$	0.201	0.125	−0.043	0.445

One speculative reason could be that the executives in non-compliant democracies do want to ratify and comply because torture practices in the past were more of a legacy of abusive government agencies. Such executives, perhaps under the pressures of the democratic public, could have an incentive to ratify the CAT and even use treaty obligations as a way to constrain domestic abusive forces. In any event, these causal tests partially challenge the conventional wisdom that poorly performing democracies are reluctant to become a treaty member because their

democratic institutions will make subsequent compliance very costly. Nevertheless, there is some evidence, though not extremely solid, that being a democracy does increase the probability of becoming a state party to the CAT by 14 percentage points among those countries that did not practice torture at all—a significantly greater effect than among those that engaged in torture in the immediate past.

Second, as indicated previously, the kind of domestic institutions that significantly improve the probability of a country being a CAT member is not democracy in general, but rather the presence of *de facto* multiple political parties. However, contrary to Vreeland [2008], among the subset of authoritarian regimes, the existence of multiple political parties does not seem to have a highly significant causal impact on treaty ratification. This presents an interesting finding: the causal effect of multiple political parties on CAT ratification can vary significantly, depending on the regime types. It also suggests for further inquiries into the potentially heterogeneous causal effects of different components within the definition of democracy on treaty ratification.

Third, I rescale and dichotomize the CIRI torture index (with zero indicating no torture and one indicating occasional or frequent torture) and test the selection effect argument by directly estimating the causal impact of torture practices on CAT ratification in the following time period. States that engage in occasional or even frequent torture practices are actually 11.4 percentage points *more* likely than those engaging in no torture at all to be a state party to the CAT in the following year. In other words, this is evidence of an *adverse* selection effect. Governments whose prior human rights practices do not conform to international standards tend to self-select into, not away from, the CAT.

For a closer look at this surprising finding about the adverse selection effect, I further disaggregate the sample observations into democracies and dictatorships based on their regime classification during the time period when their human rights practices are recorded so as not to introduce a post-treatment bias. It turns out that among democracies, engaging in torture practices would cause only a small 1.8

percentage points decrease in their chance of being a CAT member the following year. This comports with my previous findings that democracy and rights practices do not significantly interact to determine CAT ratification.

Among dictatorships, though, the estimates are highly variable and uncertain. The point estimate suggests that authoritarian regimes that practice torture are, on average, 20 percentage points more likely to ratify the CAT the following year, which seems to support a claim in the literature that "[t]he empirical record has shown fairly consistently that among non-democracies, the less compliant are as likely (and in some cases even more likely) to ratify" [von Stein, 2016, 661]. However, the high variability of causal effect estimates mean that we do not find solid empirical support for the counterintuitive claim by Hollyer and Rosendorff [2011] that authoritarian leaders may be signaling their strength to opposition groups by way of a CAT ratification. In short, my causal effect estimation indicates that prior torture practices do not significantly make CAT ratification more likely even though it points to a potential existence of an adverse selection effect. This, by implication, reiterates the need to take into account prior rights practices if one wants to single out and estimate the causal impact of CAT ratification on human rights practices. Otherwise, the causal effect of the CAT would be biased *downward* towards zero or even negative and CAT ratification would likely appear to exacerbate human rights violations.

In short, part of this study also speaks to and contends with a substantial segment of the literature surrounding the CAT. To summarize, I find that although only two main factors—regional socialization and the existence of multiple political parties—that drive CAT ratification, few other variables that causally prevent states from joining the CAT. This is reflected in the steady increase in the number ratifiers over the last three decades, going from zero in 1985 to more than 150 in 2013. Neither democracy nor previous human rights abuses represents a causal barrier to CAT ratification. Nor is any combination of these two factors. This is relatively consistent with some of the findings by Goodliffe and Hawkins [2006] despite sig-

nificant differences in terms of inferential approaches, modeling choices, and even measurements. In other words, the CAT as an international human rights institution seems to be very inclusive although not perversely so by only attracting bad state actors with abusive records. Still, it presents a challenge down the road when one examines how much of a causal impact that CAT ratification has on state behavior.

2.4 Conclusion

Machine learning in many respects has outpaced statistical theory in terms of modeling reality [Efron and Hastie, 2016]. Empirical scientists could leverage these powerful prediction methods to make robust causal inference about political behavior and institutions. One area of application is to conduct a causal variable importance analysis [Díaz et al., 2015, Hubbard et al., 2013, Pirracchio et al., 2016, Ahern et al., 2016], in which one replaces traditional measures of variable importance by a more substantively relevant measure: the causal effects of predictor variables. In this chapter, I use causal variable importance as a new test of major theories of treaty ratification. This is the methodological motivation of this research.

The substantive motivation is to use this novel test to help settle or at least provide new insights into the ongoing debate as to why states ratify international human rights treaties. There are three major theoretical approaches to treaty ratification in the literature, proposing different sets of explanatory variables, ranging from economic covariates to normative factors to domestic institutional variables. The best and most substantively relevant theories should be able to identify the most causally important variables for treaty ratification. With that reasoning, my analysis, by estimating the causal effect of each explanatory variable, offers a empirical basis to adjudicate and contribute to the debate about human rights treaty ratification in the quantitative human rights literature.

Based on its causal findings, my research in this chapter casts doubt on the instrumental explanations of human rights treaty ratification. It finds little to no

causal impact by economic variables such as economic development, foreign aid, and international trade participation. There is no evidence that human rights treaty ratification is driven by economic concerns or financial interests. The analysis also finds some mixed support for institutional models of treaty ratification. On the one hand, it questions some of the popular institutional theories that explain treaty ratification as an interactive function of compliance cost and regime types or as a function of democratic transitions or domestic judicial independence. On the other hand, however, it does find that democracy has a significant causal effect on the ratification of the ICCPR and the CEDAW whereas *de facto* existence of multiple political parties is causally important for the CAT ratification. Finally, my causal analysis finds more support for the norms-based theories of human rights treaty ratification that highlight the role of regional socialization, normative conformity, or at least an emulation of the ratification behavior of neighboring countries.

In summary, my theory testing from a causal inference perspective confirms a number of previous findings in the literature while challenging some commonly accepted conventional wisdom. Some of these new findings indicate that democracy and state practices do not interact to determine ratification decisions as often expected in the literature and that states do not self-select into human rights treaty regimes based on their prior compliance. Importantly, my findings have a causal, rather than an associational, interpretation. This causal interpretation is made possible by framing the research questions as causal queries, formulating the corresponding causal quantities of interests within a structural causal model, explicitly representing the underlying causal structure in a graphical form, and linking the interventional distribution of the outcome to the observational distribution of the data via an application of the backdoor criterion.

It should be reiterated that key to this study, as is in any other causal analyses, is an assumption about the causal structure that generates the observed data. In an observational study like this, this assumption has to be justified based on sufficient background knowledge gleaned from the current literature and any causal findings

will depend on the validity of this assumption. Other than that causal structure assumption, the data-adaptive, machine learning-based estimation methods that I use are much less dependent upon distributional and functional form assumptions than other traditional statistical models. This is one of the places where my approach improves upon existing studies by increasing the ability of my machine learning models to accommodate potentially complex relationships among the covariates that may not be captured by standard parametric statistical models.

Methodologically, despite the great promises of machine learning and the structural causal model framework, the dearth of applied research that combines these two methods suggests that there is a gap to bridge between methodological advances in causal inference and machine learning on the one hand and substantive applications in political science research on the other hand. One obvious solution is more collaboration between domain experts and methodologists who are able to apply flexible machine learning methods to different domains. Moreover, in the absence of collaborative research, given that any causal analysis requires a sufficient understanding of the empirical literature, applied researchers who have a firm grasp of the substantive background knowledge are probably better positioned to bridge this gap by adopting machine learning methods in their own research.

Finally, there is a critical need to openly embrace the structural causal model framework in political science given that a lot of research questions in the discipline are explicitly causal queries. This framework has been developed significantly in the last decade or so [Pearl, 2014a] and has been adopted very successfully in sociology, epidemiology, and biomedical research. My application of this framework to the issue of human rights treaty ratification shows that it can help researchers clarify confusion about the assumed underlying causal process, identify incoherence in causal assumptions, and modify our causal models to increase their substantive plausibility. Employing this structural causal inference framework could be extremely beneficial to applied political science research going forward.

3. A MACHINE LEARNING-BASED CAUSAL MEDIATION ANALYSIS OF HUMAN RIGHTS TREATIES

3.1 Introduction

Over the last decade, an expansive body of research has not only investigated the average causal effect of human rights treaties but also attempted to peer into their metaphorical black box of causal mechanisms. Major causal mechanisms are believed to involve institutional and legislative constraints on the executive [Simmons, 2009, Lupu, 2015], domestic judicial litigation and enforcement [Simmons, 2009, Dancy and Sikkink, 2012, Abouharb et al., 2013], political mobilization of non-governmental organizations [Simmons, 2009, Smith-Cannoy, 2012], and international emulation and socialization [Keck and Sikkink, 1998, Goodman and Jinks, 2013, Clark, 2013]. Unfortunately, existing studies in the literature have yet to investigate causal quantities such as natural direct effect, natural indirect effect, and controlled direct effect that are specifically conceived and designed for analyzing causal mechanisms, leaving the task of examining the theorized causal pathways of human rights treaties mostly to case study research.

The substantive motivation of this chapter is a lack of quantitative understanding that remains in the literature regarding the mechanistic operation of human rights treaties. I thus conduct a causal mediation analysis of human rights treaties, using graphical causal models and machine learning methods to empirically investigate the causal pathways of three major United Nations (UN) human rights treaties—the International Covenant on Civil and Political Rights (ICCPR), the Convention on the Elimination of Discrimination against Women (CEDAW), and the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT). These three treaties are selected as the object of study because

they are the human rights treaties that are most commonly examined in the literature and, as a result, we tend to have better background knowledge about them than is the case with other UN treaties. My goal is to estimate the causal effect of each treaty and, more importantly, to determine how much of their total causal effect is transmitted through the causal mechanisms that have been suggested in the literature.

It should be noted that, despite an abundance of theories about the causal mechanisms of human rights treaties, the lack of any concrete quantification of these mechanisms remains mostly due to the fact that no empirical research in the substantive literature has employed a causal inference framework to inquire about causal pathways. Traditional approaches such as path analysis are often used to quantify the proportion of a treatment effect that goes through a mediating variable. Specifically, they rely on decomposing the regression coefficients of linear models into direct and indirect effect estimates [Judd and Kenny, 1981, Baron and Kenny, 1986]. However, these approaches do not generalize to nonlinear data-generating processes or allow a causal interpretation of their findings either [Shpitser, 2013, 1013–1017]. Recent studies of causal mediation analysis in areas other than human rights research overcome these limitations [Imai and Yamamoto, 2013], but they still require certain some functional form assumptions that could be relaxed further to make more robust causal effect estimates.

The substantive findings of the causal mediation analysis in this chapter have important implications. Estimating the natural direct effect and natural indirect effect of human rights treaties gives us more granular details about the impact of major international human rights institutions. It not only provides an estimate of the treaty effect, but also evaluates how much the UN treaties directly and indirectly influence government respect for human rights. The results, as a whole, improve our mechanistic understanding of human rights treaties and provide a policy-making basis for enhancing treaty effects and reducing human rights violations. If the direct effect of human rights treaties is substantial, for example, that would strengthen the case

for universal ratification of human rights treaties regardless of what causal mechanisms that may be needed to transmit treaty effects. If most of the treaty effect is mediated through intermediate mechanisms, the policy implications would be different. Human rights activism and the political mobilization of non-governmental organizations would be critical if a treaty ratification would have any effect at all on state behaviors. Moreover, if much of the causal effect of a human rights treaty is mediated through legislative constraints and the domestic courts, then human rights defenders and international actors should apply special scrutiny and be vigilant against government efforts to undermine the judicial system and the legislative institutions that can check and balance the executive power. Otherwise, international human right treaties, even if ratified, would be rendered ineffective.

Methodologically, my causal analysis also makes three innovative contributions. First, the causal mediation analysis in this chapter is conducted in the setting of panel data structure with repeated measures of the outcome. This kind of data structure is not commonly seen in the literature on mediation analysis.

Second, my causal analysis demonstrates the powerful advantage of causal graphs in assisting identification through multi-step adjustment. Instead of conditioning on a single set of covariates, this divide-and-conquer strategy, also referred to as piecemeal deconfounding [Pearl, 2014b], conditions on separate sets of variables to identify different components of the counterfactual distribution (the distribution of potential outcomes) on a sequential basis. This strategy helps increase identification power, especially given that in this case using a single adjustment set would fail the backdoor requirement for identification [Pearl, 2014b].

Finally, my causal mediation analysis demonstrates how recent advances in machine learning and complex predictive modeling can be leveraged and incorporated into a causal inference framework to produce effect estimates that are not only causally interpretable but also more robust than those generated by standard regression models in social sciences [Hofman et al., 2017]. It should be cautioned, however, that the substantive payoffs of this machine learning-based causal infer-

ence approach to mediation analysis is contingent on the observed data we have. The issues of missing data and measurement errors are beyond the scope of this analysis. Ultimately, the quality of the data imposes unavoidable constraints upon any data analyses.

In terms of the structure of this chapter, I begin by reviewing the literature and summarizing the background knowledge upon which I construct a structural causal model of the underlying causal process. This causal model encodes the causal assumptions and provides the framework within which I can define and formulate my causal queries about the mediation process [Pearl, 2009a, 2012].

I then represent the structural causal model in the form of a causal directed acyclic graph (DAG) to aid causal effect identification. I investigate a coterie of causal effects, including (a) the total causal effect, (b) the natural direct effect and the natural indirect effect in the context of multiple causally connected mediators, and (c) the controlled direct effect. This variety of causal quantities reflect the inherent trade-off between causal assumption plausibility and causal effect identifiability. Some causal quantities such as the natural direct effect and the natural indirect effect could be of greater substantive interest, but they tend to require stronger, more restrictive causal assumptions for identification. Other causal quantities such as the total effect and the controlled direct effect are estimable under relatively weaker assumptions, but they are only indirectly related to our substantive queries.

Finally, once we have established identification of these causal effects, I use two machine learning-based estimators, the weighting estimator and the substitution estimator, to compute robust causal effect estimates from observational data. In combination, these estimates provide a more comprehensive picture of the mechanistic operation of human rights treaties.

3.2 Theory

The causal pathways through which human rights treaties influence state behavior are often referred to in the human rights literature as mechanisms of influence. The existing literature has identified four major mechanisms involving institutional and legislative constraints, domestic judicial enforcement, mass mobilization, and international socialization. First, ratified human rights treaties could change the domestic agendas of participating countries. Treaty obligations modify the set of politically feasible policy options and even alter the domestic settings and institutional constraints within treaty member countries. The direction of this causal relationship is rather unambiguous. UN human rights treaties "are exogenous to most individual countries' policy agendas" [Simmons, 2009, 127]. They invariably require member states to enact administrative and legislative changes to implement treaty obligations. States parties to the Optional Protocol to the CAT (OPCAT), for example, are legally obligated to create a new institution in the form of a National Preventive Mechanism within three years of ratification. The treaty monitoring body of the OPCAT—the Subcommittee on Prevention of Torture—encourages all national preventive mechanisms to operate as an independent national institution that monitors government compliance and involves in government policy-making.

Obligations under human rights treaties also create a rallying point in the legislature to potentially constrain abuses by the executive. As Lupu [2015, 6] explains, legislative veto players, potentially including the opposition parties in the national legislature, can exploit information gathered from treaty monitoring activities "in conjunction with the activities of NGOs and the media". Legislative veto players can then take advantage of their legislative agenda-setting power and budget control power to expose and constrain repressive behaviors of the executive as are the case in the Knesset in Israel and the parliament of Zimbabwe Lupu [2015, 6]. Legislative constraints thus raise the cost of repression, thereby reducing violations and improving human rights practices of the government. In this causal mecha-

nism, human rights treaties, together with their associated monitoring procedures, could facilitate and enhance institutional and legislative constraints only when the veto players have divergent rights preferences from those of the executive as well as a sufficient amount of veto power to begin with.

Second, the influence and impact of a human rights treaty could also be felt in domestic judicial litigation [Simmons, 2009] and human rights prosecutions [Dancy and Sikkink, 2012]. A 2004 report by the International Law Association Committee on International Human Rights Law and Practice found that domestic courts in many countries such as Australia, Canada, the Czech Republic, Japan, and Nigeria, among others, have referred to findings by the treaty bodies of the ICCPR and the CAT when they issued decisions on domestic human rights cases [Scheinin, 2004]. Treaty obligations can provide and reinforce the legal basis of judicial litigation, enhance the effectiveness of domestic courts in constraining state abuses, and thereby improve human rights practices of member states. In other words, the causal effect of human rights treaties could be mediated by the effectiveness of the domestic judicial system [Powell and Staton, 2009, Crabtree and Fariss, 2015].

Third, in a prominent study of human rights treaties, Simmons [2009] argues that the most important causal mechanism of human rights treaties is through the social and political mobilization of ordinary citizens. Treaty ratifications inform and heighten people's awareness of their rights, increase the receptivity of governments to rights demands, and galvanize the population into social movements for rights protection. Rights mobilization around state obligations under international treaties often occurs through the action and advocacy of human rights nongovernmental organizations (NGOs) and civil society groups, particularly through their tactics of naming and shaming governments into compliance [Murdie and Davis, 2012]. Smith-Cannoy [2012], for example, offers case study evidence of how human rights NGOs in Hungary and Slovakia took advantage of the complaints procedure under two UN human rights treaties—the Convention on the Elimination of Discrimination against Women (CEDAW) and the Convention on the Elimination of

All Forms of Racial Discrimination (CERD)—to mobilize citizens to push for change in the government's human rights policies.

Finally, in addition to the causal mechanisms that operate within the domestic politics of treaty members, international emulation and socialization also represent a major causal pathway from a global treaty to the domestic behavior of participating states. After ratification, member states are subject to scrutiny by treaty monitoring bodies, among others. These are panels of independent experts whose job is to monitor state practices and hold regular dialogues with government officials to advise, persuade, and challenge them to better their governments' human rights records. While interacting with representatives of states parties on a regular basis, members of the treaty bodies use their legal expertise and human rights information to pressure abusive states to emulate rights-respective practices of other countries. Members of the treaty monitoring bodies may even change the hearts and minds of government officials about human rights norms by "contribut[ing] to community expectations of appropriate state behaviour under human rights treaty obligations" [Rodley, 2013, 639]. When norms and standards are internalized, they constitute new understandings of state interest and, as a result, change the behavior of state agents. A similar process also occurs at other international venues such as the UN's special procedures under the auspices of the Human Rights Council [Clark, 2013]. The essence of this causal pathway is that treaty ratification opens up new opportunities and increases incentives for member states to deepen their repeated interactions at the international level through which emulation [Goodman and Jinks, 2013] and persuasion [Keck and Sikkink, 1998] occur.

In summary, the existing literature has proposed at least four major causal pathways from treaty ratification to human rights outcome [Risse and Sikkink, 2013]. Systematic studies are still lacking, however, in terms of investigating and systematically quantifying the efficacy of these causal mechanisms. A causal mediation analysis could tell us if these mechanisms are dominant in terms of transmitting

the causal impact of human rights treaties or whether the direct effect is the most important part of an international human rights institution.

3.3 Empirical Analysis

This section estimates a variety of theoretical causal quantities that combine to illuminate the mechanistic operation of three major UN human rights treaties, including the ICCPR, the CEDAW, and the CAT. I begin by formulating a structural causal model of the data-generating process, which is a set of equations that formalize our background knowledge and causal assumptions [Pearl et al., 2016, 26]. Counterfactual expressions derived from this structural causal model are then used to define the causal effects of interest. Additionally, a graphical representation of the structural causal model in the form of a causal directed acyclic graph (DAG) will assist identification of these causal effects.

At the outset, however, a set of practical considerations relating to functional form assumptions, data structure, and causal assumptions critically inform our analysis. First, out of a concern about correct functional forms, I decided against using parametric statistical models and instead apply flexible machine learning methods for estimation. As standard practice, researchers regularly make parametric assumptions when modeling an outcome of interest. A linearity assumption, for example, is especially helpful for both identification and estimation in a causal mediation analysis [Daniel et al., 2011, VanderWeele, 2015]. In substantive terms, what a linearity assumption implies is that causal mediators such as political constraints, judicial effectiveness, and international socialization are a linear function of treaty ratification status and other covariates. Similarly, government respect for physical integrity rights or women's political empowerment is assumed to change linearly as a function of treaty membership, causal mediators, and potential confounding variables. Even if the linearity assumption could be relaxed in recent proposed estimators to allow for more flexibility [Imai et al., 2010, Imai and Yamamoto, 2013],

restrictive assumptions of no interactions and additivity of covariate effects are still required.

In human rights research, however, I am skeptical that the literature has accumulated enough concrete knowledge to specify a linear or any exact functional forms that characterize the true relationships between human rights treaties, their mediators, potential confounders, and human rights outcomes. No studies even attempts to justify the plausibility and accuracy of their functional form assumptions, except for occasional inclusion of an interaction term. If modeling assumptions such as linearity of parameters and additivity of covariate effects are inaccurate, effect estimates are mostly likely biased and the obtained inferences are easily invalid.

Second, I aim to make causal inference in the context of observational data with repeated measurements. Variables in this research are not measured in a single point in time as in a cross-sectional study. Rather, the treatments, mediators, outcomes, and confounders may vary over time. It should be noted that, by conventional measurement practice, we view treaty ratification status as a recurring yearly commitment even though once a country ratifies a human rights treaty, it is very unlikely to withdraw from the treaty. In the case of the CAT, the substantive rationale for viewing treaty ratification as a time-varying treatment variable is that Article 31 of the CAT provides a denunciation provision that allows states to exit the treaty. Legally speaking, therefore, any country members can exit from the treaty after one year of depositing their withdrawal notifications. While the ICCPR and the CEDAW do not have any similar denunciation clauses or provisions, that did not prevent some states from unsuccessfully attempting to withdraw from the ICCPR before [Tyagi, 2009]. As a result, I generally conceive treaty membership as an implicit annual ratification as opposed to a terminal event. While it is not clear how one estimates the causal effects of human rights treaties if treaty ratification is conceived as a terminal event since it has never been done before in the literature, identification and estimation would look very different and most likely would be potentially more complicated as well.

The panel data structure also differs from another more common data structure in mediation analysis where only the treatment regimens and mediators are time-dependent and repeatedly measured, but the target causal quantity involves the outcome measured in the final time period only. Using the causal inference framework and methods developed for this data structure [Blackwell, 2013, Bacak and Kennedy, 2015, VanderWeele and Tchetgen Tchetgen, 2017] would not only lose information about the outcome but also ignore the time-varying outcome's impact on subsequent measures of the treatment and the mediators and possibly undermine identification of the causal effects.

Third, until recently methods for identification and estimation in causal mediation analysis were mostly applicable in the context of a single mediator. In real-world politics, however, rarely does a causal process take place through a single causal mechanism. Recent methodological advances enable us to partially overcome this constraint of single mediator [VanderWeele and Vansteelandt, 2013]. The complex reality of multiple mediators, however, gives rise to a different identification problem—the treatment-induced mediator-outcome confounding. This problem, as later explained, could seriously complicate and even invalidate mediation analysis by rendering some of our causal effects non-identifiable [Pearl, 2014b, Tchetgen and VanderWeele, 2014, VanderWeele et al., 2014].

In summary, this set of considerations fundamentally shapes my analysis and determines how much we can learn about the underlying causal process through which human rights treaties influence state behavior. In an observational setting, the same probability distribution could be generated by different structural causal models [Peters et al., 2017, 10], which implies that one can never infer causality from observed data alone. Instead, to learn about the effect of interventions, I have to first specify a causal model with certain causal and usually untestable assumptions. I then define my causal quantities in terms of the properties of this structural causal model and establish their estimability. Finally, I propose a procedure for estimating these quantities using flexible machine learning techniques. In

more general terms, the methodological contribution of this chapter is to integrate a causal model-based approach that can represent and reason about a causal process (the structural causal model framework) and a function-based approach that can approximate a potentially complex function (machine learning) to estimate causal quantities that closely correspond to substantive theories.

3.3.1 Causal model formulation and effect definition

To formalize our background knowledge about the underlying causal process, I construct a causal model that describes how human rights outcome Y_t causally depends on contemporaneous treaty ratification status A_t both directly ($A_t \rightarrow Y_t$) and indirectly through the mediators ($A_t \rightarrow M_t \rightarrow Y_t$). I specify four mediators as suggested in the literature, including institutional and legislative constraints $M1$, judicial litigation $M2$, mass mobilization $M3$, and international socialization $M4$. I further include in my causal model time-invariant confounders X (legal origins, constitutional treaty ratification rules, and the types of electoral systems) and time-varying confounders W (population size, gross domestic product (GDP) per capita, levels of participation in international trade, democracy, the presence or absence of *de facto* multiple parties, regime durability, and involvement in militarized interstate disputes).

The variation of the four causal mediators—legislative constraints, judicial effectiveness, NGOs mobilization, and international socialization—are respectively measured by (a) a political constraints index, which measures the feasibility of policy change based on the veto power and alignment among government branches and degrees of preference heterogeneity within the legislative branch [Henisz, 2002]; (b) a judicial independence index measuring the independent power of the judiciary to constrain choices of the government [Linzer and Staton, 2015]; (c) a naming and shaming index, a composite measurement of reporting on human rights abuses by major media outlets, Amnesty International, and the UN Human Rights Council

(formerly the UN Commission on Human Rights) [Cole, 2015, 423]; and (d) the treaty commitment preference first coordinate based on state ratifications of 280 universal treaties across a large number of policy areas [Lupu, 2016], which is a good indicator of the extent to which countries participate, interact, and socialize internationally.

It should be reiterated and will be justified in more details later that specification of which variables to include and which variable set these variables belong to is entirely informed by existing studies of human rights treaties in the literature. Ultimately, one cannot empirically validate this causal structure specification without untestable assumptions because, as previously mentioned, the same probability distribution can be generated by different underlying causal structures. This causal structure specification, it is worth emphasizing, is separate from functional forms specification, the latter of which we are able to empirically address using machine learning.

The functional relationships between treaty ratification, intermediate variables, human rights outcome, and confounders are represented by a non-parametric structural causal model in Equation set 3.1. From this generative model, we observe a random sample of n country–year observations $O_n = (X, W_t, A_t, M1_t, \ldots, M4_t, Y_t) \sim P_O$ where P_O is the joint probability distribution. Table 3.1 lists the model variables and refers to studies in the literature that examine similar relationships between these variables. I make no assumptions about the functional forms f's and thus my structural causal model is non-parametric. In other words, time-invariant confounders X, time-varying confounders W, treaty ratification status A, intermediate variables M, and human rights outcome Y are causally connected according to the functions f's, but we are agnostic as to the forms of these functions.

$$
\begin{aligned}
X &= f_X(U_X) \\
W_t &= f_W(X, W_{t-1}, U_W) \\
A_t &= f_A(X, A_{t-1}, M1_{t-1}, M2_{t-1}, Y_{t-1}, W_t, U_A) \\
M1_t &= f_{M1}(X, M1_{t-1}, Y_{t-1}, W_t, A_t, U_{M1}) \\
M2_t &= f_{M2}(X, M2_{t-1}, Y_{t-1}, W_t, A_t, U_{M2}) \\
M3_t &= f_{M3}(X, M3_{t-1}, Y_{t-1}, W_t, A_t, U_{M3}) \\
M4_t &= f_{M4}(X, M4_{t-1}, Y_{t-1}, W_t, A_t, U_{M4}) \\
Y_t &= f_Y(X, Y_{t-1}, W_t, A_t, M1_t, M2_t, M3_t, M4_t, U_Y)
\end{aligned}
\tag{3.1}
$$

I do not assume any knowledge about the distribution of exogenous variables U's. However, all $U = \{U_X, U_W, U_A, U_{M1}, U_{M2}, U_{M3}, U_{M4}, U_Y\}$ are assumed to be jointly independent, suggesting that there are no hidden variables outside our model. It means, for example, no other variables will likely confound the relationship between treaty ratification and human rights outcome. This admittedly strong assumption is nonetheless critical and generally unavoidable for any observational studies that aim to make causal inference. While causal graphs make this causal assumption transparent, the only justification one can have in an observational setting is to rely on the literature and hope it has sufficiently identified the relevant variables.

Structural causal models are most effectively communicated in the form of causal DAGs. More importantly, based on the topology of a causal DAG, we can use identification methods, including the backdoor criterion [Pearl, 2009a] and the causal mediation formula [Pearl, 2012], to determine non-parametrically the estimability of the causal effects of interest. In a nutshell, a causal DAG [Koller and Friedman, 2009, Pearl, 2009a, Elwert, 2013, Drton and Maathuis, 2017] contains nodes denoting random variables and directed edges denoting one variable's direct causal influence on another. A path in a causal DAG is a sequence of directed arrows that connect one node to another regardless of the directions of the arrows. Any paths

Table 3.1.: Causal mediation model variables

Sets	*Time frame*	*Variables, References, and Data sources*
A	1976–2015	ICCPR ratification status (OHCHR).
	1981–2015	CEDAW ratification status (OHCHR).
	1987–2015	CAT ratification status (OHCHR).
M	1800–2016	$M1$: Institutional and legislative constraints [Lupu, 2015] measured by the political constraints index (Polcon iii) [Henisz, 2002].
	1948–2012	$M2$: Judiciary effectiveness [Powell and Staton, 2009, Conrad, 2013] measured by the judicial independence index [Linzer and Staton, 2015]
	1981–2007	$M3$: Political mobilization [Murdie and Davis, 2012, Simmons, 2009] measured by the naming and shaming index [Cole, 2015].
	1950–2008	$M4$: International socialization [Clark, 2013, Goodman and Jinks, 2013] measured by the treaty commitment preference coordinate [Lupu, 2016].
Y	1976–2015	$Y1$: Political Terror Scale [Gibney et al., 2016].
	1900–2015	$Y2$: Women's political empowerment index [Sundström et al., 2017].
	1981–2011	$Y3$: CIRI torture index [Cingranelli et al., 2013].
W	1966–2015	Population size [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators.
	1966–2015	Gross domestic product (GDP) per capita [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators.
	1960–2014	Participation in international trade [Hafner-Burton, 2013] measured by the World Bank Indicators.
	1946–2008	Regime type [Hathaway, 2007, Chapman and Chaudoin, 2013, Neumayer, 2007] measured by Cheibub et al. [2010].
	1946–2008	De facto multiple parties [Vreeland, 2008, Hollyer and Rosendorff, 2011] measured by Cheibub et al. [2010]
	1946–2008	Regime durability [Goodliffe and Hawkins, 2006] measured by age in current regime [Cheibub et al., 2010].
	1966–2013	Involvement in militarized interstate disputes [Chapman and Chaudoin, 2013]. measured by MID dataset [Themnér, 2014].
X		Legal origin [Mitchell et al., 2013] measured by the legal origins data [La Porta et al., 2008].
		Treaty ratification rule [Simmons, 2009] measured by the ratification rules dataset [Simmons, 2009].
		Electoral system [Cingranelli and Filippov, 2010] measured by the database of political institutions [Cruz and Scartascini, 2016].

between two nodes that consist of arrows all pointing to the same direction are directed or causal paths. Otherwise, they are non-causal paths. A DAG is also acyclic in the sense that there are no directed paths that connect a node to itself.

Figure 3.1 depicts a causal DAG that compactly represents our structural causal model in Equation set 3.1. It is assumed to exhibit the Markov property with respect to the causal graph in the sense that the causal graph encodes all the independences in the probability distribution [Peters et al., 2017, 101–102]. The implication is that, according to the causal graph, the values of a child node are strictly a function of only its parent nodes, which are nodes that emanate arrows into the child node, and the exogenous variables. Our causal DAG also has a dynamic structure with different time periods being represented by separate shaded blocks. This topology indicates a temporal order, according to which there are no directed arrows going from the future (the block on the right) back to the past (the block on the left). For clarity of presentation, I only represent two time periods with two causal mediators $M1$ and $M2$. A graphical model with all four mediators over a longer time span could be represented in a similar fashion.

As previously stated, any causal analysis to learn about the effect of intervention from observational data is crucially dependent upon the way a causal model is constructed. Therefore, the topology of a graphical causal model should encode our causal assumptions and sufficient background knowledge about the underlying data-generating process. First, I make the routine assumption in the context of repeated measures of time-varying variables that the immediate past influences the present. In notation, I include the set of directed arrows $W_{t-1} \rightarrow W_t$, $A_{t-1} \rightarrow A_t$, $M_{t-1} \rightarrow M_t$, and $Y_{t-1} \rightarrow Y_t$.

Second, I encode what is known in the literature as the selection effect argument, which is that state decisions to ratify and remain a party to an international treaty are based in part on their prior compliance records [Downs et al., 1996, von Stein, 2005]. I graphically represent this selection effects argument using the causal arrow $Y_{t-1} \rightarrow A_t$. Note that this is unrelated and orthogonal to any statistical arguments in connection to the estimation efficiency of using the lagged dependent variable.

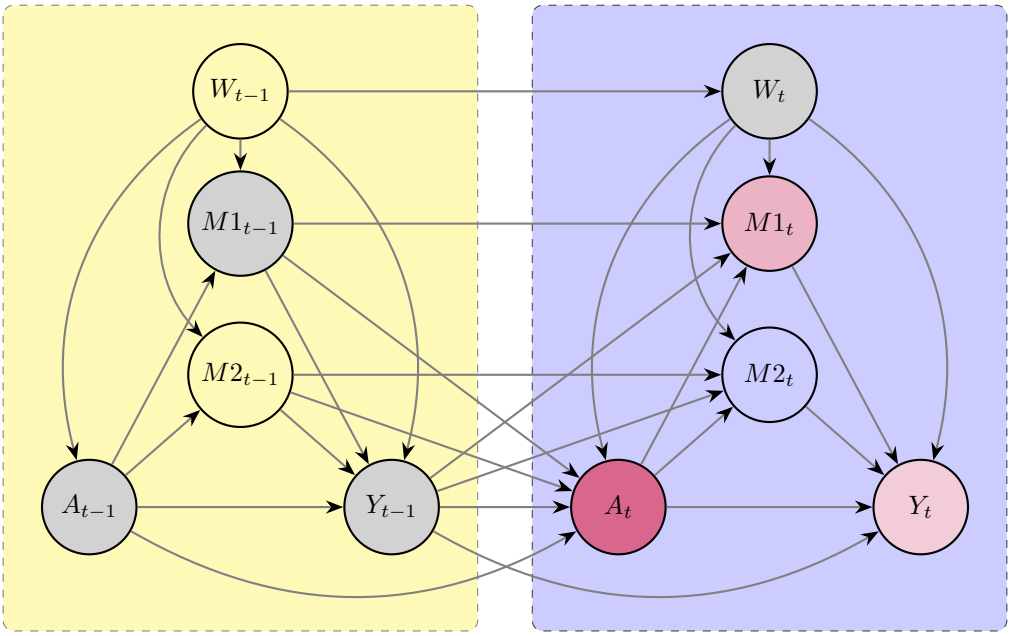


Fig. 3.1.: A DAG that represents the causal process involving time-varying confounders W, treaty ratification status A, two intermediate variables $M1$ and $M2$, and human rights outcome Y. Two color blocks denote two successive time periods. Time-invariant confounders X, which precede and potentially affect all other variables, are not represented. Exogenous variables U are assumed to be jointly independent and are not represented in the causal graph.

Third, the arrows $Y_{t-1} \rightarrow M1_t$ and $Y_{t-1} \rightarrow M2_t$ indicate the possibility that human rights outcome could affect the values of the intermediate variables in the following time period. Substantively, it means that torture, violations of civil liberties, and other repressive measures by the government may have the effect of weakening legislative constraints, undermining the court system, suppressing political mobilization, and provoking condemnations and social pressures by the domestic and international media, human rights NGOs, and the UN treaty bodies.

Fourth, I specify the directed arrows from the mediators to the treatment in the next time period $M1_{t-1} \rightarrow A_t$ and $M2_{t-1} \rightarrow A_t$. This allows for the possibility that intermediate variables could affect treaty ratification status in the following year. Substantively, the arrows suggest the scenarios where the executive ratifies and remains committed to human rights treaties in order to satisfy the demands of

the opposition parties [Vreeland, 2008], intimidate domestic opposition [Hollyer and Rosendorff, 2011], relieve pressures from social movements [Simmons, 2009], or engage in international emulation and respond to socialization [Finnemore and Sikkink, 1998]. Similarly, the number of legislative veto players and effective domestic judiciary might affect treaty ratification decisions as well [Conrad, 2013, Hill, 2016a].

Finally, other potential confounding factors, either time-invariant or time-varying, could influence both ratification decisions and human rights outcome. These potential confounding is represented by the directed arrows $W_t \rightarrow A_t$ and $W_t \rightarrow Y_t$. It should be noted that in a graphical causal model, a directed arrow indicates a possible, but not necessarily an actual, causal link. A missing arrow, on the other hand, is equivalent to ruling out any direct causality.

Given the graphical causal model with its encoded assumptions, we now translate mediation queries into various causal effects that are defined and expressed in terms of counterfactual quantities. They include the total effect, the natural direct effect, the natural indirect effect, and the controlled direct effect. First, the total effect (TE), often known as average causal effect, measures the average change in human rights outcome Y if we fix treaty ratification status A uniformly across all country–year observations. This average change, denoted by $TE = E\big[Y_{1,M_1} - Y_{0,M_0}\big]$, is the average difference between Y_1 and Y_0 where the subscript $a = \{1, 0\}$ denotes an intervention to fix treaty ratification status A at ratified ($a = 1$) and non-ratified ($a = 0$) and the average is taken over the entire sample of observations. In this formulation, the values of the mediators $\{M1, M2, M3, M4\}$ naturally change in response to treaty ratification status. Accordingly, the subscripted mediator M_1 denotes the value that a mediator will naturally obtain if we fix the treatment at $a = 1$ and M_0 similarly denotes the mediator value if the treatment value is set at $a = 0$. Computing the quantity $TE = E\big[Y_{1,M_1} - Y_{0,M_0}\big]$ helps answer our query about the average causal effect of ratifying a human rights treaty.

Second, the natural direct effect, denoted by $NDE = E\big[Y_{1,M_0} - Y_{0,M_0}\big]$, measures the average change in human rights outcome Y as a result of treaty ratification status when the values of all the mediators are set at M_0, that is, the values the mediators would obtain if the treaty was not ratified. The NDE quantity represents the portion of the total effect that is transmitted directly to the outcome without any of the four causal mechanisms we previously described. By estimating the NDE, we could learn how much a human rights treaty can change state behavior in the absence of the specified mechanisms of influence.

Third, the natural indirect effect, denoted by $NIE = E\big[Y_{1,M_1} - Y_{1,M_0}\big]$, measures the average change in human rights outcome Y when the treaty ratification status is fixed at $a = 1$ (ratified) across the board, but the mediators now alternate between the values they would obtain for each observation under $a = 1$ (ratified) and $a = 0$ (non-ratified). The NIE therefore quantifies only the portion of treaty effect that is transmitted through the mechanism under inquiry, which is said to best "capture our notion of mediation" [VanderWeele et al., 2014, 301]. Substantively, estimates of the NIE quantify the impact of incorporating a human rights treaty into domestic judicial litigation, the ability of legislative veto players to use treaty obligations to constrain human rights violations, and the capacity of human rights NGOs and international actors to use treaty obligations and information about state compliance to pressure and ultimately change a government's human rights practice.

The TE could be unpacked into a sum of the NDE and the NIE according to Equation 3.2. Equivalently, causal queries involving the TE, the NDE, and the NIE could be expressed in terms of three counterfactual quantities: $E\big[Y_{1,M_1}\big]$, $E\big[Y_{1,M_0}\big]$, and $E\big[Y_{0,M_0}\big]$.[1]

$$\begin{aligned} \text{TE} &= E\big[Y_{1,M_1} - Y_{0,M_0}\big] \\ &= E\big[Y_{1,M_1} - Y_{1,M_0}\big] + E\big[Y_{1,M_0} - Y_{0,M_0}\big] \\ &= \text{NIE} + \text{NDE} \end{aligned} \tag{3.2}$$

[1]The causal quantity $E\big[Y_{0,M_1} - Y_{0,M_0}\big]$ also corresponds to similar substantive queries about natural indirect effect. However, it is less elegant since we will be unable to neatly unpack the total effect into a sum of direct and indirect effects.

It is worth noting that when we compute the counterfactual outcome value Y_{1,M_0}, we are nesting the mediator value that corresponds to one treatment intervention (M_0 is the value of the mediator when the treatment is fixed at $A = 0$) under a different treatment intervention (when the treatment is fixed at $A = 1$). Since Y_1 and M_0 only occur under different worlds in which the treatment values differ, the quantity Y_{1,M_0} is often referred to as a cross-world counterfactual and is generally unobservable even under experimental conditions.

Fourth, another causal effect, though not directly central to our mediation analysis but could nonetheless offer additional insights into the mechanistic operation of human rights treaties, is the controlled direct effect. It is denoted by $CDE(m) = E[Y_{1,m} - Y_{0,m}]$. The CDE measures the average change in human rights outcome Y when countries become states parties to a human rights treaty but the values of the mediator such as judicial independence and legislative constraints are fixed at a specific value $M = m$ across the entire population. CDE estimates can certainly vary across different fixed mediator values. The NDE can also be summarized as the weighted average of the CDE [Pearl et al., 2016, 123] with the weighting proportional to the distribution of the mediators under non-ratification [Petersen and van der Laan, 2008, 24]. In my estimation using the demediation function [Acharya et al., 2016], all mediator values are set at their empirically lowest value $M = 0$. This is different from $M = M_0$, which is the mediator value under treaty non-ratification. In substantive terms, the CDE quantifies the direct impact of human rights treaties under rather unfavorable conditions when a mediator is set at its observed lowest value.

3.3.2 Causal identification

Total effect

In a non-experimental setting, the question of causal identification arises when we want to know whether and under which conditions a causal effect can be

uniquely computed from the joint probability distribution of the observed variables, which is assumed to be compatible with a graphical causal model. For the causal model in Figure 3.1, identification of the total effect of A_t on Y_t via adjustment requires a conditioning set Z_{TE} that satisfies the following backdoor criterion [Pearl et al., 2016, 61–64]:

(a) Z_{TE} leaves open all directed paths from A_t to Y_t;

(b) Z_{TE} blocks all backdoor paths from A_t to Y_t, that is, all paths that have an arrow entering A_t; and

(c) Z_{TE} does not open any spurious paths by including a collider (a node on a path from A_t to Y_t and has two arrows entering it) or a descendant of a collider (a node connected to a collider through a directed path).

The adjustment set $Z_{TE} = \{X, A_{t-1}, M1_{t-1}, M2_{t-1}, Y_{t-1}, W_t\}$ in Figure 3.1 is sufficient to identify the total effect of A_t on Y_t. It permits the transition from the *do*-operator in Equations 3.3 and 3.4 below, which denotes an intervention to fix treaty ratification status at $A_t = 1$ (ratified) and $A_t = 0$ (non-ratified), to a function of the observed probabilities. Once I derive and condition on the sufficient adjustment set Z_{TE}, I effectively break the non-causal paths between the treatment A_t and the outcome Y_t and render these two nodes conditionally independent from each other [Pearl et al., 2016, 46–48]. Any remaining association between them will be evidence of a causal relationship. To compute $TE = E\big[Y_{1,M_1}\big] - E\big[Y_{0,M_0}\big]$, it should be noted, we do not condition on any of the intermediate variables $\{M1, M2, M3, M4\}$ in violation of rule (a) of the backdoor criterion. Nor should we concern with any possible interactions among them.

$$\begin{aligned} E\big[Y_{1,M_1}\big] &= E\big[Y_t|do(A_t = 1)\big] \\ &= E_{Z_{TE}}\big[Y_t|A_t = 1, Z_{TE} = z\big]P(Z_{TE} = z) \end{aligned} \tag{3.3}$$

$$
\begin{aligned}
E\big[Y_{0,M_0}\big] &= E\big[Y_t|do(A_t=0)\big] \\
&= E_{Z_{TE}}\big[Y_t|A_t=0, Z_{TE}=z\big]P(Z_{TE}=z)
\end{aligned}
\tag{3.4}
$$

Natural direct effect and natural indirect effect

Estimating the natural direct and indirect effects requires computing the cross-world counterfactual quantity $E\big[Y_{1,M_0}\big]$ in addition to the counterfactuals $E\big[Y_{1,M_1}\big]$ and $E\big[Y_{0,M_0}\big]$ that form the computation of the total effect. Natural effects identification via adjustment has to satisfy the following set of conditions [Pearl et al., 2016, 122]:

(d) The first adjustment set $Z_{NE1} = \{X, Y_{t-1}, W_t\}$ does not include any descendants of A_t;

(e) The adjustment set Z_{NE1} blocks all backdoor paths from M_t to Y_t after removing the arrows $A_t \rightarrow M_t$ and $A_t \rightarrow Y_t$;

(f) Conditioning on the first adjustment set Z_{NE1}, the effect of A_t on M_t is identifiable through the second adjustment set $Z_{NE2} = \{A_{t-1}, M_{t-1}\}$ that blocks all backdoor paths from A_t to M_t ; and

(g) Conditioning on the first adjustment set Z_{NE1}, the joint effect of $\{A_t, M_t\}$ on Y_t is identifiable, which requires any confounders of the effect of M_t on Y_t not be affected by A_t.

Assuming the observed data are sample observations randomly drawn from the causal process in Figure 3.1, a combination of two separate adjustment sets Z_{NE1} and Z_{NE2} will be able to identify the natural effects. This is an example of the divide-and-conquer strategy of using two different adjustment sets on a piecemeal basis, as opposed to a single set, to satisfy the deconfounding requirements for identification. It highlights the greater facility of causal graphs than algebraic expressions and manipulations in the potential outcomes framework in terms of assisting

identification [Pearl, 2014b]. Equation 3.5, which relies on the mediation formula that Pearl [2012] develops, shows us how to compute the counterfactual $E\big[Y_{1,M_0}\big]$ as a function of the observational conditional probability.

$$E\big[Y_{1,M_0}\big] = \sum_{M} \sum_{Z_{NE1}} E\big[Y|A=1, M=m, Z_{NE1}=z\big] P(Z_{NE1}=z) \\ \sum_{Z_{NE2}} P(M=m|A=0, Z_{NE2}=z) P(Z_{NE2}=z) \tag{3.5}$$

Treatment-induced mediator-outcome confounding

A complication arises, however, with respect to criterion (g) for identification of the natural effects. This criterion implies conditional independence among the mediators, which we would have to reject based on the domain knowledge. In substantive terms, one could argue that human rights NGOs and the civil society not only galvanize and mobilize citizens against government abuses ($M3$). They also advocate for and bring lawsuits involving internationally protected human rights before domestic courts. This scenario is denoted by the additional causal arrow $M3 \rightarrow M2$. Non-governmental and civil society organizations may also pressure legislators to enact domestic laws and policies in conformity with treaty obligations and to constrain government repression by the executive. This implies the causal arrow $M3 \rightarrow M1$. Finally, independent and legal experts on treaty bodies routinely employ evidence and information about government abuses that human rights NGOs have gathered on the ground and in the field to confront and persuade government officials at the UN and other international forums. This suggests another causal link $M3 \rightarrow M4$. In other words, NGOs mobilization, which is influenced by treaty ratification status A, may very well confound the causal relationships between each of the other three mediators and the outcome. This

treatment-induced ($A_t \rightarrow M2_t$) mediator-outcome confounding is represented by the fork $M1_t \leftarrow M2_t \rightarrow Y_t$ in the causal graph in Figure 3.2.

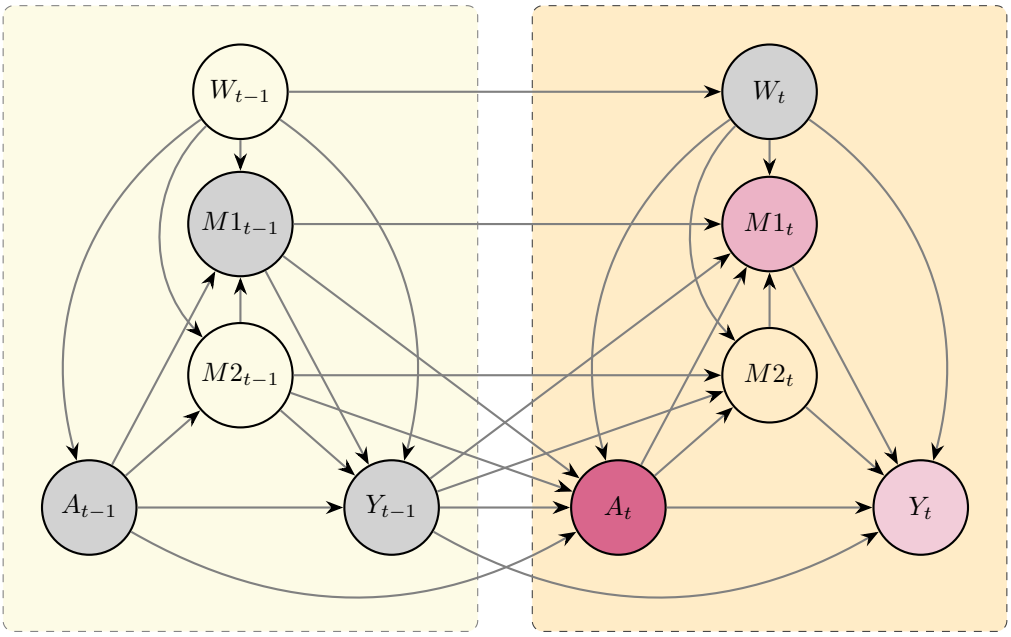


Fig. 3.2.: A causal DAG that represents the causal process with treatment-induced ($A_t \rightarrow M2_t$) mediator-outcome confounding ($M1_t \leftarrow M2_t \rightarrow Y_t$). The DAG includes time-varying confounders W, human rights treaty ratification status A, two mediators $M1$ and $M2$, and human rights outcome Y with two shaded blocks indicating two successive time periods. Time-invariant confounders X, which precede and potentially affect all other variables, are not represented. All exogenous variables U's are assumed to be jointly independent and are not represented in the causal graph.

One might be prompted to then include $M2_t$ in the adjustment set Z_{NE1} to satisfy rule (e) of the backdoor requirement that Z_{NE1} should block all backdoor paths from $M1_t$ to Y_t. Including $M2_t$ in the adjustment set Z_{NE1}, however, would violate rule (d) that $M2_t$ *not* be conditioned on because it is a descendant of the treatment A_t. As a result, when we take into account the more realistic treatment-induced direct causal dependence among the mediators, the NDE and the NIE generally become non-identifiable without strong parametric assumptions [Pearl, 2014b, 471–472].

The natural effects could still be identified under some special conditions, including (a) linear functional forms that characterize the relationships among the variables [Imai and Yamamoto, 2013]; (b) monotonicity of treatment effect A_t on a binary confounding mediator $M2_t$ [Tchetgen and VanderWeele, 2014, 285–286]; or (c) no additive interaction of the effects of multiple mediators on the outcome [Tchetgen and VanderWeele, 2014, 286–287]. Nevertheless, these conditions are still exceedingly restrictive. They assume concrete knowledge about functional forms that is usually unavailable or highly suspect. They also are not applicable in this case because all four mediators are measured as continuous variables.

A more realistic solution that I adopt here to circumvent the problem of direct causal influence among the mediators is to jointly consider all mediators as though they were a single intermediate variable. The resulting NDE estimate will be the portion of treaty effect that is transmitted through none of the intermediate variables whereas the NIE is the portion transmitted through any or all of the mediators [VanderWeele et al., 2014, 302–303]. The downside of this solution is that I cannot tease out the exact portion of treaty effect that is transmitted through each individual mediator. Fundamentally, this situation concretely illustrates an inherent trade-off in any causal analysis between causal effect identifiability on the one hand and causal assumption plausibility on the other hand. Identifiability is easier to establish if one is willing to make stronger and less plausible assumptions about the absence of certain causal links among variables. Conversely, the cold hard truth is that the more likely that more variables are causally connected, the less likely we are able to identify and estimate their independent causal effects.

Controlled direct effect

While the natural effects generally are not identifiable without overly restrictive assumptions, we could nevertheless estimate the CDE. The upside is that the CDE is estimable under weaker assumptions than the natural effects. All we need is

two adjustment sets that could respectively (a) block all backdoor paths from each mediator M_t to the outcome Y_t and (b) block all backdoor paths from the treatment A_t to the outcome Y_t after removing all arrows entering the one mediator M_t that is under consideration [Pearl et al., 2016, 77].

Note that the two adjustment sets for CDE identification in this case do not have to coincide. The causal graph in Figure 3.2 reveals that the first set $Z_{CDE1} = \{X, Y_{t-1}, W_t, A_t, M2_t\}$ blocks all backdoor paths from $M1_t$ to Y_t while the second set $Z_{CDE2} = \{X, Y_{t-1}, M2_{t-1}, W_t\}$ blocks all backdoor paths from A_t to Y_t after removing all arrows entering $M1_t$. The ability to use two separate adjustment sets for identification of a single causal effect makes it possible to use the sequential g-estimator to estimate the CDE of each confounded mediators. It is worth noting that a single adjustment set that encompasses both Z_{CDE1} and Z_{CDE2} would fail the backdoor requirement because of its inclusion of $M2_t$. This again highlights the benefits of causal graphs and the greater flexibility of a divide-and-conquer adjustment strategy for identification.

The downside in estimating the CDE is that this causal quantity does not exactly answer our original query about the portion of treaty effect that each mediator transmits. However, as Acharya et al. [2016, 6] observe, a significantly non-zero CDE estimate implies that some of the treaty effect is *not* due to the involved causal mechanism. That means a CDE estimate that significantly differs from zero indicates that a non-negligible portion of the treaty effects traverses through *other* causal pathways apart from the controlled mediator. To be fair, though, CDE estimation probably only provides confirmation rather than novel insights into human rights treaty effects given that it is highly likely that treaty effects are always partially transmitted through our theoretically identified mediators.

3.3.3 Machine learning-based estimation

To compute various causal effects of human rights treaties as described above, we use a separate dataset for each of the three treaties. The datasets for estimating the causal quantities relating to the ICCPR and the CEDAW have the same temporal coverage from 1981 to 2008. The dataset for estimating the causal quantities relating to the CAT has the temporal coverage from 1987 when the CAT went into effect until 2008. Since the adjustment sets for identification tend to include more variables to reasonably satisfy more demanding conditions for natural effects identification, these relatively short time frames are selected to avoid a high, unsustainable level of missing data. Appendix B.1 gives a more detailed description of how variables are measured as well as the data sources. Appendix B.2 provides the summary statistics. To handle missing data, I implement imputation using the Amelia II program [Honaker et al., 2011]. Information about the imputation process is provided in Appendix B.3.

Since I do not know the true functions f's in my structural causal model (Equation set 3.1), I learn these functions inductively rather than adopt *a priori* a specific functional form. This is where my analysis differs from previous approaches in the causal mediation literature [VanderWeele, 2009, Vansteelandt, 2009, Imai et al., 2011, Imai and Yamamoto, 2013, VanderWeele, 2015, Acharya et al., 2016]. Specifically, I use machine learning to work out the functional forms that minimize the empirical risks (the loss function). To demonstrate the applicability and advantage of this machine learning-based approach to our specific domain, I conduct a predictive analysis by fitting multiple algorithms to predict three different human rights outcomes Y_t measured by the Political Terror Scale, the Women's Political Empowerment Index, and the CIRI Torture Index. The goal is to compare the performance of these models in terms of learning the three underlying generative functions $Y_t = f_Y(\cdot)$ in Equation set 3.1. The performance metrics is to minimize the empirical risk, that is, the average cross-validated mean-squared error (MSE). The

lower the cross-validated MSE, the better the algorithms in approximating the true generative functions. It should be emphasized, however, that for my subsequent machine learning-based causal effect estimation, feature selection is based on the identifiability results, that is, relevant predictors are selected based on whether they are included in the adjustment sets I previously derived.

Cross-validation helps make sure the models generalize well in terms of predicting unknown outcome values. Four-fold cross-validated MSE are computed for each algorithm that predicts human rights outcome Y_t as a function of the set of covariates $\{X, Y_{t-1}, W_t, A_t, M1_t, M2_t, M3_t, M4_t\}$, which includes time-invariant confounders X, the lagged outcome value Y_{t-1}, time-varying confounders W_t, treaty ratification status A_t, and all four intermediate variables M_t. According to the causal model in Figure 3.2, this set of predictors are assumed to exert direct causal influence on the outcome Y_t. The empirical risk function to minimize is $E\big[Y_t - Q(\cdot)\big]^2$ and the true generative functions are $Y_t = Q(X, Y_{t-1}, W_t, A_t, M1_t, M2_t, M3_t, M4_t)$.

Table 3.2 describes the algorithms I use in this predictive analysis with all continuous variables standardized into a bounded 0–1 range for learning stability. Other data transformations are specified in the Appendix B.1. The list of algorithms covers a diverse array of algorithms that make different tradeoffs between interpretability and flexibility and between bias and variance [James et al., 2013]. They range from ordinary least squares regression and regularized linear regression [Tibshirani, 1996] to generalized additive models [Hastie and Tibshirani, 1990] and ensemble trees-based non-linear algorithms such as random forest [Breiman, 2001b] and boosting [Friedman, 2001].

A particularly powerful algorithm on the high-end of the flexibility spectrum is extreme gradient boosting [Chen and He, 2015, Chen and Guestrin, 2016], which is a faster implementation of gradient boosting machine [Friedman, 2001, Natekin and Knoll, 2013]. Extreme gradient boosting (XGBoost) is a non-parametric ensemble method and its tree-based nature allows it to capture non-linear, interactive dynamics among a large number of predictors. To enhance the performance of

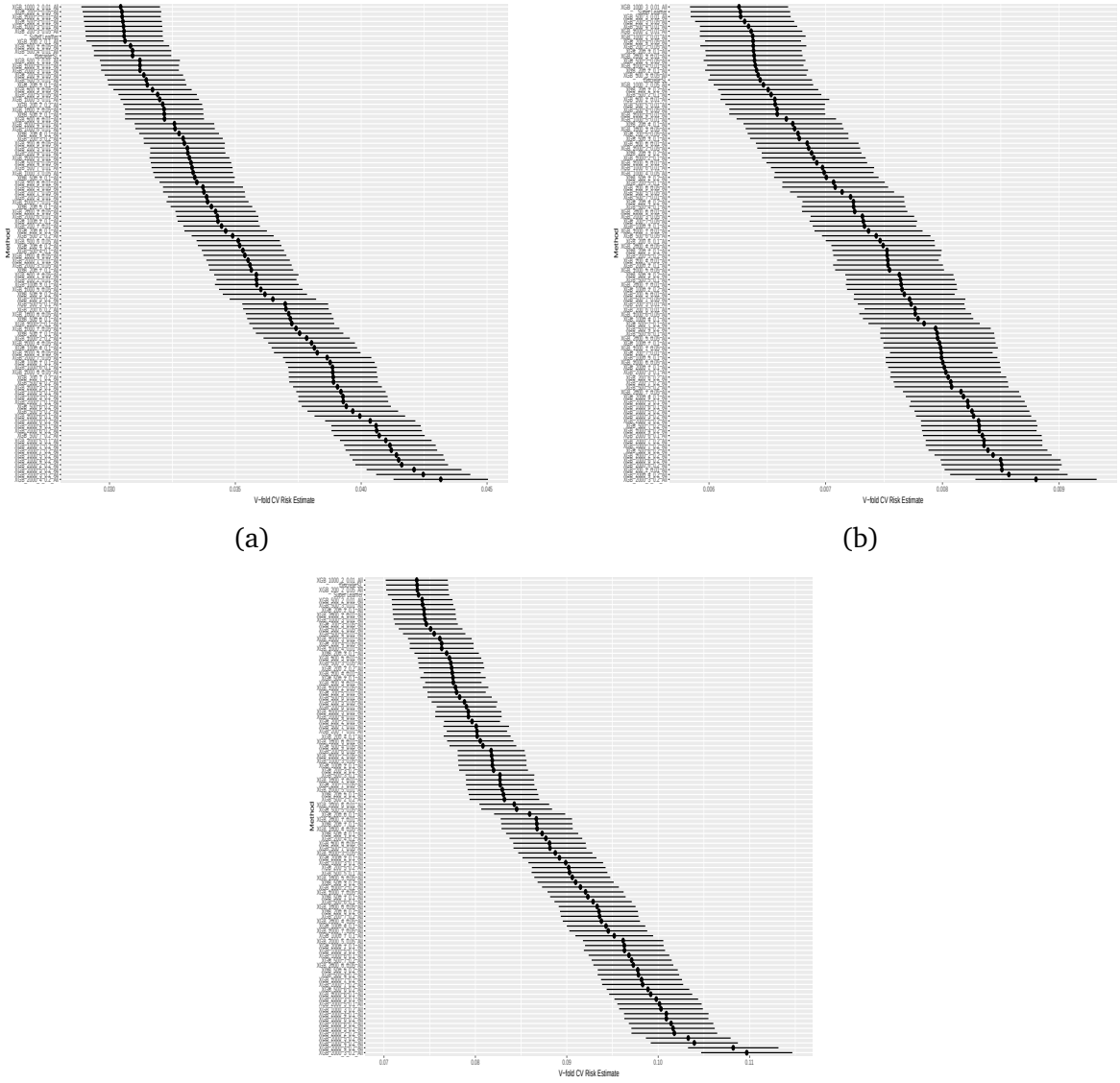
XGBoost, I use a combination of cross-validation and grid search to fine-tune its hyper-parameters to each of the three datasets in Figure 3.3 and select the best configurations of varying learning rates, tree depths, and numbers of trees.

Table 3.2.: Algorithms used in Super Learner-based predictive analysis

Algorithm	*Description*
glm	Main-term linear model
glmnet	Cross-validated penalized linear regression with lasso penalty $\sum_{j=1}^{p}\|\beta_j\|$
gam	Generalized additive model (degree of polynomials = 2)
polymars	Multivariate adaptive polynomial spline regression
randomForest	Random forest (ntree = 1,000)
xgboost (default)	Extreme gradient boosting (default hyper-parameters)
xgboost (tuned)	Extreme gradient boosting (fine-tuned hyper-parameters)

Figure 3.4 reports the cross-validated MSEs of all predictive algorithms, including the top XGBoost algorithm, for each of the three datasets. This predictive analysis casts doubt on the sufficiency of OLS linear regression models given that they yield a meager performance in predicting human rights outcomes. More flexible models tend to yield a better performance. In all three cases, the fine-tuned XGBoost algorithm consistently scores the best predictive performance.

This predictive analysis also underscores the advantage of incorporating the Super Learner ensemble technique for robust effect estimation [van der Laan et al., 2007, Polley and van der Laan, 2010]. Super Learner employs a user-selected library of algorithms, each of which is weighted by its relative cross-validated predictive performance. A weighted combination of these algorithms is then used to produce a hybrid prediction function that performs as well as, and usually even better than, the best algorithm in the library. I select linear regression, regularized linear regression with lasso, generalized additive model, and XGBoost to create the Super Learner-based variants of both (a) the weighting estimator that VanderWeele and Vansteelandt [2013] propose for total effect and natural effects estimation and (b) the g-estimator that Vansteelandt [2009] and Acharya et al. [2016] propose for controlled direct effect estimation. These four parametric, semi-parametric, and



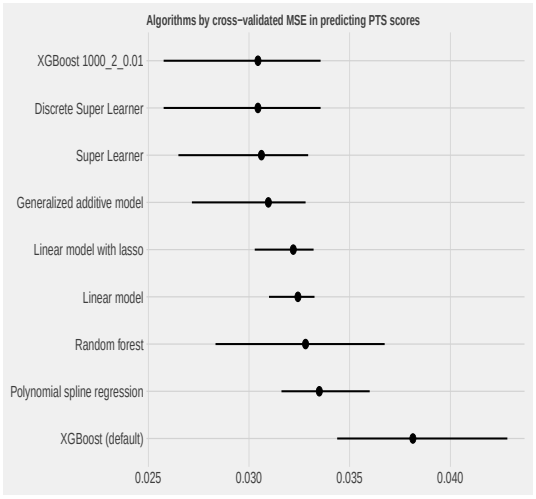
(a) (b)

(c)

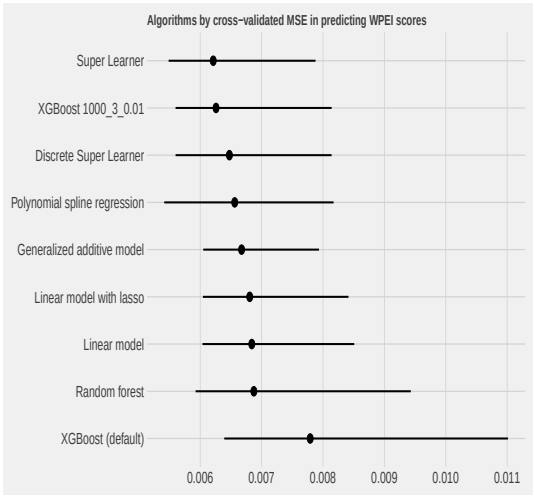
Fig. 3.3.: Minimum, maximum, and average values of 4-fold cross-validated MSE of XGBoost algorithms in predicting (a) Political Terror Scale score, (b) Women's Political Empowerment score, and (c) CIRI Torture score. Cross-validation helps prevent overfitting and provides a more accurate assessment of the abilities of these different XGBoost configurations in predicting the outcomes. The smaller the average MSE, the better that XGBoost configuration is presumably able to approximate the data-generating function that generates the human rights outcome values.

non-parametric algorithms have different degrees of complexity and make different bias-variance tradeoffs. Note that if the true generative functions happen to be lin-

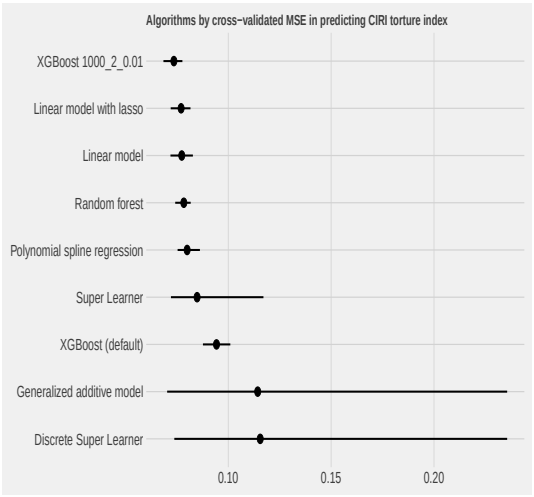
ear, Super Learner will be able to recover that as well and we therefore do not have to decide and justify whether a linear functional form assumption is appropriate or not.



(a)



(b)



(c)

Fig. 3.4.: Minimum, maximum, and mean values of 4-fold cross-validated MSE of algorithms in predicting (a) Political Terror Scale score, (b) Women's Political Empowerment score, and (c) CIRI Torture score. All three measures are rescaled into a bounded 0–1 range. Rescaling the outcome measurements changes the absolute values of the MSE, but does not affect the relative rankings of the algorithms in terms of predictive performance.

The g-estimator or substitution estimator [Robins, 1986, Robins et al., 1999] computes the causal effect of a treatment $\tau = E\big[Y|do(A = 1)\big] - E\big[Y|do(A = 0)\big]$ by using the estimate $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\big[Q_n(1, Z) - Q_n(0, Z)\big)\big]$ where Q_n is the predictive outcome model and Z is a sufficient adjustment set. For the g-estimator, the key to obtaining consistent effect estimates is to fit a correctly specified outcome model Q_n that approximates the (unknown) data generating mechanism. The inverse probability of treatment-weighted (IPTW) estimator (the weighting estimator) for natural effects estimation has the crucial benefit of not having to model the conditional density of multiple continuous mediators. However, its consistency depends on (i) the ability of a treatment model (that computes the inverse probability of treatment weights) to approximate the true treatment mechanism and (ii) a correctly specified outcome model.

Both the g-estimator and the weighting estimator require correctly specified models of various different generating functions. The ensemble machine learning technique Super Learner offers an effective solution to meet this requirement [Kreif et al., 2015, Pirracchio et al., 2015, Samii et al., 2016]. Uncertainties around point estimates are quantified using the nonparametric bootstrap method [Efron and Tibshirani, 1994] with $B = 500$. Similar to Daniel et al. [2011, 491] and suggested by Tsiatis [2007, 362–371], I combine nonparametric bootstrap with single stochastic imputation to make efficient and valid inference, obtaining distribution-free confidence intervals from the 2.5% and 97.5% quantiles of the bootstrap distribution. For each dataset that includes a different human rights treaty and its corresponding outcome measure (the ICCPR and the Political Terror Scale score, the CEDAW and the Women's Political Empowerment index, and the CAT and the CIRI Torture index), I implement the following machine learning-based estimation procedure.

1. Total effect estimation

   (a) Fit Super Learner treatment prediction model $g(Z_{TE}) = E\big[A_t|Z_{TE}\big]$ where the predictors are the adjustment set $Z_{TE} = \{X, A_{t-1}, M_{t-1}, Y_{t-1}, W_t\}$.

(b) Use the treatment prediction model $g(Z_{TE})$ to generate inverse-probability-of-treatment weights $ipw_0 = 1/P(A_t = 0|Z_{TE})$ for observations with observed treatment value $A = 0$ and $ipw_1 = 1/P(A_t = 1|Z_{TE})$ for observations with observed treatment value $A = 1$. One could bound the predicted probabilities above the threshold of 0.01 to constrain excessive variability of the estimates if necessary [Cole and Hernán, 2008].

(c) Compute the counterfactuals $E\big[Y_{1,M_1}\big] = E\big[Y * ipw_1\ A = 1\big]$ as the weighted outcome mean among observations with observed treatment value $A = 1$ and $E\big[Y_{0,M_0}\big] = E\big[Y * ipw_0|A = 0\big]$ as the weighted outcome mean among observations with observed treatment value $A = 0$.

(d) Compute the total effect $TE = E\big[Y_{1,M_1}\big] - E\big[Y_{0,M_0}\big]$.

2. Natural effects estimation with all mediators considered jointly

(a) Fit Super Learner outcome prediction model $Q_{NE1} = E\big[Y_t|A_t, M_t, Z_{NE1}\big]$ with the adjustment set $Z_{NE1} = \{X, Y_{t-1}, W_t\}$.

(b) Subset the full sample and use only observations with observed treatment value $A = 0$. This will obviate the more difficult task of modeling the joint density of multiple continuous mediators under $A = 0$. Substitute $A = 1$ into Q_{NE1} and use the observed values of all four mediators M_t and variables in the set Z_{NE1} from the subsetted sample to generate predicted outcome values $\widehat{Y}_t$.

(c) Compute $E\big[Y_{1,M_0}\big] = E\big[\widehat{Y}_t * ipw_0\big]$ as the weighted mean of predicted outcome values $\widehat{Y}_t$. Note that I use two adjustment sets sequentially to identify the joint natural effects: (1) $Z_{NE1} = \{X, Y_{t-1}, W_t\}$ in the outcome model Q_{NE1} blocks all backdoor paths from all four mediators M_t (jointly considered) to Y_t; and (2) conditioning on Z_{NE1}, the set $Z_{NE2} = \{A_{t-1}, M_{t-1}\}$ in computing the weights ipw_0 in step (1b) above blocks the backdoor paths from A_t to Y_t and from A_t to all M_t.

(d) Compute causal effect estimates $NDE_{\text{joint}} = E\big[Y_{1,M_0}\big] - E\big[Y_{0,M_0}\big]$ and $NIE_{\text{joint}} = E\big[Y_{1,M_1}\big] - E\big[Y_{1,M_0}\big]$.

3. Controlled direct effect estimation

(a) For each of the three confounded mediators $M1$ (political constraints, judicial independence, and international socialization):

i. Fit Super Learner demediation model $D_{M1} = E\big[Y_t|M1_t, Z_{CDE1}\big]$ where the set $Z_{CDE1} = \{X, Y_{t-1}, W_t, A_t, M2_t\}$ blocks all backdoor paths from $M1_t$ to Y_t.

ii. Substitute $M1_t = 1$ and $M1_t = 0$, which have been already standardized into a bounded 0–1 range, into D_{M1} to compute the maximum effect of $M1_t$ on Y_t, using $D(m1) = E\big[Y_t|M1_t = 1, Z_{CDE1}\big] - E\big[Y_t|M1_t = 0, Z_{CDE1}\big]$.

iii. Compute the demediated outcome values $\widetilde{Y}_t = Y_t - D(m1) * M1_t$, that is, subtracting the effect of the confounded mediator $M1_t$ from the outcome Y_t.

iv. Fit Super Learner prediction model of the demediated outcome $\widetilde{Q}(A_t, Z_{CDE2}) = E\big[\widetilde{Y}_t|A_t, Z_{CDE2}\big]$ where $Z_{CDE2} = \{X, Y_{t-1}, M2_{t-1}, W_t\}$ blocks all backdoor paths from A_t to Y_t after removing all arrows entering $M1_t$. Sequential adjustment using separate sets Z_{CDE1} and Z_{CDE2} satisfies the backdoor requirement for CDE effect identification.

v. Substitute $A = 1$ and $A = 0$ into the demediated model $\widetilde{Q}(A_t, Z_{CDE2})$ to compute $CDE(m1_t = 0) = E\big[\widetilde{Y}_t|A = 1, Z_{CDE2}\big] - E\big[\widetilde{Y}_t|A = 0, Z_{CDE2}\big]$.

(b) For the confounding mediator $M2_t$ (mass mobilization), repeat step 3(a) but use two sequential adjustment sets $Z_{CDE1_{M2}} = \{X, Y_{t-1}, M2_{t-1}, W_t, A_t\}$ and $Z_{CDE2_{M2}} = \{X, Y_{t-1}, M1_{t-1}, W_t\}$.

3.3.4 Results and interpretation

Estimates of various causal quantities relating to the three UN human rights treaties are reported in Table 3.3 together with their nonparametric bootstrap-based 95% confidence intervals. They provide answers to several of our causal inquiries. First, do human rights treaties reduce government abuses and protect and promote individual human rights? In contrast to many previous findings in the literature, my answer is affirmative across all three human rights treaties although the causal effect magnitudes vary. Participating in the ICCPR reduces state violations of physical integrity rights by 13.6 percentage points or, equivalently, about 0.54 points in the 5-level Political Terror Scale. Participating in the CAT leads to a more modest decrease of government's torture practice by about 7.7 percentage points—roughly 0.23 points in the 3-level scale of the CIRI Torture Index. For the CEDAW, the average causal effect is significantly more substantial, raising women's political empowerment by 22 percentage points measured by an aggregate index of women's civil liberties and political participation.

Second, how much do human rights treaties influence state behavior directly and indirectly through causal mediators? In the case of the ICCPR and the CAT, there is something concerning about their direct causal effects. Participating in these two treaties leads to *more* torture and violations of physical integrity rights. If all four mediators do not change their values in response to treaty ratification, being a member of these treaties exacerbates human rights practices by 0.8 and 3.4 percentage points, respectively. However, both the ICCPR and the CAT exert a positive indirect causal influence on state behavior that is both statistically significant and substantively larger, averaging about 14.4 and 11 percentage points, respectively. In other words, their indirect effects are in the opposite direction and about 18 times and 3.2 times larger than their respective direct effects. The case for ratifying the CEDAW is much more clear-cut and stronger. CEDAW participation improves women's empowerment both directly and indirectly through its causal mediators,

but the indirect causal impact is nine times larger than the direct effect. In fact, the four causal mechanisms I have examined are jointly responsible for transmitting roughly 90% of the CEDAW effect. Overall, these findings underscore the importance of causal mediators and suggest that the efficacy of human rights treaties is mostly about causal mechanisms—both domestic institutions and social movements and international socialization.

Table 3.3.: Total effect (TE), joint natural direct effect (CDE), joint natural indirect effect (NIE), and controlled direct effect (CDE) estimates of the ICCPR, the CEDAW, and the CAT. Super Learner-based point estimates and bootstrap ($B = 500$) quantile-based 95% CI with single stochastic imputation. All measures of human rights outcomes are rescaled into a bounded 0–1 range.

	ICCPR (PTS score)	CEDAW (WPE index)	CAT (CIRI index)
TE $= E\big[Y_{1,M_1} - Y_{0,M_0}\big]$	**0.136**	**0.220**	**0.077**
	[0.093, 0.196]	**[0.197, 0.250]**	**[0.028, 0.123]**
NDE$_{\text{joint}}$ $= E\big[Y_{1,M_0} - Y_{0,M_0}\big]$	**−0.008**	**0.022**	**−0.034**
	[−0.018, −0.001]	**[0.016, 0.027]**	**[−0.054, −0.013]**
NIE$_{\text{joint}}$ $= E\big[Y_{1,M_1} - Y_{1,M_0}\big]$	**0.144**	**0.199**	**0.110**
	[0.098, 0.202]	**[0.177, 0.228]**	**[0.063, 0.157]**
CDE (*political constraints = 0)*	−0.004	**0.021**	**−0.032**
	[−0.013, 0.000]	**[0.015, 0.028]**	**[−0.049, −0.015]**
CDE (*judicial independence = 0)*	−0.004	**0.028**	**−0.036**
	[−0.012, 0.000]	**[0.021, 0.035]**	**[−0.052, −0.018]**
CDE (*international socialization = 0)*	−0.004	**0.021**	**−0.022**
	[−0.015, 0.000]	**[0.014, 0.027]**	**[−0.040, −0.004]**
CDE (*mass mobilization = 0)*	−0.004	**0.022**	**−0.029**
	[−0.013, 0.000]	**[0.015, 0.028]**	**[−0.046 , −0.013]**
N. of countries	192	192	192
N. of years	28 [1981–2008]	28 [1981–2008]	22 [1987–2008]
N. of observations	5,268	5,268	4,290

It is worth noting that the literature remains divided when it comes to quantifying the consequences of participating in the CAT for human rights protection. While many research has indicated either a negative [Hathaway, 2002, Hill, 2010] or a positive effect [Simmons, 2009, Fariss, 2014], most have found an ambiguous or context-specific treaty effect [Hafner-Burton and Tsutsui, 2007, Conrad and Ritter,

2013, Lupu, 2013a, Conrad, 2013, Clark, 2013]. My analysis moves this ongoing debate forward by showing that it is the direct causal effect that worsens human rights practices whereas the indirect effect is substantially more positive. This finding is only possible once we conceptualize, identify, and then estimate the natural direct and indirect effects, using the framework that Pearl [2001] proposed in 2001. Further research could bring a deeper understanding as to why the treaty's direct effect is not in the direction we expected and whether the CAT might serve as a cover for participating states to ramp up their repression and abuses.

The importance of causal mediators is further supported by my CDE estimates. Once we have demediated the outcome and removed the effect of each individual intermediate variable, the positive causal effect of CEDAW ratification diminishes significantly, ranging from 2.1 to 2.8 percentage points, depending on the mediators. For the ICCPR, all CDE estimates are essentially zero. In other words, setting each of the causal mediators at its empirically lowest value and the causal effects of the ICCPR and the CEDAW will decrease so much as to be no longer very meaningful. I therefore conclude that all four causal mechanisms—political constraints, domestic judicial enforcement, mass mobilization, and international socialization—have a critical role to play in transmitting the treaty effects. Without them, human rights treaties will lose most of their causal impact.

Causal mediators are especially important for the CAT. The CDE estimates of the CAT suggest that crippling domestic institutions and blocking international socialization will effectively open the way for member states to potentially use treaty ratification as a cover to *increase* torture by somewhere between 2.2 to 3.2 percentage points. A note of caution is that the CIRI torture index that I use to measure the outcome might have a biased tendency against recorded improvements in human rights practices and thus potentially understate the positive impact of the CAT [Clark and Sikkink, 2013, Fariss, 2014].

Finally, an interesting finding is that individual CDE estimates are relatively similar to each other across all four causal mediators and to the overall NDE, suggesting

that the mediators are highly closely related. This could be interpreted as indicating an interplay and entanglement among the causal mechanisms, supporting my previous assumption about direct causal dependence among the mediators. Were the intermediate variables causally independent from each other, we would probably have seen a much greater variation among CDE estimates. Suppose, for example, domestic judicial enforcement was irrelevant while international socialization was a highly causally important mechanism and both of them were conditionally independent, their CDE estimates would diverge in both magnitude and statistical significance. The finding also implies that any causal analysis to untangle the causal importance of individual causal mechanisms would be very difficult and that these mediators should be included together in future research on human rights treaties.

3.4 Conclusion

Conducting causal mediation analysis to learn about causal mechanisms has become increasingly popular in many fields, including political science [Imai et al., 2011, Imai and Yamamoto, 2013]. A research area where causal mediation analysis could offer much needed insights is relating to the causal impact of human rights treaties. Multiple theories in the literature have articulated various causal pathways along which the effect of human rights treaties could be transmitted. Yet, besides a number of qualitative research that uses case study methods to illustrate the causal logics and examine the causal process, there is no quantitative inquiries in the existing literature that empirically investigate these causal pathways on a systematic basis.

This research gap is the substantive motivation that has led me to leverage recent advances in both machine learning and the causal inference literature to define, identify, and estimate the total effect, the natural direct effect, the natural indirect effect, and the controlled direct effect of three major UN human rights treaties, including the ICCPR, the CEDAW, and the CAT. These effect estimates help

decompose the causal effect of human rights treaties, indicate how much of the treaty effect is mediated through the causal mechanisms, and quantify the importance of causal mediators in preserving the effectiveness of international human rights institutions. Overall, the causal mediation analysis in this chapter advances our collective understanding of the ways in which human rights treaties constrain and influence state behaviors.

The results indicate that an overwhelming portion of human rights treaty effect is mediated through four causal mechanisms. As a limitation of my analysis, I am not able to tease out the exact portion of treaty effect that each intermediate variable mediates. This limitation is due to empirically valid concerns about the plausibility of assuming that all four causal mechanisms are causally independent from each other. The causal analysis nonetheless indicates that the four causal mechanisms, including legislative constraints, domestic judicial litigation and enforcement, human rights NGOs advocacy and mobilization, and international socialization, are all critical. Furthermore, it is highly likely that these four causal mechanisms are intertwined with each other.

The broad implications are clear that these four causal mediators are extremely critical to the success and efficacy of the UN human rights treaties. Without them, all three human rights treaties under examination here would lose most, if not all, of their positive causal impact and, in the case of the ICCPR and the CAT, might even become a negative influence on human rights protection. As a result, advocating for universal ratification of human rights treaties is not even nearly enough. To protect and defend the efficacy and the causal impact of international human rights treaties requires domestic and international efforts to guard against government attempts to control the legislature, undermine the rule of law, and restrict the space for NGOs and the civil society to operate.

In addition to the substantive contributions, my machine learning-based causal mediation analysis demonstrates several aspects of an innovative application. First, I apply the graph-based structural causal model framework to a new setting of panel

data with repeated outcome measures. This data structure is very common in many areas of international relations and political science research, but causal inference research has mostly been done using cross-sectional data. My application is an attempt to apply this causality framework to this new kind of data structure.

Second, to perform the causal analysis in this chapter, I exploit the facility of causal graphs to assist identification via separate adjustment sets of covariates. My analysis demonstrates an application of this piecemeal, independent adjustment approach to identification in a real-world research. It exemplifies the great benefits of the divide-and-conquer strategy for causal effect identification that Pearl [2014b] has proposed. More broadly, independent adjustment has the potential to increase the number of scenarios under which various causal effects can be identified and estimated.

Third, my causal mediation analysis also illustrates an inherent trade-off between causal assumption plausibility and causal effect estimability, which is rarely found or mentioned in empirical research for causal inference. In the case of multiple causally connected mediators that are examined here, this trade-off is made particularly apparent. The approach that I adopted to deal with this tradeoff is to avoid making unrealistic causal assumptions while striving to produce causal findings about the mechanisms of human rights treaties that are at least as substantively useful and relevant as possible.

Finally, in this analysis I employ machine learning for robust causal effect estimation. Unless there are compelling reasons to believe that a specific model specification truly reflects the unknown underlying data-generating process, it is often preferable to use flexible and powerful machine learning methods so that the consistency of our estimates and the validity of our inferences are not dependent upon the accuracy of restrictive functional form assumptions.

4. UNPACKING TREATY IMPACT: THE DIFFERING CAUSAL EFFECTS OF HUMAN RIGHTS MONITORING PROCEDURES

4.1 Introduction

For many decades, the United Nations (UN) human rights treaty system has been a crucial endeavor to protect human rights across the world. An expansive body of research, including an ongoing research project by a network of independent researchers [Kolb, 2016], has examined the impact of individual human rights treaties on state practices [Hathaway, 2002, Landman, 2005, Neumayer, 2005, Hafner-Burton and Tsutsui, 2007, Simmons, 2009, Hill, 2010, Lupu, 2013b, Clark, 2013]. Yet, few have evaluated the comparative effectiveness of monitoring procedures under these treaties. Two following anecdotal examples highlight the impact of a treaty monitoring procedure under two different UN human rights treaties—the inclusion of an individual communication mechanism—in addressing human rights abuses and potentially changing state behaviors. In both cases, the states, when ratifying the human rights treaties, opted to recognize the competence of the treaty monitoring bodies to receive and consider complaints of human rights violations by individuals against these states and to issue their findings and decisions.

On July 29, 1997, the Human Rights Committee, the treaty body under the International Covenant on Civil and Political Rights (ICCPR), in the case of *Arhuaco* v. *Colombia* (Communication No. 612/1995) found the Colombian government responsible for the torture and arbitrary detention of Jose Vicente, Amado Villafane Chaparro, and others. Following the Human Rights Committee's decision, the Colombian government, under its own Law 288/96, issued an opinion in favor of compliance with the decision and later let the case proceed in national courts [Ulfstein and Keller, 2012, 365]. This adjudication, called Views by the Human

Rights Committee, was adopted under the First Optional Protocol to the ICCPR, to which Colombia was a party, that establishes the individual communication mechanism.

Similarly, on May 11, 2001, the Committee against Torture, the treaty body under the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT), issued its decision in *Ristic* v. *Yugoslavia* (Communication No. 113/1998), concluding that the government had violated its obligations under the CAT and recommending remedial measures by the state party. Even though the Committee's decision was not legally binding, the country's Supreme Court later endorsed the decision and ordered reparations for the victims [Ulfstein and Keller, 2012, 369–370]. The Committee's decision was made under Art. 22 of the CAT, according to which the Committee has the competence to receive and consider communications from or on behalf of the victims and the Yugoslavia had accepted this competence.

Going beyond anecdotal evidence like these two examples, this chapter focuses on treaty compliance monitoring procedures and examines their causal impact on a more systematic basis, yielding more insights into the empirical effectiveness of the UN human rights treaties in more granular details. Human rights treaties not only set the standards of behavior for states parties; they also engage in compliance monitoring using a variety of monitoring procedures. The motivating idea of this research is to unpack treaty monitoring practices into monitoring procedures and estimate the causal effect of each procedure on human rights outcome, thereby providing a more detailed picture about the causal impact of international human rights treaties. Overall, there are five types of treaty monitoring procedures, including state reporting, inquiry, state communication, individual communication, and country visit. I exclude the state communication procedure from my investigation because this procedure has never been used in practice before.

State reporting refers to the procedure according to which a state party submits periodic self-reports on the measures it has taken and the progress it has made to

fulfill its treaty obligations. State reporting is also the only mandatory procedure under all UN human rights treaties. Participating in other monitoring procedures is optional, which allows treaty members to opt out by way of making a reservation (the inquiry procedure) or requires them to opt in through a unilateral declaration (the state communication procedure and the individual communication procedure) or specifically demands a separate formal ratification (the country visit procedure under the Optional Protocol to the CAT).

Under the state communication procedure, the treaty body—a committee of independent experts—hears complaints that one participating state may bring against another participating state. The individual communication procedure such as the one under Art. 20 of the CAT allows the treaty body to receive and adjudicate complaints brought against a state party by individuals within that state's jurisdiction. In the absence of a declaration to opt out the inquiry procedure, treaty members have to answer questions and inquiries that the treaty body may have regarding allegations of systematic violations. The treaty body may also conduct an inquiring visit under the inquiry procedure to investigate allegations of treaty violations.

The only operative country visit procedure in the UN human rights treaty system as of this writing is the one under the Optional Protocol to the CAT (OPCAT).[1] It permits a separate monitoring body—the Subcommittee on Prevention of Torture and other Cruel, Inhuman or Degrading Treatment or Punishment (SPT)—to visit, investigate, report, and even publish its findings on the torture practices of a member state. Table 4.1 summarizes the monitoring procedures under the CAT [Egan, 2011, Keller and Ulfstein, 2012, Rodley, 2013, Bassiouni and Schabas, 2011]. Appendix C.1 provides similar information for all nine UN core human rights treaties as well as a detailed list of these treaties and their current status of ratification. Figure 4.1 then shows the number of states that participate in each of the four monitoring

[1]Under the UN Charter-based human rights bodies system, including the UN Human Rights Council and its subsidiaries such as the Special Procedures, country visits do take place [Kothari, 2013]. However, they are beyond the scope of examination in this study, which focuses on the UN treaty-based human rights system.

procedures under the CAT and the OPCAT from 1984 when the CAT was open for ratification until 2015.

Table 4.1.: Monitoring procedures under the Convention against Torture (CAT)

Procedure	State reporting	Inquiry	State communication	Individual communication	Country visit
Provision	Art. 19	Art. 20	Art. 21	Art. 22	Optional Protocol
Participation	Mandatory	Opt-out allowed	Opt-in required	Opt-in required	Ratification required

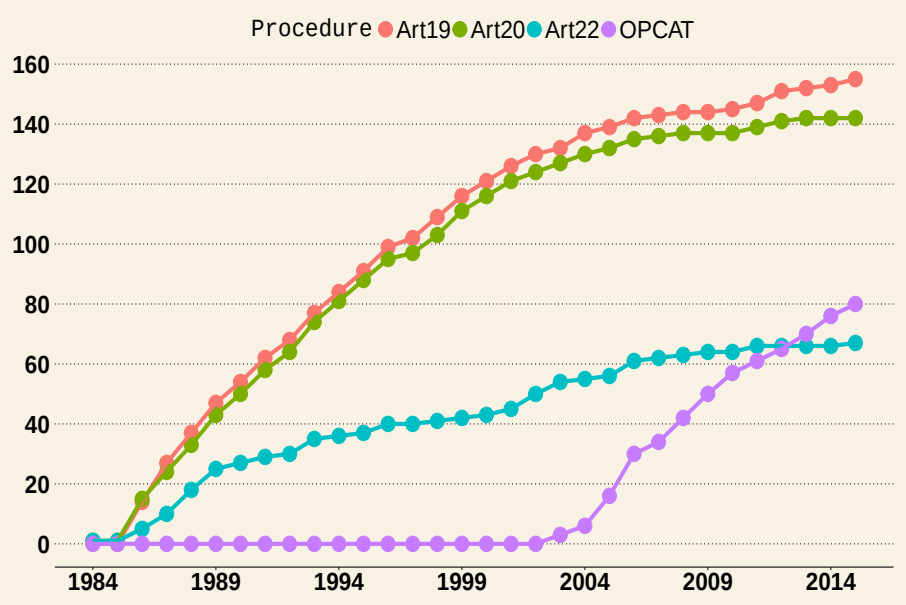


Fig. 4.1.: The number of states that participate in each of the four monitoring procedures under the CAT from 1984–2015, including the state reporting procedure under under Art. 19, the inquiry procedure under Art. 20, the individual communication procedure under Art. 22, and the country visit procedure under the OPCAT. Since the state reporting procedure under Art. 19 is mandatory, the number of states participating in this procedure is also the number of CAT ratifiers. The CAT was open for ratification in December 1984 and went into effect in 1987. The OPCAT was open for ratification in December 2002 and entered into force in 2006.

I investigate the causal effects of the monitoring procedures under the CAT and the OPCAT for several reasons. First, the CAT is one of the most important in-

ternational treaties that protects a universal and non-derogable human right—the right not to be subject to torture. Second, the CAT, together with the OPCAT, is the only UN human rights treaty currently in effect that has all five monitoring procedures. Especially important among them is the country visit system under the OPCAT, which is designed and operated by a separate monitoring body with clearly designated authority. As a result, among all the UN human rights treaties, the CAT provides the only case study where we can evaluate and compare the causal effects of all treaty monitoring procedures. Finally, it is worth noting that the CAT aims to address a single type of human rights violations (torture) and protect a highly specific human right (the right not to be subject to torture). This treaty is different from other UN human rights treaties that address a wide range of rights such as civil and political rights, women's rights, and children's rights. We can therefore more easily and properly compare the causal effects of monitoring procedures under the CAT since they operate in the same specific area of human rights.

Substantively, my causal analysis indicates that only the country visit procedure significantly and consistently reduces torture and improves government respect for physical integrity rights. Other procedures demonstrate no such causal impact. In fact, the state reporting procedure tends to have a negative, though not always significant, impact on human rights protection. These differing causal effects, I argue, are most likely the result of the variation in intrusiveness among the monitoring procedures. Future research should explore the causal effects of monitoring procedures under other human rights treaties in order to draw a more definitive conclusion as to whether intrusive monitoring procedures with stronger external oversight lead to greater improvements in relevant human rights outcomes. More broadly, a similar systematic relationship could be detected with respect to monitoring mechanisms in other domains of international relations such as weapons inspection and nonproliferation.

Methodologically, this study also presents an innovative application of the graph-based structural causal model framework and machine learning-based estimation to

a substantive research question in political science. A combination of a transparent causal inference framework and flexible machine learning methods has great potential to advance political science research much further in the future.

4.2 Theoretical Proposition

I develop a proposition to explain why intrusive monitoring procedures may have a greater causal impact on human rights outcome than less intrusive ones. It starts with the empirical observation that monitoring procedures vary in their intrusiveness to state sovereignty. This intrusiveness has two implications. The first one is that participating in an intrusive procedure sends a credible signal to international audiences about a state's intent to protect human rights. Second, an intrusive monitoring procedure is likely to generate more information about the actual behavior of participating states. By providing more credible signals *ex ante* and more information *ex post*, intrusive procedures improve state practices by raising both the cost of treaty violations as well as the probability of getting caught violating treaty obligations.

Where my argument differs from, and contributes to, the literature is mainly with respect to the key independent variable of interest. Unlike the existing literature that mainly focuses on treaty ratification, I instead (1) disaggregate treaty ratification into separate state decisions to participate in different treaty monitoring procedures; (2) classify these procedures by their intrusiveness; (3) explain the signaling value and the informative power of monitoring procedures as a function of their intrusiveness; and (4) empirically estimate the causal effects of treaty monitoring procedures. A stylized illustration of the causal process according to this theoretical proposition is presented in Figure 4.2.

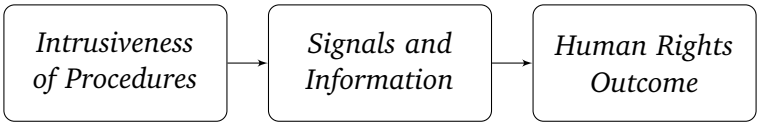


Fig. 4.2.: Causal model of monitoring procedures

4.2.1 Intrusiveness of monitoring procedures

A key mandate of human rights treaty bodies involves monitoring treaty compliance by member states, using a variety of procedures that can vary in their intrusiveness. One way to classify the intrusiveness of monitoring procedures is to use the three criteria under the legalization framework, including obligation, precision, and delegation, all of which are defined "in terms of key characteristics of rules and procedures, not in terms of effects" [Abbott et al., 2000, 402]. Obligation concerns whether a particular rule imposed upon states parties is legally binding. Precision refers to the degree to which obligations are unambiguously and clearly specified. Delegation measures "the extent to which states and other actors delegate authority to designated third parties [. . .] to implement agreements" [Abbott et al., 2000, 415]. Hafner-Burton et al. [2015b, 11] apply this three-criterion framework to develop a ten-indicator measure of the sovereignty costs of a large number of human rights institutions. In the case of monitoring procedures under the CAT, I adapt these criteria to develop a more targeted and descriptive ranking based on the treaty language as well as the actual implementation of each procedure.

First, obligations range from being explicitly non-binding to merely hortatory to unconditionally binding. In a state reporting system, states voluntarily submit self-reports for review by the treaty bodies. Many treaty members, however, either substantially delay their initial and periodic submissions or simply ignore this obligation altogether. According to a 2012 report by the UN High Commissioner for Human Rights, around 20% of the states parties to the CAT have never submitted any reports at all [Pillay, 2012, 22]. When participating in other monitoring pro-

cedures, however, states do not get to choose whether and when they are subject to scrutiny by the treaty monitoring bodies. Under the inquiry procedure, for example, the treaty bodies can actively initiate inquiries into allegations of systematic violations without having to wait for any state reports.

Obligation is more demanding under the individual communication procedure. Monitoring no longer depends on a state's periodic submission of self-reports. Rather, it is in response to submitted individual complaints. Unlike their reviews and comments on state reports under the reporting system, adjudication by the treaty bodies under the individual communication procedures also carries a "great weight"—a status that has gained wide consensus among international legal scholars and treaty bodies and is also acknowledged and affirmed by the International Court of Justice [Ulfstein and Keller, 2012, 92–94]. An indication that participating states take a treaty body's adjudicating decisions seriously is that many of them choose to respond and defend themselves against allegations. States accused of treaty violations regularly argue in front of the treaty bodies not only against the merits but also against the admissibility of submitted complaints and the standing of alleged victims.

The country visit procedure that the SPT operates under the OPCAT likely imposes the most onerous and binding obligations. This procedure takes the form of regular and unannounced visits to places of detention, broadly defined, within any territories under the jurisdiction of OPCAT members (Art. 4). According to its First Annual Report (p. 25), for example, the SPT visited not only police facilities and prisons, but also juvenile and shelter centers, children's homes, and drug rehabilitation centers. Its Fifth Annual Report (p. 13) also states that the SPT tried "to increase its activities in relation to non-traditional places of detention during 2011, including immigration facilities and medical rehabilitation centres." As an indication of the unrestricted nature of the SPT's visits, OPCAT members are obligated to grant the SPT access to any visit sites even in the case of a state emergency (Art. 14). The SPT is able to request any relevant information and interview in private

any persons it believes can supply relevant information (Art. 14 and Art. 15). States parties are usually notified and consulted about an upcoming visit, but the purpose is "to facilitate the visit, not to prevent it from occurring" and, in fact, "no State has objected to a visit proposed by the SPT" [Steinerte et al., 2011, 98]. By its own account, over the last ten years since the SPT started its work in February 2007, it has conducted visits to 50 member states [SPT, 2018]. In addition to the SPT, Articles 17–23 of the OPCAT also require participating states to establish or designate a national preventive mechanism, preferably in the form of an independent national human rights institution, as both a liaison and an oversight body that has similar mandate and authority as the SPT. Finally, states parties are not able to make reservations to any of the provisions in the OPCAT (Art. 30).

The second criterion is precision. The basic idea is that precise and specific rules make it easier for monitoring bodies to determine whether alleged violations are factual and accurate and which reparations are merited. Monitoring procedures that operate according to more precise rules are therefore more likely to make concrete determination about state compliance. The individual communication procedure fare best according to this evaluative criterion. Under this procedure, the competence to adjudicate submitted complaints, the composition, and the rules of operation of the treaty bodies are highly precise and are even deemed comparable to those of international courts [Ulfstein and Keller, 2012, 98]. Decisions by the treaty bodies with respect to individual complaints also have the effect of clarifying the legal content and contributing to more precise interpretation of international rules for national governments and domestic courts. This semi-judicial role of the treaty bodies under the individual communication procedure, while not producing legally binding and directly enforceable decisions, could nonetheless prove highly intrusive to the domestic governance of states parties, especially compared to the state reporting procedure. As a former UN Special Rapporteur on Torture has observed, the individual communication procedure "is the most court-like function of the treaty bodies" whereas "[the state reporting system] is a mode of reviewing

compliance with the treaties' obligations in minimally intrusive manner" [Rodley, 2013, 634].

The third criterion concerns delegation. In the context of human rights treaty monitoring procedures, it refers to the amount of authority that treaty members delegate to the treaty bodies to carry out monitoring compliance and promote implementation of treaty obligations. One way to assess the amounts of delegated authority across multiple monitoring procedures is to examine the rules that treaty bodies have developed on their own to carry out their different mandates. Delegated authority is minimal under the state reporting procedure because under that system the treaty bodies are in a passive position to receive and make comments on the self-reports that states parties care enough to submit. The inquiry procedure under Art. 20 of the CAT grants more authority to the treaty body, including the ability to initiate inquiries into allegations of systematic violations. However, Art. 20(5) requires the treaty body to seek cooperation from the state under inquiry "at all stages of the proceedings," placing a significant limitation in terms of their delegated authority.

Under the individual communication procedure, there is more delegated authority since treaty bodies could receive and deliver judgments on complaints by alleged victims or those acting on their behalf. States parties cannot halt or delay the process even if they refuse to participate in arguments or respond to allegations. Similarly, treaty bodies can also reach a judgment on the admissibility and merits of submitted complaints with or without the inputs and defense put up by the accused governments. The SPT, the treaty body under the OPCAT, arguably has the greatest amount of delegated authority to implement the country visit procedure. According to the OPCAT, the SPT could randomly select a member state in which to conduct an investigative visit without having to secure any authorization. In fact, according to its First Annual Report, the SPT selected its first batch of country visits by random. Selected states are obligated to grant the SPT access to any places within

their jurisdiction and the SPT could later publish its findings by a simple majority decision in the CAT Committee, the treaty monitoring body of the CAT.

When aggregating over these three criteria, I approximately classify monitoring procedures under the CAT and the OPCAT by their intrusiveness. The country visit system under the OPCAT ranks first by the two criteria of obligation and delegation whereas the individual communication procedure ranks first by the precision criterion. Similar to Hafner-Burton et al. [2015b], I assume that obligation, precision, and delegation contribute roughly equally to the overall metrics of intrusiveness and therefore classify the country visit procedure as more intrusive than the individual communication and inquiry procedures and the state reporting procedure is the minimally intrusive procedure among them.

4.2.2 Signal about intent and information about compliance

Two logical implications follow the variation in intrusiveness among treaty monitoring procedures. First, ratifying an intrusive monitoring procedure will credibly reveal a state's intent to comply with treaty obligations [Farber, 2002]. The logic is straightforward. Intrusive monitoring procedures impose a high sovereignty cost, defined as constraints on a state's freedom of action, that only a state genuine about compliance could ratify and maintain its ratification status without having to withdraw by Art. 31 of the CAT and Art. 33 of the OPCAT. According to this logic, by subjecting itself to the country visit procedure, for instance, a state sends a credible signal about its intent to comply than is the case if it merely participates in the state reporting system.[2] The reason is that hosting unannounced and unrestricted

[2]Hollyer and Rosendorff [2011] offer a different signaling logic. According to their argument, dictators sign the CAT to signal their strength rather than their intent to comply. Domestic opposition groups perceive a dictator's treaty commitment and subsequent ostentatious treaty violations as credibly signaling her strength. As a result, they are less likely to mount a challenge, in effect prolonging the survival of the authoritarian regime. Although this argument "has some plausibility problems on its face" [Simmons, 2012, 743], it has not been contested empirically. I note three key differences between my analysis and Hollyer and Rosendorff [2011]. First, my target population for inference is all countries, not just the subset of dictatorships. Second, my independent variable of interest is participation in monitoring procedures under the CAT, not just a commitment to the

visits by international independent experts, and therefore running the risk of having state violations and abuses exposed, is significantly more costly than submitting self-reports at the discretion of the government. Even if fulfilling the state reporting obligation might not be entirely costless [Goodman and Jinks, 2003, Cole, 2009, Hafner-Burton et al., 2015b], the cost of participating in an intrusive procedure could be significantly higher. It is this higher cost that makes the intent to comply more credible.

The history of the OPCAT is instructive of how concerns about sovereignty costs could delay or even stymie the adoption of a comparatively more intrusive monitoring procedure. The formulation of the OPCAT was modeled after a proposal by the Swiss Committee against Torture, which was later renamed as the Association for the Prevention against Torture, during the negotiation for the CAT in the late 1970s and early 1980s. At that time, this proposal, known as the "Swiss model," was considered so intrusive to state sovereignty, especially compared to the "Swedish model" that the CAT eventually adopted, that the "Swiss model" had to be rejected [Clark, 2009, Evans, 2011]. It took almost 15 years after the adoption of the CAT for the "Swiss model" to be revived, modified, and adopted as the OPCAT on December 18, 2002. In other words, both the CAT and the OPCAT used to be considered at the same time, but they were eventually adopted almost two decades apart because they differ significantly in terms of their intrusiveness. As a result, their ratification sends signals of different levels of credibility.

The second implication is that intrusive monitoring is likely to generate more information about state compliance. The SPT's visit to Paraguay in 2009, for example, produced a 313-paragraph report on that country's torture practice alone. This is because, compared to the CAT Committee, the SPT is under fewer restrictions when

---

treaty and its mandatory state reporting system. Third, the hypothetical policy intervention (or the treatment) in my analysis is ratification as opposed to signature as in the analysis by Hollyer and Rosendorff [2011]. The third point is critical because, under international law, signing a human rights treaty does not activate any treaty monitoring activities. Nor does it create any treaty obligations for the signatories, except for the legally vague obligation "not to defeat the object and purpose of a treaty prior to its entry into force" according to Article 18 of the Vienna Convention on the Law of Treaties [Jonas and Saunders, 2010].

monitoring and exposing government abuses. The SPT also has more clout because of its ability to publish a state's substandard record. Conversely, governments with a good record, by Art. 16(2) of the OPCAT, could ask the SPT to publish its positive assessment. According to the Fourth Annual Report of the SPT, for instance, five visit reports have been published following requests by Honduras, Maldives, Mexico, Paraguay, and Sweden. This design enhances the ability of OPCAT participation to effectively separate compliant states from violating countries, thus reducing the risk of OPCAT ratification being used merely as a cover for human rights abuses. The individual communication procedure is also likely to generate more compliance information because individual complaints provide more detailed information about state abuses and the treaty body's decisions in response will likely remove remaining ambiguity about state compliance.

In summary, a state's voluntary decision to subject itself to an intrusive monitoring procedure credibly signals its intent to comply. That procedure is also likely to produce more information about a state's actual compliance. This *ex ante* signal and *ex post* information decrease government repression through a couple of mechanisms. First, ratifying an intrusive procedure that imposes a significant sovereignty cost will establish a higher baseline expectation by international audiences. Other countries, inter-governmental organizations, and non-governmental organizations (NGOs) perceive a strong and credible signal from the ratifying state. A greater public expectation results in a higher reputational cost for treaty violations [Brewster, 2013]. In other words, by raising the reputational stake, participation in an intrusive monitoring procedure raises the cost of violations and thereby reduces the incentive to commit treaty violations in the first place.

Second, intrusive monitoring is also more likely to detect non-compliant behaviors, increasing the probability that a member state gets caught violating its treaty obligations. This is especially important in the case of serious state torture, which tends to occur in countries for whom protective domestic institutions are already weak or ineffective. Furthermore, as Lupu [2013b, 477-481] points out, even inde-

pendent domestic courts could encounter enormous difficulties when enforcing international commitments to protect physical integrity rights. A major reason is that repressive governments have a considerable capacity to interfere with witnesses, hide the victims, and destroy evidence, thereby raising the cost of producing legally admissible evidence. The informative power of intrusive monitoring could be especially useful under such difficult circumstances and can provide evidence that is hard to obtain otherwise.

The unannounced and unrestricted nature of the SPT's visits also produces the kind of evidential information that could be instrumental for legislative opposition parties [Lupu, 2015], social movements, and civil society pressure [Simmons, 2009] to constrain and hold abusive state officials accountable. The SPT's Second Annual Report, for example, mentions that the SPT has "carried out unannounced visits to places of detention [and] had interviews in private with persons deprived of their liberty" (p. 8). Its Third Annual Report reiterates that "confidential face-to-face interviews with persons deprived of liberty are the chief means of verifying information and establishing the risk of torture" (p. 9). The same report also raises concern that many detainees whom the SPT spoke with may become a target of reprisal afterward (p. 11). The risk that detainees have taken in talking to the SPT suggests that the kind of information the SPT gathers is highly unlikely to obtain in voluntary state reports and constructive dialogues that states parties occasionally have with treaty body experts under the state reporting procedure.

To summarize, in addition to raising the reputational cost of human rights abuses, intrusive monitoring procedures also increase the probability of detecting violations in participating states by producing compliance information that is usually not available under less intrusive procedures. This information factors into domestic institutions such as the legislature, domestic courts, and social movements that exert a constraining effect on state behavior. Here I make no assumption as to which causal pathway—international reputation or domestic institutions or social movements—is more effective. They may operate more or less independently

or, more likely, there could be a complex interplay between them. Regardless, it should not prevent us from estimating the causal effects of treaty monitoring procedures. If a relationship truly exists between the intrusive design of treaty monitoring procedures and their causal effects, we expect more intrusive procedures such as individual communication and especially country visit will have a larger, more consistent causal effect than do less intrusive ones such as state reporting. It is noteworthy that the variation in intrusiveness among monitoring procedures does not only exist; it is by design. In fact, it is probably not a coincidence that more intrusive monitoring procedures are almost always presented in an optional protocol or an optional provision of a human rights treaty.

4.3 Empirical Analysis

This section estimates the causal effects of participating in the state reporting procedure under Art. 19 of the CAT, the inquiry procedure under Art. 20, the individual communication procedure under Art. 22, and the country visit procedure under the OPCAT. I use observational data on 192 countries from 1987 when the CAT went into effect until 2015. Adopting the causal inference framework that Pearl [2000, 2009a] develops, I begin by formulating a structural causal model that formalizes the background knowledge about the data generating process. I then use a causal graph to represent this structural model, encoding the necessary causal assumptions, and establishing the estimability of the causal effects of interest. Establishing estimability, known as causal identification, is essential because it specifies the conditions for translating an interventional distribution of the outcome when we intervene to fix the treatment values into the observational distribution of the outcome that we observe. It permits, assuming identification conditions are satisfied, an estimation of the causal effect using observational data. Finally, I employ the machine learning-based targeted maximum likelihood estimator to compute effect estimates that are more robust than those obtained via standard parametric

statistical models. I conclude with our interpretation of the effect estimates by mapping them back into my substantive research question.

4.3.1 Causal model formulation

To make causal inference in a non-experimental setting, one has to start by assuming that the observed data come from a data-generating system that is compatible with a structural causal model. A structural causal model is simply a set of equations that make explicit our notion about "how nature assigns values to variables of interest" [Pearl et al., 2016, 27]. Our structural model in Equation set 4.1 describes the functional relationships between human rights outcome Y, a set of four treatments $A = \{A1, A2, A3, A4\}$, mediators M, and a set of time-invariant confounders X and time-varying confounders W. The subscript t indicates the time period during which the observed variables are measured.

The treatments, whose causal effects are our target of estimation, include $A1$ (the state reporting procedure), $A2$ (the inquiry procedure), $A3$ (the individual communication procedure), and $A4$ (the country visit procedure). To measure the treatment variables, I use a state's formal ratification of the CAT, the presence or absence of a reservation to Art. 20, a declaration of intent to be bound by Art. 22, and formal ratification of the OPCAT. Since the state reporting procedure is mandatory under the CAT, a measure of participation in this procedure completely overlaps with ratification of the CAT. I do not measure how participating states actually engage with the periodic review process under state reporting system [Creamer and Simmons, 2015] or with other procedures. Measurements of actual monitoring activities, while seemingly more intuitive, may miss the signaling value of formal ratification. Equally important, the lack of data would present an insurmountable challenge.

$$
\begin{aligned}
X &= f_X(U_X) \\
W_t &= f_W(X, W_{t-1}, U_W) \\
A1_t &= f_{A1}(X, A1_{t-1}, M_{t-1}, Y_{t-1}, W_t, U_{A1}) \\
A2_t &= f_{A2}(X, A2_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t, U_{A2}) \\
A3_t &= f_{A3}(X, A3_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t, U_{A3}) \\
A4_t &= f_{A4}(X, A4_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t, U_{A4}) \\
M_t &= f_M(X, M_{t-1}, A1_t, A2_t, A3_t, A4_t, W_t, U_M) \\
Y_t &= f_Y(X, Y_{t-1}, A1_t, A2_t, A3_t, A4_t, M_t, W_t, U_Y)
\end{aligned}
\tag{4.1}
$$

Informed by the existing literature about the causal mechanisms through which a human rights treaty could influence state behavior, I specify four potential causal mediators, including institutional and legislative constraints [Lupu, 2015], domestic judicial enforcement [Powell and Staton, 2009, Conrad, 2013], civil society mobilization [Simmons, 2009], and international socialization [Keck and Sikkink, 1998, Clark, 2013, Goodman and Jinks, 2013]. They are believed to be transmitting and mediating the causal effects of human rights treaties. The mediators are respectively measured by (a) a political constraints index, which measures the feasibility of policy change based on the veto power and alignment among government branches and degrees of preference heterogeneity within the legislative branch [Henisz, 2002]; (b) latent judicial independence estimates, which measure the independent power of the judiciary to constrain choices of the government [Linzer and Staton, 2015]; (c) a naming and shaming index, an aggregation of reporting on human rights abuses by major media outlets, Amnesty International, and the UN Commission on Human Rights, [Cole, 2015, 423] which reflects the work of a civil society, particularly domestic NGOs, in calling out for attention to state abuses; and (d) treaty commitment preference coordinates based on state ratifications of 280 universal treaties across a large number of policy areas [Lupu, 2016], which is a good indicator of the extent to which countries interact and socialize in-

ternationally, particularly when it comes to transmitting and adopting human rights norms [Greenhill, 2016].

To investigate the robustness of my causal effect estimation, I also use three different measures of human rights outcome, including the Political Terror Scale [Gibney et al., 2016], human rights protection scores [Fariss, 2014], and the CIRI torture index [Cingranelli et al., 2013]. Each of these datasets has a different measurement timeframe. I therefore right-censor my analysis accordingly. Finally, I include a number of time-invariant covariates (legal origin, treaty ratification rule, and electoral system) and time-varying covariates (population size, gross domestic product (GDP) per capita, participation in international trade, regime type, regime durability, and involvement in international conflicts). As indicated in the literature, these covariates are the usual suspects that may confound the relationship between treaties and human rights practices. Table 4.2 lists the observed variables, indicates the timeframe of measurement for each variable, refers to studies in the literature that similarly examine these variables, and identifies the data sources. Appendix C.2 provides a more detailed description of data sources, variable measurements as well as the recoding, preprocessing, and transformation of variables for data analysis.

From the model of the generative system in Equation set 4.1, we observe a sample of n country–year observations $O_n = (X, W_t, A_t, M_t, Y_t) \sim P_O$ where P_O is the joint probability distribution of the observed variables. In estimating the contemporaneous causal effect of each treatment, I compute the average change in human rights outcome Y_t as if we could physically intervene to alternate the ratification/-participation status of a monitoring procedure for all country–year observations. The effects of these interventions are expressed in terms of the mean of the interventional outcome distribution: $E_{P_O}\big[Y|do(A = 1)\big]$ and $E_{P_O}\big[Y|do(A = 0)\big]$. The do-operator indicates an active intervention to fix the treatment value at $A = 1$ (ratified) and $A = 0$ (non-ratified) and the expectations are taken over the entire

Table 4.2.: Causal model variables

Sets	*Timeframe*	*Variables, References, and Data Sources*
A	1986–2015	$A1$: Art. 19 ratification status (OHCHR).
	1986–2015	$A2$: Art. 20 reservation status (OHCHR).
	1986–2015	$A3$: Art. 22 declaration status (OHCHR).
	1986–2015	$A4$: OPCAT ratification status (OHCHR).
M	1986–2016	$M1$: Institutional and legislative constraints [Lupu, 2015] measured by political constraints index (Polcon iii) [Henisz, 2002].
	1986–2012	$M2$: Judiciary effectiveness [Powell and Staton, 2009, Conrad, 2013] measured by judicial independence index [Linzer and Staton, 2015].
	1986–2007	$M3$: Political mobilization [Murdie and Davis, 2012, Simmons, 2009] measured by naming and shaming index [Cole, 2015].
	1986–2008	$M4$: International socialization [Clark, 2013, Goodman and Jinks, 2013] measured by treaty commitment preferences [Lupu, 2016].
Y	1986–2011	$Y1$: CIRI torture index [Cingranelli et al., 2013].
	1986–2013	$Y2$: Human rights protection scores [Fariss, 2014].
	1986–2015	$Y3$: Political Terror Scale [Gibney et al., 2016].
W	1986–2015	Population size [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators.
	1986–2015	Gross domestic product (GDP) per capita [Hafner-Burton and Tsutsui, 2007] measured by the World Bank Indicators.
	1986–2015	Participation in international trade [Hafner-Burton, 2013] measured by the World Bank Indicators.
	1986–2015	Regime types [Hathaway, 2007, Chapman and Chaudoin, 2013, Neumayer, 2007] measured by Polity scores [Marshall Monty et al., 2016].
	1986–2015	Regime durability [Goodliffe and Hawkins, 2006] measured by age in current regime [Cheibub et al., 2010].
	1986–2015	Involvement in militarized interstate disputes [Chapman and Chaudoin, 2013] measured by MID dataset [Themnér, 2014].
X		Legal origin [Mitchell et al., 2013] measured by legal origins data [La Porta et al., 2008].
		Treaty ratification rule [Simmons, 2009] measured by ratification rules dataset [Simmons, 2009].
		Electoral system [Cingranelli and Filippov, 2010] measured by database of political institutions [Cruz and Scartascini, 2016].

population. The difference of the two expected values is the average causal effect of participating in a monitoring procedure $\tau = E_{P_O}\big[Y|do(A=1)\big] - E_{P_O}\big[Y|do(A=0)\big]$.

In summary, to compute the average causal effect of a monitoring procedure, we would not simply observe the treatment values that nature generates. Rather,

we would need to intervene to disable the treatment assignment mechanism, in my structural causal model, for instance, the treatment mechanism that generates the ratification status for the state reporting procedure is the equation $A1_t = f_{A1}(X, A1_{t-1}, M_{t-1}, Y_{t-1}, W_t, U_{A1})$, fix the treatment values at $A1 = a$ for $a = \{0, 1\}$, and then predict the outcome values under these two different interventions to compute the average causal effect.

4.3.2 Causal identification

The question of causal identification arises when we want to establish the conditions under which we can translate and compute an interventional query from an observational probability distribution. This translation is made on a transparent basis using a graphical causal model in the form of a directed acyclic graph (DAG). The DAG in Figure 4.3 is a graphical representation of the structural causal model in Equation set 4.1. A directed acyclic graph [Pearl, 2000, Koller and Friedman, 2009] comprises of nodes/vertices denoting random variables and edges/arrows denoting one variable's direct causal influence on another variable. A path in a DAG is an arrow or a sequence of directed arrows, regardless of their directions, that connects one node to another. A path between two nodes that consists of arrows of the same direction is a causal path. Otherwise, it is a non-causal path. An acyclic graph contains no cycle or feedback loop, meaning that no node in the graph can have a causal path leading to itself.

Identification of causal effects is dependent upon the causal structure, which is represented by the topology of a causal graph. Thus, any causal model should be justified on the basis of the background knowledge to maximize the chance that it accurately captures the true data-generating process. In other words, we build on the existing literature to satisfy the assumption that the structural and graphical causal model have a one-to-one relationship and consistency with the underlying

probability distribution. The causal model in Figure 4.3 is constructed based on the following justification.

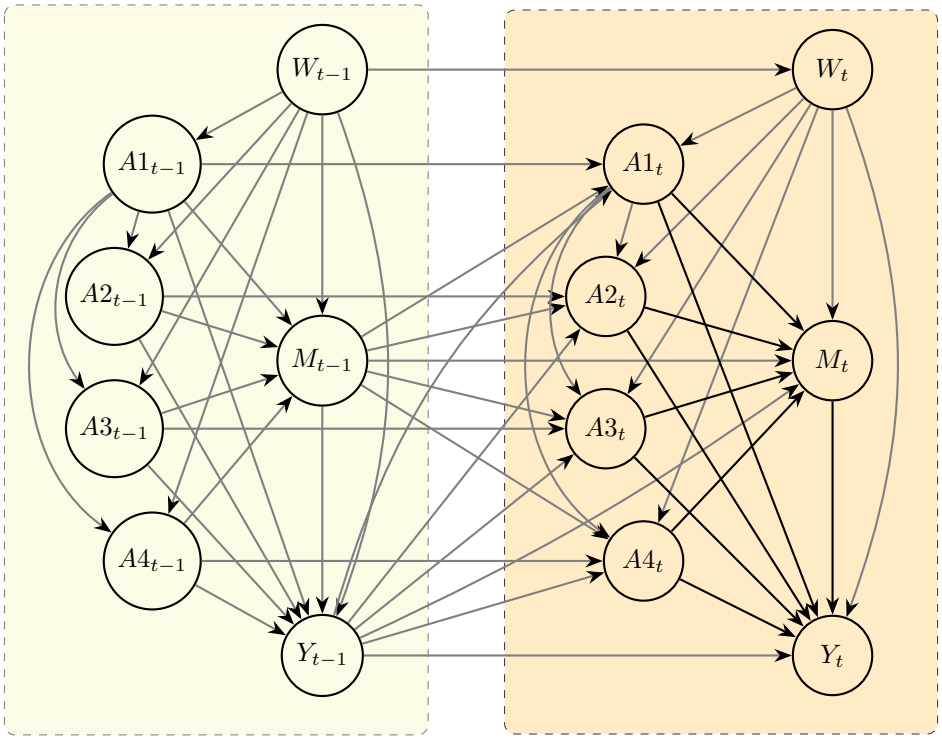


Fig. 4.3.: A causal DAG representing the causal process from treaty monitoring procedures $A1$ (state reporting under Art. 19), $A2$ (inquiry under Art. 20), $A3$ (individual communication under Art. 22), and $A4$ (country visit under the OPCAT) to Y (human rights outcome) with time-varying confounders W. Time-invariant confounders X, which precede and potentially affect all time-varying covariates, are not represented. All exogenous variables U's are assumed to be jointly independent and are not represented. Two shaded blocks indicate two time periods. The conditioning set $Z_{A1} = \{X, A1_{t-1}, M_{t-1}, Y_{t-1}, W_t\}$ is sufficient to identify the causal effect of $A1_t$ on Y_t. Similarly, the conditioning sets $Z_{A2} = \{X, A2_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t\}$, $Z_{A3} = \{X, A3_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t\}$, and $Z_{A4} = \{X, A4_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t\}$ are sufficient to, respectively, identify the causal effect of $A2_t$, $A3_t$, and $A4_t$ on Y_t.

First, the causal graph represents both the contemporaneous direct effect of each monitoring procedure on human rights outcome ($A_t \rightarrow Y_t$) and their contemporaneous indirect effects that go through all the mediators ($A_t \rightarrow M_t \rightarrow Y_t$). I make one key assumption with respect to the relationships between four treaty monitoring procedures. In order to participate in any optional monitoring procedures

(inquiry, individual communication, and country visit), a state has to be a party to the CAT in the first place. Since we have no reasons to believe in any particular causal order between the three optional monitoring procedures, I assume they are not directly causally related to each other. In other words, within the same temporal period, $A2$, $A3$, and $A4$ are independent conditional on treaty ratification $A1$ and other covariates. This assumption of conditional independence among three optional monitoring procedures is necessary for causal effect identifiability. Otherwise, if these procedures mutually cause each other, we would not be able to identify their causal effects.

Second, to represent potential confounding by various time-varying covariates W, I include directed arrows from these covariates to all treatments ($W_t \rightarrow A_t$), mediators ($W_t \rightarrow M_t$), and outcome ($W_t \rightarrow Y_t$). For a clear and concise presentation, I do not represent time-invariant confounders X, but they are assumed to precede and affect all other variables in the model.

Third, I incorporate in the graphical model the selection effect argument [von Stein, 2005, Simmons and Hopkins, 2005]. This argument claims that mostly only those states that intend to comply in the first place would join a human rights institution and, as a result, selection into a human rights treaty or, similarly in this case, a treaty monitoring procedure would be potentially biased. I represent this argument about the potential effect of the lagged human rights outcome on the treatments, using the directed arrows $Y_{t-1} \rightarrow A1_t$, $Y_{t-1} \rightarrow A2_t$, $Y_{t-1} \rightarrow A3_t$, and $Y_{t-1} \rightarrow A4_t$.

Fourth, I further include the directed arrow $Y_{t-1} \rightarrow M_t$ to denote the causal effect of the lagged outcome variable on the time-varying mediators. These causal arrows reflect the possibility that the use of torture by the executive could threaten the opposition parties and weaken legislative constraints; intimidate the judges and undermine judicial independence of the court system; suppress social movements even while potentially provoking mass mobilization; and possibly prompt international censure, criticisms, and condemnation.

Fifth, to further enable the graphical causal model to capture a potentially complex reality, I add the directed arrows $M_{t-1} \rightarrow A1_t$, $M_{t-1} \rightarrow A2_t$, $M_{t-1} \rightarrow A3_t$, and $M_{t-1} \rightarrow A4_t$. Substantively, this means the lagged mediators might have a causal influence on the ratification/participation status of treaty monitoring procedures. This is meant to incorporate the finding that states may enter into reservations to certain human rights treaty provisions based in part on the effectiveness of their domestic judicial system [Hill, 2016a]. It is also based on the research that suggests countries may ratify human rights treaties due to mobilization by human rights NGOs [Simmons, 2009] or an entrenchment of human rights norms through international socialization [Finnemore and Sikkink, 1998].

Finally, given the time-series cross-sectional structure of the data, I allow the possibility that the lagged value of a variable has an influence on its current value. Thus, for every time-varying covariate I include a directed arrow such as $A1_{t-1} \rightarrow A1_t$, $Y_{t-1} \rightarrow Y_t$, and so forth.

In summary, all potential causal relationships between variables in the model are derived on the basis of the background knowledge in the literature and then represented by directed arrows in a graphical model. It should be noted that a directed arrow in a graphical causal model does not necessarily indicate an actual causal influence, but rather a possible one. Thus, including a directed arrow is synonymous to *not* making a causal assumption whereas a missing arrow is equivalent to assuming that a direct causal relationship is absent.

As a graphical representation, a causal DAG compactly represents the causal structure of the data generating process without making any assumptions about the forms of any generative functions f or the probability distribution of the exogenous variables $U = (U_X, U_W, U_A, U_M, U_Y)$ other than that these exogenous variables are assumed to be jointly independent. The causal graph also exhibits the invariance property [Pearl, 2009a, 30], according to which a node/variable is independent from its non-descendants (nodes that are not on a causal path from that variable) conditional on its parent nodes (nodes that have a directed arrow entering that

variable). This concept of invariance lies at the heart of identification using the backdoor criterion.

The backdoor criterion is used to determine non-parametrically the identifiability of a causal effect via covariate adjustment [Pearl et al., 2016, 61–66]. To identify the causal effect of $A1_t$ on Y_t, for example, the backdoor criterion requires a sufficient adjustment set that:

(a) blocks any non-causal paths between $A1_t$ and Y_t that have an arrow entering $A1_t$;

(b) leaves open all causal paths from $A1_t$ to Y_t; and

(c) creates no spurious paths when conditioning on a collider (a node that lies on a path between $A1_t$ and Y_t and has two arrows coming into it) or a descendant of a collider.

Sufficient adjustment sets for causal effect identification are derived based on the structure of the causal DAG in Figure 4.3. According to criterion (a), the set $Z_{A1} = \{X, A1_{t-1}, M_{t-1}, Y_{t-1}, W_t\}$ is sufficient to identify the causal effect of $A1_t$ on Y_t. To satisfy criterion (b), a sufficient adjustment set should *not* include any of the mediators M_t. It also should not include any interaction terms between a mediator and the treatment or a mediator and other covariates. Nor should the adjustment set include any of the optional monitoring procedures $\{A2_t, A3_t, A4_t\}$ when estimating the causal effect of $A1_t$. The reason is that these optional procedures are the child nodes of $A1_t$ since participation in any of these three procedures is legally premised on being a state party to the CAT ($A1$) in the first place. Criterion (c) is automatically satisfied since we do not have any colliders on any paths emanating from the treatment A_t to the outcome Y_t. Applying the same backdoor criterion, I derive three other adjustment sets $Z_{A2} = \{X, A2_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t\}$, $Z_{A3} = \{X, A3_{t-1}, M_{t-1}, Y_{t-1}, A_t, W_t\}$, and $Z_{A4} = \{X, A4_{t-1}, M_{t-1}, Y_{t-1}, A1_t, W_t\}$ to respectively identify the causal effects of the other three monitoring procedures.

The idea of the backdoor criterion is to find an adjustment set of covariates that blocks all non-causal paths (usually known as backdoor paths because they include arrows that enter the treatment variable/node) between the treatment and the outcome while leaving open the causal path from $A1_t$ to Y_t [Pearl, 2009a, 79–81]. When we condition on a sufficient adjustment set to make the treatment and the outcome (for example, $A1_t$ and Y_t) conditionally independent, we have effectively removed all non-causal paths between $A1_t$ and Y_t and any remaining association is evidence of a causal relationship. In short, a successful backdoor adjustment will render an interventional query $E(Y|do(A = a))$ equivalent to an observational query $E(Y|A = a)$. The causal effect $\tau = E_{P_O}\big[Y_t|do(A1_t = 1)\big] - E_{P_O}\big[Y_t|do(A1_t = 0)\big]$ then becomes estimable as $\psi(P_n) = E(Y_t|A_t = 1, Z_{A1} = z) - E(Y_t|A_t = 0, Z_{A1} = z)$.

4.3.3 Machine learning-based estimation

For causal effect estimation, I employ the machine learning-based targeted maximum likelihood estimator (TMLE) [van der Laan and Rose, 2011]. The estimation procedure has three steps. First, we fit an initial predictive outcome model $Q_n^0(A, Z) = E(Y|A, Z)$ and the predictive treatment model $g_n^0(Z) = P(A|Z)$ where Z is the sufficient adjustment set for causal identification. Second, we modify the initial model Q_n^0 into an updated outcome model Q_n^1 using the updating equation $\text{logit}(Q_n^1) = \text{logit}(Q_n^0) + \epsilon_n H_n$ and the "clever covariate" $H_n(A, Z) = \frac{I(A=1)}{g_n(A=1|Z)} - \frac{I(A=0)}{g_n(A=0|Z)}$. The coefficient ϵ_n is estimated from a separate logistic regression model $\text{logit}(Y) = \text{logit}(Q_n^0) + \epsilon_n H_n$. Finally, we plug in the two binary treatment values $a = \{1, 0\}$ into the updated model Q_n^1 to compute the causal effect estimate $\hat{\psi}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\big[Q_n^1(A_i = 1, Z_i) - Q_n^1(A_i = 0, Z_i)\big]$. Statistical uncertainties around the estimate are approximated by the variance of the efficient influence function [van der Laan and Rose, 2011].

The case for this targeted estimator is made in more details by van der Laan and Rose [2011, 101–118]. Suffice here to note that this estimator has many de-

sirable properties that few others can match. First, it is doubly robust, producing unbiased estimates if either the initial outcome model Q_n^0 or the treatment model g_n^0 is consistent. It is more robust than, for example, propensity score-based estimators, the consistency of which depends on the correct specification of the propensity score model. If both Q_n^0 and g_n^0 are consistent, the targeted maximum likelihood estimates are maximally precise.

Second, when using standard parametric regression models for effect estimation, researchers often make assumptions about the forms of functions that characterize the relationships between the variables. These functional form assumptions include, for example, linearity of parameters and additivity of covariate effects. Given that many social and political dynamics are non-linear, these assumptions are likely unwarranted. If a model is not correctly specified in terms of its functional form, the estimates will be biased. The TMLE method circumvents this limitation by incorporating a machine learning ensemble technique called Super Learner that adapts to the data and better approximates the true underlying functions [Polley and van der Laan, 2010].

Super Learner uses a collection of parametric (in our case, ordinary least squares linear regression and regularized linear regression with lasso), semi-parametric (generalized additive models and spline regression models), and non-parametric algorithms (random forest and gradient boosting). They are then assembled in a weighted combination with an individual weight for each algorithm that is proportionate to its cross-validated predictive performance. The use of cross-validation helps make sure that the algorithms should generalize well and avoid overfitting. This cross-validated combination of algorithms creates a hybrid and much more powerful predictive function, which is then used to build both the initial outcome model Q_n^0 and the treatment model g_n^0. Super Learner-based models are much more likely to approximate the true underlying data generating process and satisfy the assumption of correct model specification. Table 4.3 lists the algorithms I use for causal effect estimation.

Table 4.3.: Algorithms used in Super Learner-based targeted maximum likelihood estimation

Algorithm	*Description*
glm	Main-term generalized linear models.
glmnet	Regularized linear models with lasso penalty $\sum_{j=1}^{p}\|\beta_j\|$.
gam	Generalized additive models (deg. of polynomials = 2).
polymars	Adaptive polynomial spline regression.
randomForest	Random forest (ntree = 1,000).
xgboost	Extreme gradient boosting (ntree = 1,000, max depth = 4, eta = 0.1).

For ease of interpretation, all three measures of human rights outcome are rescaled into a bounded 0–1 range with zero indicating the worst torture practice and one indicating the best human rights record. To handle missing data, I conduct multiple imputation ($m = 5$), using the Amelia II program [Honaker et al., 2011], and combine the effect estimates from each imputed data set. Appendix C.3 provides the summary statistics of the data and Appendix C.4 summarizes the imputation process.

4.3.4 Results and interpretation

Machine learning-based TMLE estimates of the causal effect of participating in each of the four monitoring procedures are reported in Table 4.4. In the top panel of Table 4.4, the results indicate that across three different measures of human rights outcome, the country visit procedure is the only monitoring procedure that has a consistently significant causal impact in terms of reducing torture and improving government respect for physical integrity rights. Its causal effect ranges from a 1.2 percentage point increase in human rights protection score to a 4.6 percentage point decrease in political terror scale to a dramatic 11.8 percentage point reduction in CIRI torture index.

The CIRI torture index specifically measures the torture practice by state officials and private individuals at the instigation of the government. It is probably the most

Table 4.4.: Average causal effect point estimates and 95% CI of treaty monitoring procedures under the CAT and the OPCAT.

Procedure	PTS score	HR protection score	CIRI torture index
Super Learner-based Targeted Maximum Likelihood Estimation Influence Function-based CI with Multiple Imputation			
State reporting (Art. 19)	**−0.023**	0.003	**−0.165**
	[**−0.044, −0.001**]	[−0.006, 0.011]	[**−0.313, −0.018**]
Inquiry (Art. 20)	**0.058**	**0.007**	**−0.064**
	[**0.031, 0.084**]	[**0.004, 0.009**]	[**−0.087, −0.041**]
Individual complaint (Art. 22)	**0.173**	0.116	**0.144**
	[**0.142, 0.204**]	[−0.348, 0.580]	[**0.033, 0.256**]
Country visit (OPCAT)	**0.046**	**0.012**	**0.118**
	[**0.026, 0.065**]	[**0.009, 0.015**]	[**0.029, 0.207**]
Linear Models of Human Rights Outcome Least Square Estimation with Multiple Imputation			
State reporting (Art. 19)	−0.001	0.002	−0.030
	[−0.031, 0.028]	[−0.001, 0.005]	[−0.081, 0.020]
Inquiry (Art. 20)	**0.058**	**0.006**	0.046
	[**0.019, 0.098**]	[**0.001, 0.010**]	[−0.021, 0.113]
Individual complaint (Art. 22)	**0.049**	**0.007**	0.040
	[**0.006, 0.092**]	[**0.002, 0.012**]	[−0.032, 0.112]
Country visit (OPCAT)	0.014	0.003	0.036
	[−0.025, 0.053]	[−0.002, 0.007]	[−0.032, 0.104]
Number of years (CAT)	29 [1987–2015]	27 [1987–2013]	25 [1987–2011]
Number of observations (CAT)	5,414	5,032	4,648
Number of years (OPCAT)	10 [2006–2015]	8 [2006–2013]	6 [2006–2011]
Number of observations (OPCAT)	1,929	1,547	1,163

targeted measure of the outcome. However, its limited timeframe of measurement leads to a smaller number of observations. I therefore use two other indicators that measure a larger variety of government abuses of physical integrity rights, including political imprisonment, extrajudicial execution, enforced disappearances, and other violations. As a result, the causal effect estimates naturally vary, but they nonetheless empirically confirm my theoretical argument about the causal impact of the country visit procedure under the OPCAT. Given the large number of deter-

minants of human rights outcome, some of which are highly resistant to meaningful changes, the finding that ratifying a single intrusive monitoring procedure can lead to a substantial improvement in human rights conditions is significant.

The next causally effective monitoring procedure is the individual communication mechanism under Art. 22 of the CAT. In terms of magnitude, its causal effect could be even greater, averaging at around a 15 percentage point improvement in human rights conditions. However, when we measure human rights outcome using the human rights protection score, its causal impact is no longer significant due to the large variation of the effect estimate. The causal effect of the inquiry procedure is smaller and inconsistent across different outcome measures. Counter-intuitively, but not without similar findings in the literature [Hill, 2010, Lupu, 2013a], the state reporting system, if anything, has a damaging impact on human rights protection. This surprising negative impact is even statistically significant, increasing torture by 16.5 percentage points when we measure state practices of torture using the CIRI torture index. It is not implausible that abusive governments may use participation in this relatively low-cost, almost symbolic reporting procedure as a cover for their domestic repression and to deflect international criticism.

In short, the monitoring procedures under the CAT and the OPCAT have substantially different effects on human rights outcome, ranging from a high of a 17.3 percentage point improvement to a low of a 16.5 percentage point decline in human rights conditions. Importantly, these findings provide the empirical evidence in support of the argument that intrusive procedures have a positive causal effect whereas the same claim cannot be made with respect to less intrusive ones. By implications, the findings suggest that one major way to improve human rights practices is to design and promote treaty monitoring procedures that are able to exercise intrusive oversight over state compliance. Among the five types of monitoring procedures available, the country visit procedure and, to a lesser extent, the individual complaint procedure represent the most effective protection mechanisms. Moreover, ongoing efforts to reform the reporting procedures of the UN human rights treaty

system should be directed toward building a more intrusive system. Otherwise, the current reporting system is unlikely to have any positive impact or even backfire.

To contrast the machine learning-based TMLE estimator with the conventional practice of effect estimation, I report in the bottom panel of Table 4.4 the effect estimates from linear regression models of human rights outcome. Covariate selection for causal effect identifiability remains the same. The results indicate that only in the case of the inquiry procedure with human rights outcome measured in the PTS scores and human rights protection scores are the effect estimates somewhat similar between the two estimation methods. In other cases, a simple linear regression model fails to produce any similar effect estimates. It is particularly off the mark with respect to the country visit procedure and when one uses the CIRI torture index to measure the outcome variable.

It is worth emphasizing that I include ordinary least squares linear regression in my user-selected library of Super Learner algorithms underlying the targeted maximum likelihood estimator. It means that if a linear regression model happened to accurately capture the underlying data-generating process, the Super Learner-based estimator would have recovered the same estimates and there would be no differences in results. This is an indication that the functional form assumptions of linearity and additivity are probably not appropriate in this case. More broadly, unless there are strong reasons to the contrary, one should employ estimators that can incorporate flexible machine learning methods to accommodate potentially nonlinear, complex data-generating processes and produce more robust estimates than is the case with parametric statistical models.

4.4 Conclusion

In one of the opening examples of this chapter, after Milan Ristic died of police brutality in Yugoslavia in February 1995, his father exhausted all domestic remedies and had to turn to the Committee against Torture under the CAT to demand for jus-

tice. Despite the state's argument to dismiss his case on the basis of inadmissibility and then to deny its responsibilities in its arguments on the merits, the Committee "finds that the State party has violated its obligations under articles 12 and 13 of the Convention to investigate promptly and effectively allegations of torture or severe police brutality." It also urged the state to "provide [...] an appropriate remedy" (Communication No. 113/1998), which was later ordered by the state's Supreme Court. This outcome was only possible because the individual complaints procedure under the CAT was designed to monitor state compliance more intrusively despite the state's protest to dismiss the case. It also underscores the greater efficacy of a relatively more intrusive treaty monitoring procedure in providing justice and extending government accountability.

The quantitative political science literature, however, rarely focuses on compliance monitoring mechanisms under the UN human rights treaties. Its focus on the issue of institutional design more generally is also mostly confined to institutions in international political economy. In this chapter, I bring to bear evidence from international institutions in the area of international human rights and examine the issue of institutional design and institutional impact from an explicit causal inference perspective. Specifically, I disaggregate treaty compliance monitoring into state participations in different monitoring procedures and address the question of whether these monitoring procedures have differing causal effects on human rights outcome because of their different designs. Answering this question has important implications.

First, it contributes to the larger body of literature that examines the empirical implications of institutional design for substantive policy outcomes [Downs et al., 1998, Koremenos, 2005, Gilligan and Johns, 2012, Abbott and Snidal, 2013]. Estimating the causal effects of four monitoring procedures under the CAT and the OP-CAT, my causal analysis suggests that human rights institutions that impose more binding obligations, operate according to more precise rules, and enjoy greater delegation of authority tend to be more causally effective and lead to better outcomes.

Second, in terms of human rights research, the contribution of this chapter involves examining the causal impact of an important human rights treaty from a different level of analysis. In the existing empirical literature, contradictory findings unfortunately abound when it comes to estimating the causal effect of the CAT ratification. One major reason could be that the CAT as well as other human rights treaties have mostly been examined at a more aggregate level. Instead of focusing on treaty ratification as a whole, I shift the investigative focus onto treaty monitoring processes. I then develop an argument as to why monitoring procedures may have differing causal impacts, arguing that the magnitude of their causal effect on human rights outcome is likely a function of the intrusiveness in their design.

Overall, the empirical evidence indicates that not all treaty monitoring procedures are created equal or have similar causal impact. Rather more intrusive procedures such as the country visit procedure may have the best ability to improve human rights conditions. My causal analysis also informs the current debate on how to reform the operation and improve the performance of the UN human rights treaty bodies. The research findings in this chapter favorably support more intensive monitoring and extensive information-gathering by international bodies and panels of independent experts. More broadly, this research is certainly only an initial step in evaluating and determining the kind of institutional design that works, has no causal impact, or even backfires in protecting human rights and improving government accountability. Further research is needed with respect to other human rights treaties as well as international institutions in other domains.

Methodologically, I employ the structural causal model framework that Pearl [2009a] and others have developed to address the question of causality in an observational setting. By putting the task of causal identification front and center, I attempt to endow a causal interpretation to the effect estimates on a transparent basis. Many, if not most, research questions in political science are indeed queries about cause and effect. Researchers should therefore openly embrace and employ a causality framework within which to conduct their research. In this chapter, I also

apply the machine learning-based targeted learning methodology for effect estimation. Its purpose is to relax the assumption of correct model specification, which is required in parametric statistical models. Because of its desirable properties, targeted learning is likely to become more popular and widely used across different fields of scientific inquiry in the future.

5. WHAT CAUSES STATE REPRESSION? A PREDICTION-BASED CAUSAL INQUIRY

5.1 Introduction

The previous two chapters have investigated the causal effects of human rights treaties and treaty monitoring procedures on state practices of torture and other human rights violations. In addition to international human rights law, however, the literature has also identified many other covariates that are shown to have a statistically significant relationship with human rights outcomes, ranging from demographic factors to economic indicators and from domestic institutional features to international variables. However, the way these covariates are identified could very well render their importance an artifact of statistical modeling and a product of a specific methodological approach. In this chapter, I re-examine from two different methodological perspectives the covariates that may also have an impact on state repression and human rights violations. The goal is to identify the covariates of state repression and human rights violations that are truly predictive and, more importantly, causally significant.

This chapter starts from the argument that empirical evaluations of different theories as to why governments violate human rights should go beyond the null hypothesis significance testing framework [Hill and Jones, 2014].[1] It is also sympathetic to the prediction-based approach, which argues that that empirical inquiries should focus more on estimating the predictive power of covariates that correlate with and potentially determine the levels of state repression rather than testing their statistical correlation with the outcome [Hill and Jones, 2014, 661]. An even

[1]For a recent discussion, debate, and proposed reforms of this inference framework, see Johnson [2013], Benjamin et al. [2017], McShane et al. [2017].

better and more substantively relevant approach, however, would be to incorporate algorithmic prediction into a causal analysis to estimate the causal effects of multiple determinants of state repression. The reason is that a causal analysis produces findings that have a causal interpretation and thus could form the empirical basis for policy interventions to effect desired changes in the outcome values, which in our case is to reduce state repression and prevent human rights violations.

In summary, explanatory variables of state repression that are identified through an application of the null hypothesis statistical significance testing approach are not necessarily predictive or causally relevant to the outcome. Predictive covariates that are identified through an application of the machine learning prediction-based approach are predictive, but not necessarily causally relevant. Finally, causal determinants of state repression that are identified through an application of the causal inference approach have a causal interpretation and thus are causally relevant and more substantively useful. Taking both the prediction-based approach and the causal inference approach, I replicate and extend the study by Hill and Jones [2014] in the following three directions to further learn about what predicts and what causes state repression and human rights violations.

First, I replicate part of the predictive analysis by Hill and Jones [2014]. Their study is motivated by the skepticism about existing empirical studies in the literature that mostly rely on statistically significant coefficients of covariates in statistical models. Hill and Jones [2014, 662] express their concern about the issue of overfitting. That is, researchers, following currently prevalent research practices, fit statistical models to all of their data and therefore "have no way of knowing if the patterns they uncover are the result of the peculiarities of a particular dataset or whether they are more general." As a result, the findings in the existing literature may not be generalizable and reliable.

A related problem is that, in the vast majority of the literature, researchers often rely on data models using a simple linear functional form that may or may not be able to capture the true data-generating process. If a model's functional form

does not capture the underlying data-generating process, statistically significant estimates could be biased. Hill and Jones [2014, 662] propose to address "this deficiency in the literature through the use of cross-validation and random forests."

My replication, however, shows that, for this particular human rights dataset, random forest, a popular machine learning technique that uses an ensemble of decision trees to make predictions with each decision tree using a randomly selected subset of covariates, is actually not an effective machine learning algorithm. I empirically demonstrate that extreme gradient boosting (XGBoost) is a better algorithm in terms of predictive accuracy for their dataset. The result of my replication is nevertheless relatively consistent with the original study although the specific ranks of covariates in terms of their predictive power are slightly different.

Second, I then take the causal inference approach and convert the predictive analysis into a causal analysis. For the purpose of making causal inference, I first build a causal model, making explicit the causal assumptions about the underlying data-generating system through a causal graph. I the embed the XGBoost prediction method into the structural causal model framework to estimate and compare the causal power, as opposed to the predictive power, of time-varying variables among the same set of predictive covariates. A causal analysis, it should be noted, is not only complementary to a predictive analysis. It is also more directly useful in terms of providing the basis for policy-making decisions.

My causal analysis suggests that boosting economic development, promoting international trade, intensifying shaming on the international media, and protecting the independence of the domestic judicial system likely represent some of the most impactful measures to reduce and prevent human rights violations. Overall, though, my causal analysis paints a bleaker picture, showing how persistent state repression can be and how much it can resist meaningful changes even if one could stage significant interventions on its causal determinants. By implications, it is highly unlikely that government violations of human rights could be prevented as a result of intervening on a single covariate or implementing a single policy change.

Third, when making causal inference without the benefits of randomization of the treatment values or exogenous variation in the values of the independent variable, researchers have to assume that there are no omitted variables that cause both the outcome and the independent variable of interest. This assumption, however, cannot be empirically verified based on a scrutiny of the observed data alone. This is because, fundamentally, the same joint probability distribution of the variables can be generated by different underlying causal processes [Pearl, 2009a, Peters et al., 2017]. As a result, one has to rely on domain expertise and sufficient subject matter knowledge to make and justify the assumption about the causal structure that governs how the underlying causal process transpires. Nevertheless, I present a heuristic to investigate the validity of this crucial assumption about the underlying causal structure, using the invariant causal prediction method [Peters et al., 2016]. This heuristic offers a practical approach to diagnosing the residuals of causal prediction, thereby lessening or boosting our confidence in the accuracy of the causal model. I then apply this heuristic to diagnose the causal model in the second section and find it to be sufficient.

5.2 Predictive Model of State Repression

5.2.1 Measures, metric, and models

The empirical question that Hill and Jones [2014] set out to investigate is rather straightforward: for a certain measure of state repression (the dependent variable) across countries and over time, which covariates are most important in terms of predicting repressive practices? To answer that question, a set of "theoretically informed covariates" [Hill and Jones, 2014, 668] and their measurements are gathered from the literature (Table 1 in the original paper). These covariates are chosen because they are often included in statistical models in previous studies either as the key independent variable or as control variables. In other words, they are be-

lieved to have a causal influence either on state repression or, if they are included as control variables, on both state repression and the independent variable of interest.[2]

Hill and Jones [2014] examine various measures of state repression. I focus on one measure in particular—the human rights protection latent score [Fariss, 2014]. This outcome measure has some nice properties, being a continuous measurement with relatively few missing values. It is also a model-based composite measure that incorporates a variety of other measures of human rights violations, including two major human rights datasets: the Political Terror Scale (PTS) [Gibney et al., 2015] and the Cingranelli-Richards (CIRI) indicators [Cingranelli et al., 2013]. Most importantly, the Fariss score accounts for the changing standards of accountability in human rights reports. This property is critical to an assumption we need later for the invariant causal prediction method, which is that the same generative causal model should work consistently across different time periods. If left unaccounted for, a gradual improvement in the accountability standards and in human rights information over time [Clark and Sikkink, 2013] would introduce a systematic bias in the measurement of human rights outcome and make our causal model empirically inconsistent across temporal environments.

I use the replication dataset of 2,096 country–year observations for 154 countries from 1982 to 1999 as well as the computer code that Hill and Jones [2014] provide to conduct model-based imputation of missing values.[3] To avoid overfitting

[2]In regression models, which remain the major workhorse in the literature for inferring cause-and-effect relationships, a covariate is included as a control variable only if it is believed to be a potential confounding factor that *causes* both the independent variable and the outcome. That they "are correlated with both state repression and the variable of interest" [Hill and Jones, 2014, footnote 2], however, is not a sound justification for their inclusion. When a covariate is a mediator or an intermediate variable that both causally follows the independent variable and has a causal influence on the outcome, its inclusion as a control variable results in a post-treatment bias. Additionally, if a covariate is directly or even indirectly caused by both the independent variable and the outcome variable, its inclusion leads to a collider bias. In both of these cases, this covariate would be correlated with both state repression and the independent variable of interest, but it should *not* be included as a control variable. Determining whether a covariate is a confounder or a mediator or a collider on which specific causal pathways is usually very difficult, if at all possible, since it requires concrete causal knowledge about the underlying data-generating system, which is a point that could be emphasized more often in the literature.

[3]A number of models in the original paper do not include the lagged dependent variable and thus use observations from 1981 to 1999. Similar to the authors, I do not impute missing values of the

and increase the generalizability of my prediction models, I perform 5-fold cross-validation (instead of 10-fold cross validation as in the original paper to reduce the computational cost). To quantify the variability of my estimates, I create 500 bootstrap datasets from a single imputed dataset (instead of 100 bootstrap datasets from each of the five imputed datasets as in the original study) and run an XGBoost algorithm on each of these 500 datasets to obtain the estimates. I then use the 2.5% and 97.5% quantiles of the bootstrap distribution of the estimates to create distribution-free 95% confidence intervals. This combination of non-parametric bootstrap and single stochastic imputation is shown to be valid for making efficient inference [Tsiatis, 2007, Daniel et al., 2011]. In terms of evaluating metrics, similar to part of the original study, I use the root-mean-squared error RMSE $= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ where y_i is the observed outcome value and $\hat{y}_i$ is the predicted outcome value for each country–year observation. This is a proper scoring metrics for comparatively evaluating various prediction algorithms for a continuous outcome variable.

Finally, I set up three baseline models as in the original study by Hill and Jones [2014]. The first baseline model has only two covariates: gross domestic product (GDP) per capita and population size. The second one additionally includes the binary covariate of civil war. The third baseline model has three covariates: GDP per capita, population size, and the lagged dependent variable. I further implement a fourth baseline model that is analogous to the jackknife regression analysis, iteratively dropping a covariate rather than an observation, I . That is, to measure the predictive power of each covariate, I use a baseline model that includes all other covariates, but not the lagged dependent variable. For each of the baseline models, I iteratively add each of the predictive covariates, computing the new RMSE as a ratio of the baseline model's RMSE. A ratio smaller than one suggests that the additional covariate increases the baseline model's predictive performance. Smaller ratios indicate greater power of the covariates in predicting state repression. I use the RMSE reduction ratio to measure the predictive power of covariates because it

lagged dependent variable (the first year for each country). Thus, the time period covered in the replication dataset that I use with the lagged dependent variable is from 1982 to 1999.

is more straightforward and easily interpretable than the decision rule used in the original paper.[4]

5.2.2 Predictive algorithms

My predictive analysis uses the replication dataset from Hill and Jones [2014]. The set of predictive covariates are essentially "fixed" as well in the sense that they are chosen on the basis of some theoretical justification and the domain knowledge in the existing literature. As a result, the performance of predictive models now mostly depends on how closely the algorithms we use are able to approximate the unknown, underlying function that generates the measured outcome (state repression). I examine a variety of different predictive algorithms and comparatively evaluate their performance, using the Super Learner prediction function [van der Laan et al., 2007, Polley and van der Laan, 2010]. Super Learner computes the 10-fold cross-validated mean-squared error (MSE) when each algorithm is used to predict the human rights protection score as a function of all predictive covariates and the lagged dependent variable.

Table 5.1 lists the algorithms I use for this comparative analysis. This list covers a diverse array of algorithms that make different trade-offs between interpretability and complexity and between bias and variance [James et al., 2013]. In addition to the ordinary least squares regression and conditional random forest that Hill and Jones [2014] use in their paper, I also include regularized regression with lasso, ridge regression, generalized additive models, local regression, polynomial spline regression, random forest, and extreme gradient boosting (XGBoost). These are some of the most commonly used algorithms in the machine learning and algorithmic predictions literature.

[4]The decision rule that Hill and Jones [2014, 670] use is "if the lower bound (the .025 quantile) of the prediction error [i.e., the RMSE in the case of a continuous outcome variable] for the model including that covariate is above the upper bound (the .975 quantile) of the prediction error for the baseline model, then the covariate is marginally important."

Table 5.1.: Algorithms used in Super Learner-based predictive analysis

Algorithm	*Description*
glm	OLS linear regression.
glmnet	Regularized linear regression with lasso penalty $\sum_{j=1}^{p}\|\beta_j\|$.
glm.ridge	Ridge regression with penalty $\sum_{j=1}^{p}\beta_j^2$.
gam	Generalized additive model (degree of polynomials = 2).
polymars	Polynomial multivariate adaptive regression splines.
loess	Local regression
randomForest	Random forest (ntree = 1,000).
cforest	Conditional random forest (ntree = 1,000).
xgboost (default)	Extreme gradient boosting (default hyper-parameters).
xgboost (tuned)	Extreme gradient boosting (fine-tuned hyper-parameters).

Of particular interest is the XGBoost algorithm [Chen and He, 2015, Chen and Guestrin, 2016], a faster and more efficient implementation of gradient boosting machine [Friedman, 2001, Natekin and Knoll, 2013]. XGBoost is an especially powerful, non-parametric ensemble method that is able to capture non-linear, interactive dynamics among the predictive covariates. I use a combination of 10-fold cross-validation and grid search to separately fine-tune the hyper-parameters of XGBoost (the number of trees, the tree depth, and the learning rate) to the imputed dataset. The results are reported in Figure 5.1. I then select the three best configurations of XGBoost learners and include them in my comparative analysis of algorithmic performance.

The evaluation result in Figure 5.2 shows that XGBoost is the most powerful predictive algorithm for this replication dataset. Generalized additive model and regularized regression, including both lasso and ridge regression, come in a close second, followed by linear regression of varying degrees of flexibility. The random forest technique registers a surprisingly disappointing performance. This comparative evaluation suggests that the predictive accuracy in the analysis by Hill and Jones [2014] could be further improved, using a more effective supervised machine learning algorithm. I thus adopt the best configuration of XGBoost for my predictive analysis of state repression in the following section.

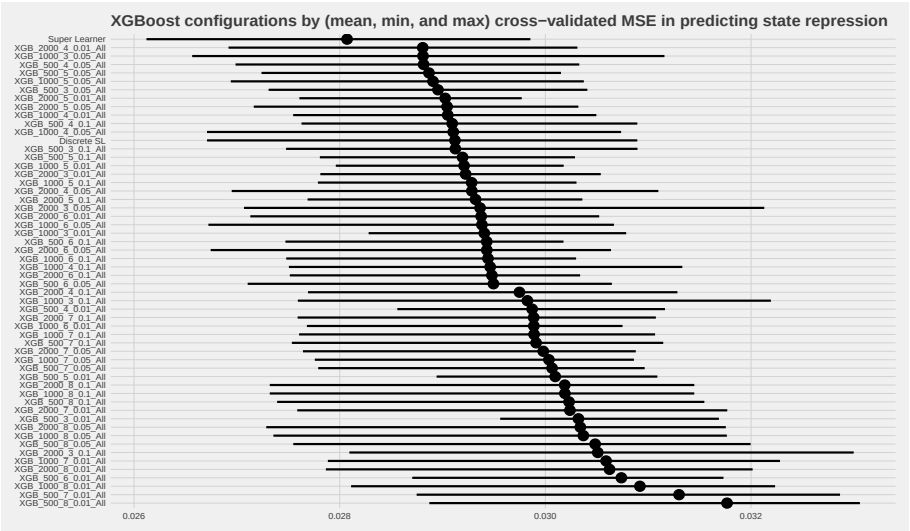


Fig. 5.1.: Predictive performance of different XGBoost configurations via ten-fold cross-validation with Super Learner.

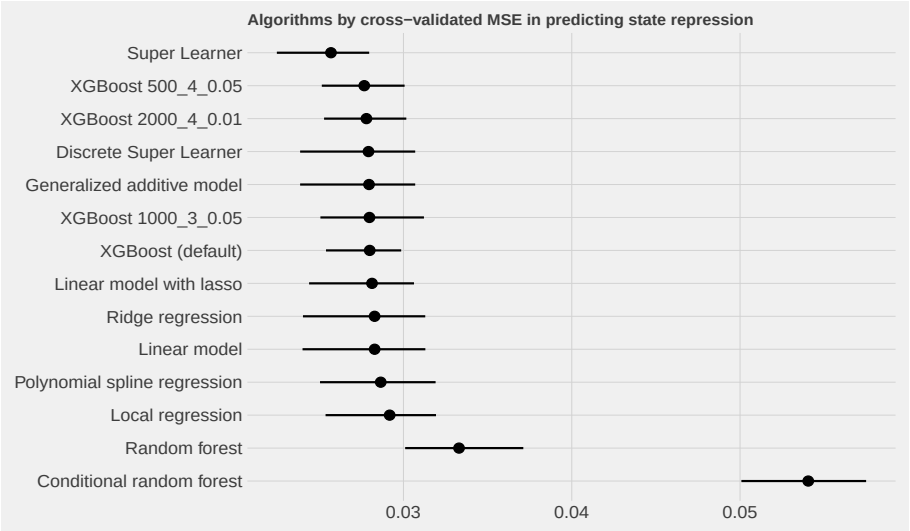


Fig. 5.2.: Predictive performance of different algorithms by ten-fold cross-validated MSE in predicting human rights protection score.

5.2.3 Predictive power of covariates

Figure 5.3 reports the predictive power of each covariate measured in terms of how much its inclusion reduces the RMSE of each of the four baseline models. First, for the baseline model that has only GDP per capita and population size, the most

important covariates, in descending order of predictive power, are the youth population, the number of international non-governmental organizations (NGOs) operating in the country, civil war, democracy, competitiveness of political participation, state reliance on oil revenues, constraints on the executive, and trade openness. Each of these covariates increases the predictive power of the baseline model by somewhere between 10% and 20% on average. Overall, this is roughly consistent with the findings by Hill and Jones [2014, 674] in their Figure 2 and especially Figure 7 although the ranks of specific covariates are slightly different.

Second, when the baseline model further includes civil war, the RMSE reduction when adding each covariate becomes smaller, indicating lesser predictive power of additional variables. In addition to democracy and its three individual components (constraints on the executive, competitiveness of political participation, and openness in executive recruitment), only three other variables can reduce the RMSE of the baseline model by more than 10% on average: the youth population, international NGOs, and state reliance on oil resources. Nevertheless, the most predictive covariates are relatively consistent across the first two predictive analyses.

Third, once I have the lagged dependent variable in the baseline model, the predictive performance of this baseline model does not change very much when additional covariates are included. In fact, only two covariates significantly reduce the RMSE of the third baseline model: the youth population and the number of international NGOs. Although Hill and Jones [2014] do not conduct a predictive analysis with the third baseline model when state repression is measured in human rights protection score, they reach a similar conclusion that the lagged dependent variable basically "dampens the predictive power that other covariates add to the model" and they interpret this finding as supporting the argument that "the governments can become habituated to the use of violence to resolve political conflict" [Hill and Jones, 2014, 674].

Finally, the "jackknife" predictive models that I introduce further underscore the limited power of individual covariates in predicting state repression. The only vari-

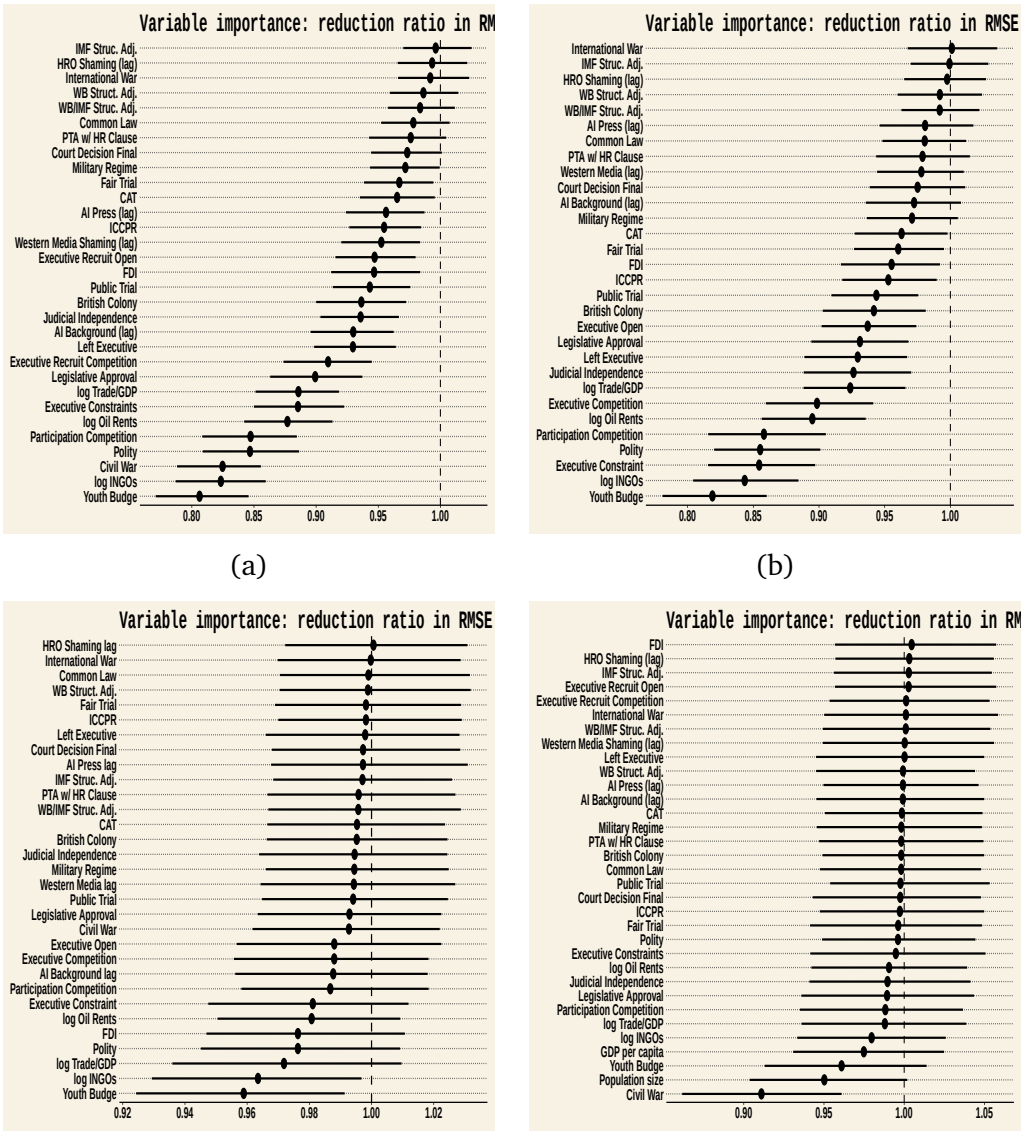


Fig. 5.3.: Predictive power of covariates measured by the reduction ratios from the baseline model's root-mean-squared error (RMSE) with bootstrap-based (B = 500) 95% confidence intervals when a covariate is added to the baseline model. The predictive algorithm is XGBoost with fine-tuned hyper-parameters. The predicted outcome is state repression measured in Fariss human rights protection score. Four different baseline models: (a) GDP per capita and population size; (b) GDP per capita, population size, and civil war; (c) GDP per capita, population size, and lagged dependent variable; and (d) "jackknife" baseline model with all covariates except for the one whose predictive power is being estimated.

able that has any significant predictive power is civil war, reducing the RMSE of its corresponding baseline model by about 9%. No other covariates consistently adds any predictive power even though the youth population comes in a close second, increasing its baseline model's predictive performance by about 5% on average.

In summary, based on the results across different predictive models, one can roughly divide covariates predicting state repression into three tiers of predictive importance. The first tier includes the youth population, the number of international NGOs, and possibly civil war. The second tier includes democracy and its various components, GDP per capita, and trade openness. The rest of the covariates, including constitutional settings and other time-invariant covariates, are not reliably significant in predicting state repression. It should be reiterated that, other than Hill and Jones [2014], few other studies have adopted the algorithmic modeling approach [Breiman, 2001a] to examine the predictive power of covariates that correlate with state repression. By employing a more effective algorithm to improve upon the study by Hill and Jones [2014], my replication presents findings about important predictors of state repression that are likely more accurate, informative, and reliable.

Substantial changes in the predictive power of covariates in the presence of the lagged dependent variable or when all other covariates are already in the baseline model are not particularly surprising. If there is indeed a self-perpetuating dynamics in state repression, the prediction results suggest that the causal impact of these covariates could vary, perhaps significantly, from their predictive power. More broadly, there are some reservations with respect to the adequacy of a predictive analysis of state repression.

First, unlike coefficient estimates from a linear regression model, for example, measures of variable importance in terms of cross-validated reduction in RMSE or in marginal permutation importance [Hill and Jones, 2014, 668–669] do not have an immediately clear substantive interpretation. It is not obvious how the predictive power estimates can be interpreted in terms of concrete real-world implications.

They may facilitate a thoughtful discussion as Hill and Jones [2014, 677] have about the measurements of certain underlying theoretical constructs and the implications for research, but measures of predictive power remain of limited utility. Most importantly, predictive power does not necessarily imply any useful insights about possible interventions for desired changes in the outcome, which is ultimately what we care about in our scientific research.

Second, the most important reason a predictive analysis is not adequate is that its findings do not have a causal interpretation and thus cannot provide a directly useful basis for policy-making decisions. For that, one has to transition from a predictive analysis to a causal analysis, predicting how the outcome would change if we could *intervene* on a predictive covariate. In other words, the object of inquiry is not the observational distribution of the outcome but rather its interventional distribution. Aside from other practical concerns about policy implementation, any proposed policy change or intervention has to address the question as to how much, for example, state repression will be reduced if the values of a covariate are intervened upon and externally modified. My causal analysis provides some answers to that question by estimating the causal effects of all time-varying predictive covariates when they are switched from their observed minimum values to their observed maximum values.

5.3 Causal Model of State Repression

5.3.1 Model formulation and causal identification

Estimating causal effects from observational data always assumes some concrete knowledge about the underlying causal system, which I will make explicit in the form of a directed acyclic graph. To formulate a model of this system, I first categorize the large number of predictors in Hill and Jones [2014, 670] into a set of time-invariant covariates W and another set of time-varying covariates. I focus on the latter set for the practical reason that they tend to be more amenable to a

policy change or intervention. Among the time-varying covariates, I further divide them into contemporaneous predictors A_t and the lagged predictors X_{t-1} as in the original study. The predictors X_{t-1} include the number of Amnesty International background reports, Amnesty International press releases, Western media shaming, and human rights NGOs shaming. These variables "are all lagged by one year" [Hill and Jones, 2014, 670] based on the justification that international shaming might possibly have an impact on state practices in the following year.

Following the recommendations by Hill and Jones [2014, 676–679], I omit the civil war covariate, the Polity measure of democracy, and its two competition component indicators due to their measurement issues. Specifically, measures of the concept of civil war likely pick up noncombatant casualties from the use of lethal violence, thus overlapping with the outcome measure of state repression. Measures of the competition components in the Polity dataset also overlap with measures of government repression as well [Hill, 2016b]. Including these variables would lead to a partially tautological causal model. The entire set of covariates in my causal analysis are summarized in Table 5.2.

A structural causal model that purports to represent an underlying causal system expressed is simply a collection of functions that generate the values of the variables in the model [Peters et al., 2017, 33–39]. I then further express the structural model using the directed acyclic graph (DAG) in Figure 5.4. In a DAG, each variable is denoted by a node and its values are strictly a function of its parent nodes and an error term where the parent nodes are the variables that have a direct causal influence on it [Elwert, 2013, Pearl, 2009a]. A direct causal influence is graphically represented by a direct arrow or edge from the parent node to the child node. DAGs are also acyclic because within the same temporal period, there are no directed paths (arrows or a sequence of arrows that have the same direction) going from one node to itself.

Table 5.2.: State repression model variables

Sets	*Variables from Hill and Jones [2014]*
W	Constitutional provisions for a fair trial. Constitutional provisions for a public trial. Constitutional provisions for final decisions by constitutional courts. Constitutional provisions for legislative approval of liberties suspension. Common law legal system. Former British colony.
A_t	*Demographics*: Population size. Youth population. *Macroeconomic factors*: GDP per capita. Oil revenue per capita. *Violent conflict*: International war. *Political institutions*: Military regime. Left/right regime. Executive constraints. Executive recruitment openness. *Domestic legal institutions*: *De facto* judicial independence. *International law*: ICCPR ratification. CAT ratification. *International economic factors*: Trade openness. Foreign direct investment. Structural adjustment programs (World Bank and IMF). Preferential trade agreements with human rights clauses. *Civil society/INGOs*: INGO presence.
X_{t-1}	*Civil society/INGOs*: Amnesty International background reports (lagged). Amnesty International press releases (lagged). Western media shaming (lagged). Human rights organization shaming (lagged).
Y_t	Human rights protection latent score [Fariss, 2014].

In our case, for example, the state repression variable at time t is Y_t and its values are generated as a function of its parent nodes PA_{Y_t} and an error term. That is, the generative function for the outcome variable is $Y_t = f_Y(PA_{Y_t}, U_Y)$. The set of parent nodes of Y_t are $PA_{Y_t} = \{W, Y_{t-1}, X1_{t-1}, X2_{t-1}, A1_t, A2_t\}$ where W are time-invariant covariates, Y_{t-1} is the lagged dependent variable, and all $X1_{t-1}$, $X2_{t-1}$, $A1_t$, and $A2_t$ are the lagged and contemporaneous covariates. To create a clear causal graph, I do not include the nodes W and, without loss of generality, I only represent two lagged predictors ($X1$ and $X2$) and two contemporaneous predictors ($A1$ and $A2$). The generative function f_Y as well as the generative functions for all other variables are assumed to work consistently across temporal environments.

While I make no assumptions about the form of any generative functions f's for any variables, my causal model nonetheless makes several assumptions as follows. First, I assume the underlying joint probability distribution of state repression and its predictors are Markov and faithful to the causal DAG, ensuring a one-to-one interchangeability and consistency between the distribution and the causal graph [Peters et al., 2017, 101–109]. The Markov and faithfulness assumptions mean that any conditional independencies in the probability distribution are encoded in the DAG and vice versa. As a result, conditional independencies in the distribution can be read off directly from the DAG, using the concept of d-separation [Pearl et al., 2016, 45-48]. For example, conditional on the set of its parent nodes $PA_{Y_t} = \{W, Y_{t-1}, X1_{t-1}, X2_{t-1}, A1_t, A2_t\}$, Y_t is independent from its non-descendant nodes such as $A1_{t-1}$, $A2_{t-1}$, and Y_{t-2}, that is, those variables that do not lie on a directed path emanating from Y_t. The importance of the Markov and faithfulness assumptions is that they connect the underlying distribution with the DAG and allow us to use graphical tools such as the backdoor criterion to establish identifiability and conduct efficient computation of causal effects.

Second, I assume the error terms of the generative functions in my structural causal model are jointly independent. Graphically, it means, for example, there are no hidden nodes that simultaneously cause the dependent variable Y_t and any of

the predictive covariates X_{t-1} and A_t. The assumption of joint independence of the error terms is equivalent to the no omitted variable bias assumption, which can only be justified based on the domain knowledge in the current literature. Absent some form of randomization, there are no empirical methods that can use observational data alone to guarantee the validity of this assumption. The role of causal graphs is to help make this assumption as well as its justification explicit and transparent.

Third, following Díaz et al. [2015, 6], I assume time-varying covariates during the same time period such as $A1_t$ and $A2_t$ do not mutually cause each other. This assumption is necessary because otherwise no causal effects would be identifiable. This assumption is not as restrictive as it might seem since each time-varying covariate is allowed to causally influence every other time-varying covariate in the next time period. For example, trade openness is assumed not to directly cause foreign direct investment (FDI) during the same year but it could easily increase or decrease the net flow of FDI in the following year.

To avoid cluttering the causal graph, I consider but do not represent the edges $X1_{t-1} \rightarrow X2_t$ and $X2_{t-1} \rightarrow X1_t$ in Figure 5.4. Adding these edges increases the flexibility of the causal model, but does not change any adjustment sets for causal identification. One should not assume that these potential causal links do not exist. Substantively, these arrows allow for the possibility that, for instance, NGOs reporting of human rights violations at time $t-1$ could be the basis for media shaming and Amnesty International's advocacy at time t.

Given these causal assumptions, I now briefly justify the topology of the causal DAG in Figure 5.4. In essence, a graphical model like this instills our knowledge and assumptions about the underlying, unobservable causal process. Most importantly, the directed arrows connecting the variables reflect our understanding about how the causal process unfolds. A description of these causal arrows in substantive terms therefore should look relatively reasonable and justifiable for people who are familiar with the literature.

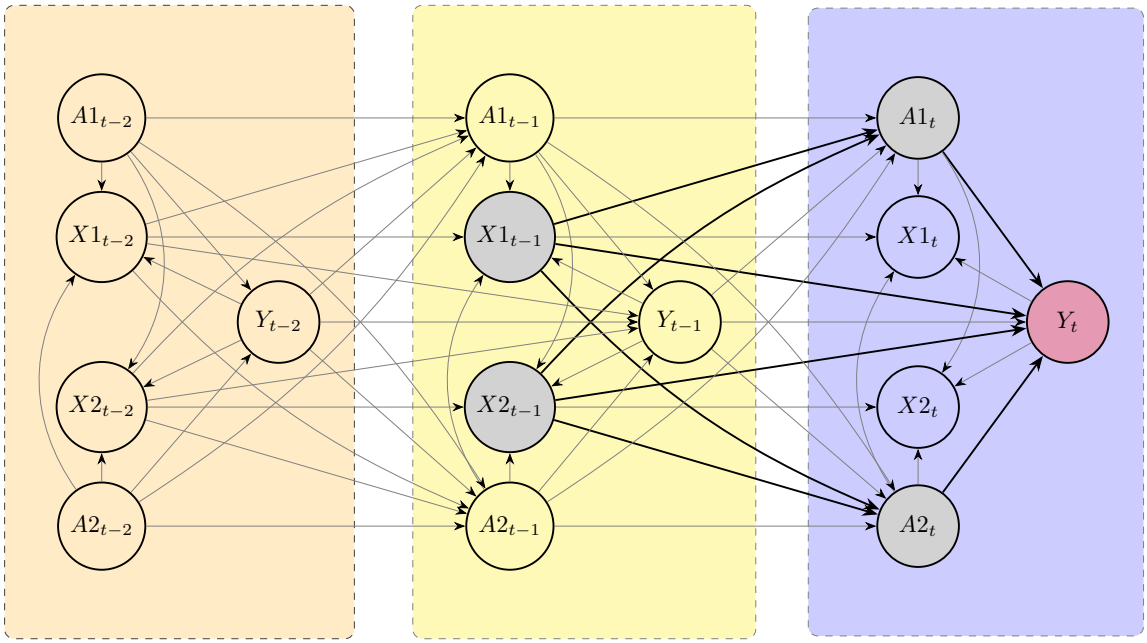


Fig. 5.4.: A dynamic graphical causal model with shaded blocks indicating temporal periods. Time-invariant covariates W, assumed to affect all other covariates, are not represented. The same sufficient adjustment set to identify both the causal effects of $A1_t \rightarrow Y_t$ and $A2_t \rightarrow Y_t$ is $Z_A = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X1_{t-1}, X2_{t-1}\}$. The sufficient adjustment sets to identify the causal effects of $X1_{t-1} \rightarrow Y_t$ and $X2_{t-1} \rightarrow Y_t$ are, respectively, $Z_{X1} = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X2_{t-1}\}$ and $Z_{X2} = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X1_{t-1}\}$.

First, given the panel data structure of the replication dataset, I allow the value of a time-varying covariate at time $t-1$ to affect its value at time t. This is represented by all the arrows $Y_{t-1} \rightarrow Y_t$, $A_{t-1} \rightarrow A_t$, $X_{t-1} \rightarrow X_t$, and so forth.

Second, state repression and human rights violations can have all kinds of effects on the predictive covariates in the following time period. Notationally, $Y_{t-1} \rightarrow A1_t$ and $Y_{t-1} \rightarrow A2_t$. State repression, for instance, could erode democracy, undermine domestic political and legal institutions, restrict the space and presence of civil society, and even lead to international economic sanctions and other repercussions. State repression could certainly provoke condemnations, criticisms, and international shaming during the same year as well, which is represented by the arrows $Y_t \rightarrow X1t$ and $Y_t \rightarrow X2_t$.

Third, the direct arrows from all A_t to all X_t reflect the recognition that international shaming by the media, civil society, and human rights NGOs in response to human rights violations does not happen in a vacuum. Rather, it could depend on the specificities of a particular country such as its political and legal institutions, macroeconomic conditions, and international economic factors. These arrows also represent the argument by Simmons [2009] and others that one of the key mechanisms through which the International Covenant for Civil and Political Rights (ICCPR) and the Convention against Torture (CAT) influence state behavior is by facilitating NGOs mobilization ($A_{t-1} \rightarrow X_{t-1}$) to pressure state officials for human rights protection ($X_{t-1} \rightarrow Y_t$).

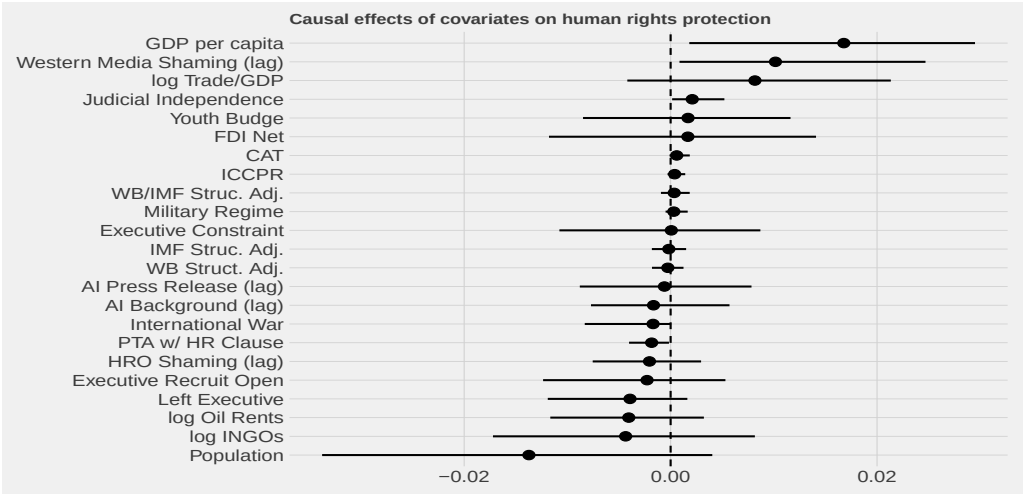
Finally, the causal effects I am estimating are those of time-varying covariates, both lagged ($X1_{t-1} \rightarrow Y_t$ and $X2_{t-1} \rightarrow Y_t$) and instantaneous ($A1_t \rightarrow Y_t$ and $A2_t \rightarrow Y_t$). To compute these causal effects, I intervene to set the value of each covariate at zero (its observed minimum value) and one (its observed maximum value) and compute the difference between the means of the two interventional outcome distributions. To do that with observational data, we need to translate the interventional distributions back into the observational distribution by making them essentially equivalent, using sufficient adjustment sets of covariates [Peters et al., 2017, 109–118]. These adjustment sets are identified by applying the backdoor criterion [Pearl et al., 2016, 61–66] to the graphical causal model in Figure 5.4. The causal identification via covariate adjustment suggests that the same adjustment set $Z_A = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X1_{t-1}, X2_{t-1}\}$ can be used to identify the causal effects of $A1_t \rightarrow Y_t$ and $A2_t \rightarrow Y_t$. Two separate adjustment sets sufficient to identify the causal effects of $X1_{t-1} \rightarrow Y_t$ and $X2_{t-1} \rightarrow Y_t$ are, respectively, $Z_{X1} = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X2_{t-1}\}$ and $Z_{X2} = \{W, Y_{t-1}, A1_{t-1}, A2_{t-1}, X1_{t-1}\}$.

5.3.2 Causal power of covariates

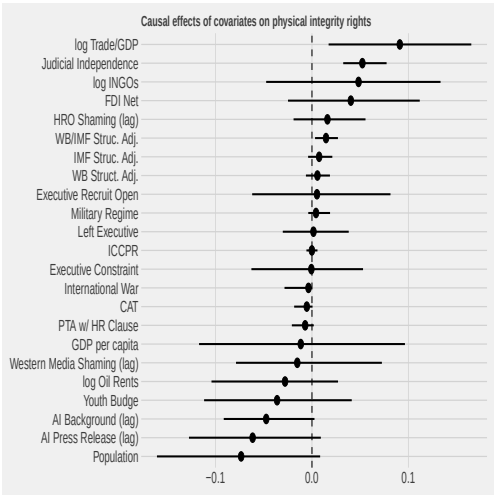
Once we have determined the sufficient adjustment sets for identifying the causal effects, I use the substitution estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\big[Q_n(1,Z) - Q_n(0,Z)\big)\big]$ to compute the causal effect $\tau_A = E\big[Y_t|do(A_t=1)\big] - E\big[Y_t|do(A_t=0)\big]$ of a contemporaneous covariate A and the causal effect $\tau_X = E\big[Y_t|do(X_{t-1}=1)\big] - E\big[Y_t|do(X_{t-1}=0)\big]$ of a lagged covariate X [Robins, 1986, Robins et al., 1999]. The do-operator indicates an intervention to fix the value of a treatment. In this estimator, Q_n is a predictive model of state repression and Z is the corresponding sufficient adjustment set. Key to the consistency of this estimator is the ability of Q_n to closely approximate the function $Y_t = f_Y(\cdot)$ that generates the outcome values.

In terms of implementation, for each covariate X_{t-1} and A_t, I fit two XGBoost predictive model of state repression (with previously fine-tuned hyper-parameters), setting the value of the treatment covariate alternately at one and zero and compute the mean difference of the predicted outcomes. The major difference between using an XGBoost predictive model for causal effect estimation and for predictive power estimation previously is that this time, rather than an relatively arbitrary baseline model, I use the adjustment set Z that is deemed sufficient for the purpose of causal identification. If the causal assumptions are satisfied, this will enable us to give our effect estimates a causal interpretation. For variance estimation, I similarly generate nonparametric bootstrap datasets (B = 500) and derive the quantiles-based 95% confidence intervals. The causal effect estimates are graphically summarized in Figure 5.5.

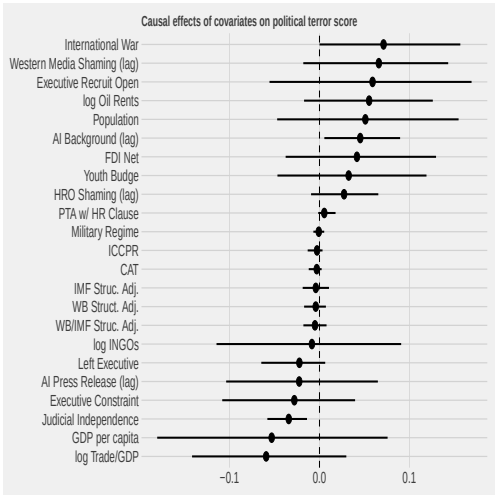
For ease of interpretation and comparison, I have standardized the human rights outcome measure into a bounded range between zero (lowest protection of human rights) and one (highest protection of human rights). The results indicate some divergence between the predictive power and the causal power of most covariates. While many covariates are predictive of state repression, at least marginally, most do not have any consequential causal impact at all. Only three variables have any



(a)



(b)



(c)

Fig. 5.5.: Causal effects of time-varying covariates on state repression when switching from their observed minimum to observed maximum values with bootstrap (B = 500) quantiles-based 95% confidence intervals. Outcome measures are standardized into a bounded 0–1 range, including (a) Fariss human rights protection score with higher scores indicating greater respect for human rights; (b) CIRI physical integrity rights with higher scores indicating better rights protection; and (c) Political Terror Scale with higher scores indicating worse repression.

significant causal power to improve human rights protection, including GDP per capita, international shaming on the Western media, and domestic judicial independence. Even then, their causal impact is quite limited. Each of them increases

human rights protection by less than two percentage points on average. This finding is certainly disappointing, but it is important to be clear-eyed about the heroic challenge of promoting and protecting human rights.

I apply the same estimation procedure to two other human rights outcome measures, including the CIRI indicators of physical integrity rights and the PTS scores. Emerging from this additional effect estimation is the marginally negative impact of international war and the positive causal influence of trade openness. Most important, though, is the consistent effect of judicial independence in improving protection of physical integrity rights by 5.2 percentage points and reducing the level of political terror by 3.4 percentage points on average. There are some similar findings in the literature about the role of an effective domestic court system in preventing torture [Powell and Staton, 2009], particularly the easily detectable forms of torture practices [Conrad et al., 2018], because of its ability to impose higher costs on abusive leaders.

In summary, the findings of my predictive and causal analyses seem to overlap somewhat with respect to the marginally important role of economic development and trade openness in improving human rights protection. Other than that, a causal analysis overturns some of the conclusions from previous predictive analyses about the importance of the youth population, international NGOs, and even some components of democracy such as constraints on the executive and openness of executive recruitment. Instead, judicial independence emerges as the most consistently impactful covariate. Permutation importance measures from random forest predictions in Hill and Jones [2014] back up this conclusion although their findings are not supposed to have any causal interpretation. The implications for the cause of human rights protection are nonetheless to focus on boosting economic development, promoting international trade participation, intensifying the scrutiny by the international media, and protecting the independence and effectiveness of the domestic court system.

5.4 Partial Diagnostics of Causal Model

Causal effect estimation using observational data requires prior causal knowledge about the underlying data-generating system, which can be assumed implicitly or represented explicitly in the form of a causal DAG. Either way, one of the questions that almost always arises is whether this causal knowledge is empirically accurate. More concretely, in the absence of a clearly exogenous variation in the predictor values, how can it be guaranteed that there is no omitted variable bias that could threaten causal inference? As I have claimed previously, there are simply no empirical methods that can properly address this concern and researchers have to rely on the domain knowledge to justify their models of the causal process. Matching methods, for example, often include a coterie of balance tests and diagnostic tools to assess how well the covariates are balanced across the treatment and control groups. However, one can only compare the balance before and after the matching process in the observed variables and still has to assume that all the unobserved covariates are somehow balanced as well.

Given the ultimately unverifiable nature of this causal assumption, in this section I nonetheless apply a heuristic to assess our confidence as to whether there are any missing causal determinants of state repression. Any confounders that can create omitted variable bias are necessarily a subset of this set of direct causes of state repression. Specifically, I rely on a recent methodological development in the causal discovery literature known as invariant causal prediction [Peters et al., 2016] to inform this diagnostics.

The invariant causal prediction method is inspired by a key insight of the structural causal model framework known as invariance [Pearl, 2009a, 22–26]. The invariance principle states that "the conditional distribution of the target variable of interest (which is often also termed the 'response variable'), given the complete set of corresponding direct causal predictors, must remain identical under interventions on variables other than the target variable itself" [Peters et al., 2016, 948].

For example, assuming a generative function for state repression $Y = f_Y(PA_Y, U_Y)$ that works consistently across different environments and an environmental variable E that is neither a parent nor a descendant of Y [Peters et al., 2016, 960], if the parental set (that is, the direct causes) of state repression PA_Y is complete—which implies no omitted variable bias—then the conditional independency $Y \perp\!\!\!\perp E|PA_Y$ should stay invariant across all environments.

The ingenuity of the invariant causal prediction method is that it exploits this invariance principle to test all possible sets of predictive covariates in order to derive sets of plausible causal predictors X_S that satisfy the invariance requirement $Y^e \perp\!\!\!\perp E|X_S^e$ for all $e \in E$. In essence, it reduces the problem of causal discovery to one of testing for statistical conditional independence across multiple sets of covariates X_S. Based on this method and making no assumptions about the functional form of the generative function f_Y, I use the time indices as the environmental variable [Heinze-Deml et al., 2017, 6] and XGBoost as a nonlinear prediction algorithm in the following diagnostic procedure [Pfister et al., 2017, 7].[5]

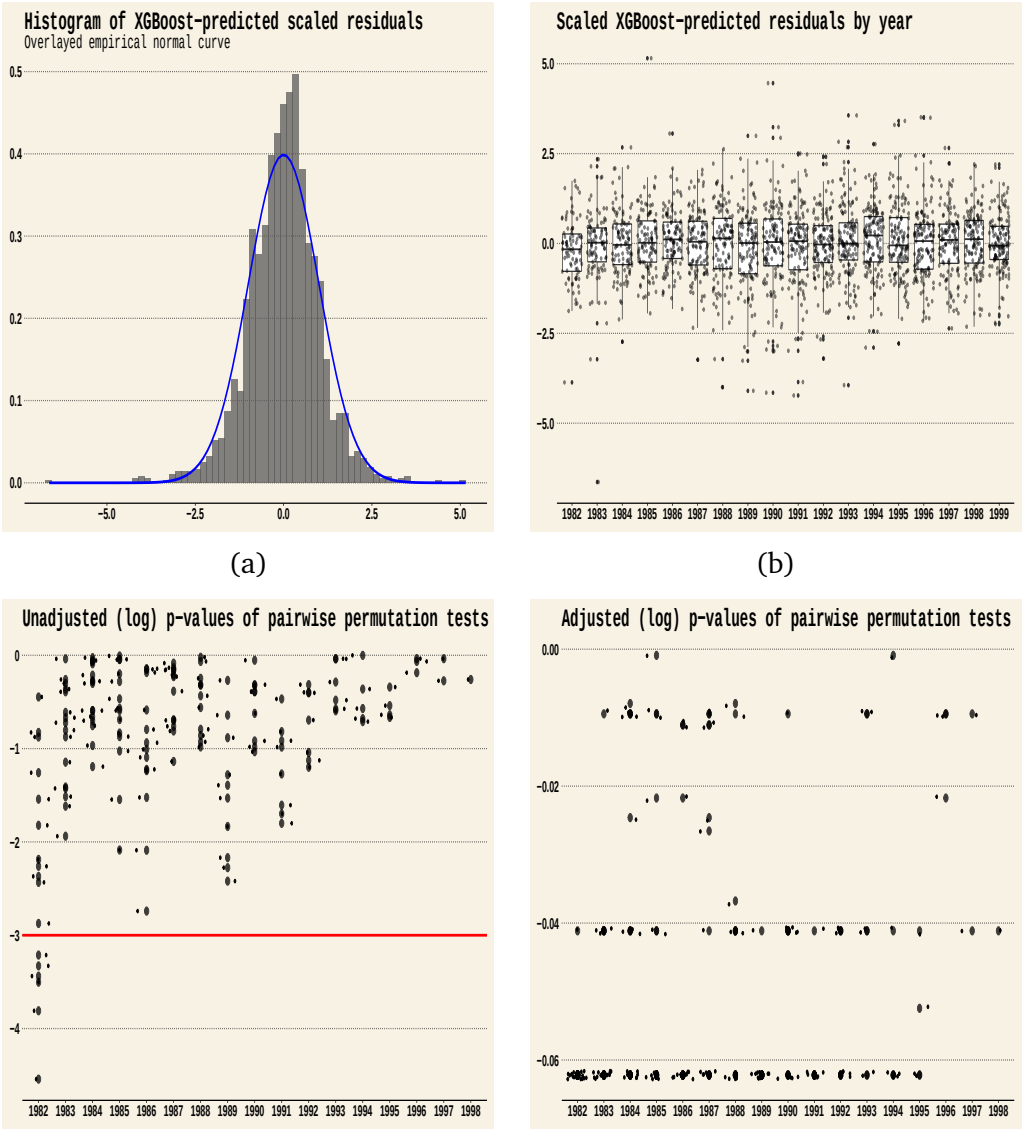
1. Use XGBoost to predict $\hat{Y}_t = f_Y(PA_{Y_t})$ from the pooled data where the theoretically informed parent set is $PA_{Y_t} = \{W, Y_{t-1}, X1_{t-1}, X2_{t-1}, A1_t, A2_t\}$.

2. Compute the scaled residuals $R_n = (r_1, \ldots, r_n)$. Assuming XGBoost approximates the generative function f_Y, R_n should be approximately independent and identically distributed.

3. Conduct pairwise permutation tests for independence between the temporal environment variable E (the time indices) and the scaled residuals R_n.

The diagnostic results are reported in Figure 5.6. They indicate that the residuals are roughly normally distributed with slightly more observations around the zero mean. Importantly, the pairwise permutation tests indicate a clear degree of

[5] I should reiterate my preference in favor of the Fariss human rights protection score as the outcome measure here for the reason that it accounts for the changing standards of human rights accountability and thus removes a potential correlation between the temporal environment variable E and the outcome Y.



(a) (b)

(c) (d)

Fig. 5.6.: Diagnostics of scaled residuals from causal invariant prediction of state repression measured in Fariss human rights protection score: (a) a histogram of scaled residuals from XGBoost causal prediction; (b) a box-plot of scaled residuals by year; (c) logged unadjusted p-values from pairwise permutation tests of XGBoost causal prediction residuals across temporal environments (years) with the horizontal line at $log(0.05) = -3$; and (d) logged adjusted p-values from pairwise permutation tests with Benjamini & Hochberg adjustment to control the false discovery rate.

independence between the environmental variable and the scaled residuals with only six unadjusted p-values smaller than 0.05 out of 153 comparisons. Once I use the Benjamini–Hochberg adjustment to control the false discovery rate [Benjamini and Hochberg, 1995], all adjusted p-values are significantly larger than the conventional threshold of statistical significance. This suggests that, conditional on the set of parental predictive covariates, the distribution of state repression does not change across temporal environments. We are therefore more confident that there are no omitted variables that causally influence state repression.

The invariant causal prediction method is certainly much more ambitious than my application. It aims to learn the underlying causal structure with a statistical guarantee. For our case, though, this is not possible just yet given that the problem of testing all possible subsets from the set of covariates theoretically identified as potentially direct causal factors of state repression is of $\mathcal{O}(2^p)$ complexity and exponential in computational time. The promise of learning the underlying causal structure from observational data remains extremely challenging to fulfill, but it also represents a huge potential for methodological advances [Pearl, 2009a, Spirtes and Zhang, 2016, Mooij et al., 2016, Eberhardt, 2017]. After all, uncovering the underlying causal structure is not just the inverse of the better known task of estimating causal effects. It actually goes to the foundation of making causal inference from observational data by providing an empirical basis for the assumed data-generating process that underlies any causal effect estimation.

5.5 Conclusion

This chapter revisits the question of what causes state repression from two methodological approaches: algorithmic prediction and causal inference. I first replicate part of a previous predictive analysis by Hill and Jones [2014], using the more recent and effective learning algorithm of XGBoost. The findings that this machine learning prediction-based approach produces suggests three key variables

that are generally most predictive of state repression, including the youth budge in the population, the number of human rights NGOs, and civil war. They are followed by democracy, economic development, and trade openness as the next tier of the most predictive covariates. The rest of other covariates under examination do not have much predictive power over state repression.

I then take the causal inference approach, converting my predictive analysis into a causal analysis and examining the causal determinants of state repression. The results of the causal analysis overlap somewhat with those of the predictive analysis, both of which underscore the critical role of economic development and international trade participation in reducing state repression. But there are also clear differences between the two analyses. The findings from the causal analysis describe a more challenging situation for human rights defenders and anyone who cares about preventing state repression. It also highlights the importance of an independent domestic court system as the most consistently significant factor in improving human rights protection. Only two other variables that have some causal power to reduce state repression are economic development and international shaming by human rights NGOs.

Methodologically, there are important tradeoffs between different approaches. The algorithmic prediction-based approach is more straightforward and does not need to make many assumptions about the underlying data-generating process. Its findings, however, are only suggestive at best and less applicable for the purpose of assisting policy-making decisions. The causal inference approach produces findings that have a causal interpretation and, by implications, are more directly useful. However, it has to rely on a set of causal assumptions, many of which are not empirically testable.

The third approach, lesser known in social sciences but also arguably very promising, is to examine questions like state repression from the perspective of causal discovery, particularly using the invariant causal prediction method. Unfortunately, this method is still not computationally feasible for my application. For now, how-

ever, I nonetheless rely on the insights from this method to develop a diagnostic procedure that can help evaluate whether a set of potential direct causes of state repression is complete and, by corollary, whether there is any potential omitted variable bias. Empirical researchers should keep a close watch on further developments in this method and, more generally, on advances in the field of causal discovery and causal structure learning.

6. CONCLUSION

For the last two decades, Judea Pearl's graph-based structural causal model (SCM) framework has gradually revolutionized his core field of artificial intelligence and computer science and had transformative ripple effects on many other disciplines such as philosophy, psychology, and epidemiology. My dissertation adopts and applies his monumental work on causality to international relations research. In each of the previous four empirical chapters, I combine machine learning and the SCM framework to investigate the causal determinants, causal mechanisms, and causal impacts of major United Nations (UN) human rights treaties as well as the causes of human rights violations. This conclusion chapter will briefly review the substantive premise of each empirical chapter; explain how the machine learning-based causal inference approach contributes to answering the research questions and provides new insights into the substantive debates in the literature; and summarize the benefits, tradeoffs, and implications of this new methodological approach.

The substantive premise of **Chapter 2** is the ongoing and still unresolved debate in the literature about why states commit to human rights treaties. Many theories of treaty ratification have been proposed, but political scientists have not yet come to any consensus as to which factors are most important in explaining human rights treaty ratification. Is it international socialization and the pressure of normative conformity? Do states commit to universal human rights treaties in exchange for foreign aid, international investment, and preferential trade agreements as the instrumental approach to treaty ratification suggests? Are democratic domestic institutions the most important factors? Furthermore, there are other theories that explain treaty ratification as an interactive function of democracy, prior treaty compliance records, and the existence of multiple political parties.

My research in this chapter weighs in on this unresolved debate in the literature. It applies a new methodological framework to conduct a causal variable importance test of three major theoretical approaches to treaty ratification. Because each theoretical approach proposes a different set of covariates as the most important determinants of treaty ratification, my test strategy is to directly assess, evaluate, and compare the causal effects of these theoretically informed covariates. Specifically, I employ the graph-based SCM framework and the ensemble prediction method of Super Learner to estimate the causal effects of multiple determinants of state ratifications in the case of three UN human rights treaties on civil and political rights (the ICCPR), women's rights (the CEDAW), and the right not to be tortured (the CAT).

The results tell us about the most important factors that cause states to ratify these three human rights treaties and provide an empirical basis to adjudicate among different theories of treaty ratification. The findings indicate that among the covariates, the density of regional ratification turns out to be the single most consistent and the second most causally important predictor across all three human rights treaties. These causal findings strongly support the norms-based theories of treaty ratification in the literature. In contrast, given the limited causal relevance of economic factors such as economic development, international trade participation, and official development assistance, my causal analysis casts doubt on the instrumental explanations that focus on the economic rationale of treaty ratification.

One of the other key results is that democracy is the most causally important variable for the ratification of the ICCPR and the CEDAW. In fact, having a democratic regime causes the probability of being a state party to the ICCPR and the CEDAW to go up by 23.7 and 11.6 percentage points, respectively. For the CAT ratification, however, it is not democracy per se that has a significant causal impact. Rather it is the *de facto* existence of multiple political parties that raises the probability of ratification by 19.2 percentage points on average. However, other domestic institutional factors such as regime transitions and judicial independence have no

significant causal impact on state commitment to human rights treaties. In addition, I estimate the causal effect of democracy conditional on prior compliance records as well as the causal effect of prior compliance records conditional on democratic regime. Contrary to many theoretical expectations in the literature, I find no significant negative causal impact by either of these two factors on treaty ratification. Overall, my causal analysis offers very mixed support for the institutional approach to treaty ratification.

In short, the machine learning-based SCM framework enables researchers to construct reasonable causal models based on sufficient background knowledge and then conduct robust causal effect estimation. My research in chapter 2 uses the ongoing debate about treaty commitment as its substantive premise, recasts theoretically predictive variables as causal determinants of treaty ratification, and applies a new methodological framework to flexibly estimate the causal effects of these variables. These causal effect estimates offer a direct test of multiple theories of treaty commitment from a causal inference perspective, advancing our understanding about the causes of human rights treaties.

Judea Pearl's causality framework has also transformed how research questions about mediation are defined and answered. In **Chapter 3**, I present a machine learning-based causal mediation analysis of the same three UN human rights treaties (ICCPR, CEDAW, and CAT) in the context of panel data structure, multiple causally connected mediators, and no functional form assumptions. The substantive premise of this chapter is a lack of any quantitative empirical evaluations of various theories in the literature about the causal pathways of human rights treaties. In fact, the existing literature has articulated multiple causal mechanisms that potentially transmit the effects of human rights treaties, including legislative constraints, domestic judicial litigation and enforcement, political mobilization of civil society organizations, and international socialization. Yet, there has never been a concrete quantification of how much the causal effect of human rights treaties is actually transmitted through these causal mechanisms.

My examination of the literature suggests that this research gap remains because human rights researchers have not taken advantage of recent advances in the causal inference literature to shed new lights on the causal mechanisms of human rights treaties. The causal mediation analysis in Chapter 3 fills in this gap and, to the best of my knowledge, is the first quantitative evaluation of the ways in which international human rights treaties constrain and influence state behaviors. Methodologically, the empirical strategy of this chapter combines (a) the mediation formula within the SCM framework, (b) the felicity of causal graphs in assisting causal identification, and (c) the flexibility of the Super Learner prediction method for robust estimation.

The substantive findings indicate that all three human rights treaties have a positive causal impact on human rights protection and promotion. Participating in the ICCPR reduces state violations of physical integrity rights by 13.6 percentage points on average while committing to the CAT leads to a more modest decrease of government's torture practices by about 7.7 percentage points. For the CEDAW, the average causal effect of treaty participation is more substantial, enhancing women's political empowerment by 22 percentage points on average.

However, there is something concerning about the *direct* effects of the ICCPR and the CAT, both of which actually lead to *more* torture and violations of physical integrity rights. If all four causal mediators do not change their values in response to treaty ratification, being a member of these two treaties exacerbates human rights violations by 0.8 and 3.4 percentage points, respectively. The good news is, at the same time each of these two treaties has a positive *indirect* causal effect that is both statistically significant and substantively larger, averaging about 14.4 and 11 percentage points. CEDAW participation, on the other hand, improves women's empowerment both directly and indirectly with its indirect causal impact being much more substantial. The four causal pathways that I examine are jointly responsible for transmitting roughly 90% of the CEDAW's total causal effect. Overall, my causal mediation analysis underscores the importance of the causal mediators in

transmitting human rights treaty effects. Without these causal mediators, which include legislative constraints, domestic judicial litigation and enforcement, civil society mobilization, and international socialization, all three human rights treaties under examination here would lose most, if not all, of their positive causal impact.

Chapter 4 is substantively premised upon an oversight in the human rights literature with respect to the issue of treaty compliance monitoring. While the quantitative literature on human rights treaties contains a lot of studies that examine the impact of UN human rights treaties as a whole, it rarely investigates the various forms and mechanisms of treaty compliance monitoring. I therefore focus on the UN treaty on torture and unpack the ongoing monitoring practices under this treaty into multiple monitoring procedures and compare the individual causal effect of each monitoring procedure on human rights outcome.

Methodologically, this chapter presents a straightforward application of the machine learning-based SCM framework to a set of relatively new human rights institutions in the quantitative empirical literature. Specifically, I use a causal graph to assist causal identification and the targeted learning methodology for robust estimation. Substantively, I estimate the causal effects of four treaty monitoring procedures under the Convention against Torture (CAT) and its Optional Protocol (OPCAT), including state reporting, inquiry, individual communication, and country visit. The results show that only the country visit procedure has a significant, positive causal impact on human rights protection. Other monitoring procedures do not. The differing causal effects, I argue, are a function of the varying intrusiveness among the treaty monitoring procedures.

The findings improve our understanding about the granular effectiveness of human rights treaties. Importantly, both my causal theory and the empirical findings suggest that more intrusive procedures tend to have a positive causal effect whereas other less intrusive procedures do not. In terms of implications, one key strategy to improve the efficacy of international human rights regimes is to design procedures that are able to exercise intrusive monitoring and oversight over state compliance.

Among the existing monitoring procedures, the country visit procedure and, to a lesser extent, the individual complaint procedure are probably the most effective protection mechanisms. Relatedly, ongoing efforts to reform the reporting procedures under the UN human rights treaty system should be directed towards designing more intrusive mechanisms. Otherwise, the current reporting system is unlikely to have any positive impact and maybe even backfires. The causal findings in this chapter also have important implications for larger body of literature on the relationship between institutional design and institutional impact, providing one more data point from a set of international institutions in the area of human rights.

Finally, **Chapter 5** bridges the gap between algorithmic prediction and causal inference in investigating the causes of state repression and human rights violations. The substantive premise is to find out which factors are most predictive of state violations of human rights, but also more substantively important, which factors are causally relevant in preventing state repression. The obvious implication is that the most causally important variables will be the best candidates to be intervened upon to reduce state repression and enhance human rights protection.

Based on that substantive premise, I first replicate part of a recent predictive analysis, perhaps the first one in the quantitative human rights literature. I use the demonstrably more effective machine learning algorithm of extreme gradient boosting to estimate the predictive power of covariates that, according to the literature, are associated with state repression. I then incorporate this prediction method into the SCM framework to evaluate and compare the causal effects of these same covariates. Finally, I present a new heuristic to partially diagnose my causal model of the underlying data-generating process.

The findings from the predictive analysis suggest three tiers of covariates in terms of their ability to predict state repression. The first one includes the youth population, the number of international NGOs, and possibly civil war. The second tier includes democracy and its various components, gross domestic product (GDP) per capita, and trade openness. The rest of the covariates that are examined, includ-

ing constitutional settings and other time-invariant covariates, are not significantly predictive of state repression.

There is a divergence, however, between the predictive power and the causal power of most covariates under examination. While many covariates are predictive of state repression, most do not have much causal impact. Only three of them have the causal power to improve human rights protection, including GDP per capita, international shaming on the Western media, and domestic judicial independence. Even then, their causal impact is quite limited with each of them increasing human rights protection by less than two percentage points on average.

I also apply the same causal effect estimation procedure, using two other measures of human rights outcome: the Cingranelli-Richards indicators of physical integrity rights and the Political Terror Scale scores. Emerging from these additional causal analyses is the marginally negative impact of international war and the positive causal influence of trade openness. Most important, though, is the consistent effect of judicial independence in improving protection of physical integrity rights by 5.2 percentage points and reducing the level of political terror by 3.4 percentage points on average. Overall, the results of both the predictive analysis and the causal analysis are generally discouraging for the cause of human rights protection, but they also identify the most impactful factors that can reduce and prevent state repression, most likely including economic development, domestic judicial independence, and the naming and shaming by human rights NGOs.

The previous four empirical chapters of this dissertation combine to demonstrate the great potential of a machine learning-based SCM framework. A more expanded research agenda going forward could be to investigate a series of substantive questions in international relations and political science, using this new methodological approach. Some of the questions I would like to revisit and investigate include:

- What are the causes of the freedom of the press and its consequences for the protection and promotion of other human rights?

- What are the causal determinants of state ratification and implementation of the UN treaty and protocols on transnational human trafficking?
- What are the causes of domestic judicial independence and its consequences for development and governance?; and
- What are the causes and (economic and political) consequences of international immigration?

To answer each of these questions and a countless number of other questions in political and social sciences is to engage in a study of causation. At the most fundamental and intuitive level, causation is defined as follows: a variable X is a cause of variable Y "if Y in any way relies on X for its values" [Pearl et al., 2016, 5]. Equivalently, if X is an input to the function that generates the values of variable Y, then X is a direct cause of Y and "X is a *cause* of Y if it is a direct cause of Y, or of any cause of Y" [Pearl et al., 2016, 26]. Adopting this definition of causation means that I subscribe to a functional theory of causation, according to which "causal relationships are expressed in the form of deterministic, *functional* equations" [Pearl, 2009a, 26], for example, $\{X \leftarrow U_X; Y \leftarrow f_y(X, U_Y)\}$ where the variables Us are exogenous variables and U_X is independent of U_Y.

A collection of functional assignment equations is called a structural causal model [Peters et al., 2017, 33–34] and its associated graphical representation is called a graphical causal model or simply a causal graph. The critical role of structural causal model and graphical causal model is to describe the causal reality, also known as the underlying causal structure, as we understand it. Causal structure is the ultimate basis of any causal study because to study the causal effect is to study the effect (or consequence) of an intervention on the causal structure.

The concept of intervention is most clearly understood in the context of a randomized controlled experiment where intervention is the act of assigning the treatment values and the causal analysis involves observing and analyzing what happens under that intervention. In observational studies, there is no actual intervention and

all the researchers have at their disposal is the observed data from the joint observational distribution. As a result, the researchers have to imagine the intervention and what would happen under the intervention, that is, the interventional or counterfactual distribution of the outcome. They then have to attempt a one-to-one translation from the imagined interventional or counterfactual distribution back to the observational distribution. If that translation is reasonable and credible and the interventional or counterfactual distribution can be expressed in terms of the observational distribution, then causal identification is said to be established. This means that the causal effect is estimable and computable from the observational data. The entire enterprise of identification is to find a way, using the front-door criterion or the back-door criterion or an instrumental variable, etc., to link together the interventional/counterfactual distribution and the observational distribution and specify the conditions under which such one-to-one mapping is possible.

Once it is determined that the effect of an intervention is estimable, different statistical methods can be used to actually compute the causal effect from the observed data. To do that, these methods attempt an approximation of the functions that Nature uses to generate the values of the variables. In a SCM framework, these generative functions are represented by the generic function notation f's. Parametric statistical models assume that we have accurate prior knowledge and information about the forms of these functions. If that prior knowledge is accurate, parametric models tend to perform well in estimating the target parameters that correspond to the causal quantities of interest. In social science research, however, usually we do not have concrete and credible knowledge about these functional forms. In that case, an application of flexible machine learning methods will be preferable because these methods do not depend as much on assumptions about the true functional forms as parametric models do. Instead, a machine learning method adopts a performance metric (loss function) and then approximates the data-generating mechanisms through a trial-and-error process by optimizing its predictive performance

measured by the chosen performance metric. Machine learning-based estimation therefore would likely produce more robust causal effects.

The above description is a summary of the machine learning-based causal inference approach that I adopt to study human rights and human rights treaty commitment and compliance. This approach adds value to political science research and empirical scientific research in general in a variety of ways. First, it offers a rigorous, yet intuitive, step-by-step workflow to make robust causal inference in applied research. One of the key benefits of this approach is that it demystifies the entire task of making causal inference. I believe the study of causality using machine learning should be as accessible as possible to every researcher rather than being the exclusive repertoire of a select group of academics sitting at the top echelon of the scientific community.

Second, my application of this methodological approach to the study of human rights and human rights treaties hopefully introduces and further promotes the use of the SCM framework in political science. The study of causality in the discipline is still being dominated by the potential outcomes framework. Political scientists not familiar with the SCM framework are thus being deprived of its benefits and more scientific progress might be delayed. It is worth noting, however, that the logical equivalence between these two frameworks has been firmly established [Pearl, 2009b] whereas the practical advantage of the SCM framework remains the subject of heated debates in the causal inference literature. There have been recent attempts to unify these two approaches as well [Richardson and Robins, 2014]. Similar empirical research that applies the SCM framework could make important contributions in and of itself while also offering useful application examples for other researchers to consider and follow.

Third, my application of graphical causal models also underscores the crucial importance of substantive domain knowledge in making causal inference. It should not be a surprise that when one is trying to make causal inference using observational data, one becomes hyper-aware of the centrality of the subject matter knowl-

edge and insights from qualitative research. It is those qualitative domain expertise that determines how one's graphical causal model should look like and whether and how the causal effect could be identified and made computable from observational data. As a result, adopting the SCM approach to causal inference more widely will bring more fruitful interactions, collaboration, and mutual appreciation between quantitative and qualitative scholars in any disciplines.

From my own experiences, however, reviewers and discussants, upon encountering a graphical causal model that purportedly represents the background knowledge about the underlying causal process, often raise immediate questions about what they believe are unsettled areas of disagreements in the human rights literature. That, I believe, is a testament to the power of transparency that graphical causal models hold over their algebraic counterparts in the potential outcomes framework. On the other hand, it also goes to show that causal inference research on human rights and international human rights law remains difficult mostly because of a lot of remaining uncertainties and disagreements in the domain knowledge.

By corollary, the kind of research questions that I believe are most amenable to, and most likely to benefit from, this machine learning-based causal inference approach are those that are supported by a sufficient amount of domain knowledge and, thus, graphical models of the underlying causal structure are credible and easy to justify. Moreover, these questions should likely involve potentially complex relationships among a large number of variables and, hence, flexible machine learning methods would prove more beneficial. Last but not least, these research questions should ideally be legitimate subjects of investigation in a discipline that is relatively open to adopting methodological advances from other fields and disciplines such as computer science and epidemiology.

There are certainly some tradeoffs, nevertheless. A machine learning-based causal inference approach might require additional investment in terms of time and resources on part of the applied researchers to get themselves familiarized and conversant with the framework and all the methods involved. It is a worthwhile

investment, though, because the skills and knowledge are transferable across fields and disciplines, in both academia and industry. Possibly it could also bring some intellectual curiosity and satisfaction when one starts to think deeply about the question of causality in the way, as quoted in Pearl et al. [2016, vii], that Virgil (29 BC) might have felt, “Lucky is he who has been able to understand the causes of things” (*Felix, qui potuit rerum cognoscere causas*).

APPENDICES

A. CHAPTER 2: APPENDIX

A.1 Variable Description

- **Treaty ratification status of the ICCPR, CEDAW, CAT**: A country–year binary variable coded 1 for ratification and 0 otherwise. Data are coded manually from the database of the Office of the High Commissioner for Human Rights.

  (http://www.ohchr.org/EN/HRBodies/Pages/HumanRightsBodies.aspx).

- **Human rights dynamic latent protection scores**: a country–year interval variable that measures respect for physical integrity rights. Rescaled to a 0–1 range from the empirical range for ease of estimation and interpretation. The scores were generated by Fariss [2014] using a dynamic ordinal item-response theory model that accounts for systematic change in the way human rights abuses have been monitored over time. The human rights scores model builds on data from the CIRI Human Rights Data Project, the Political Terror Scale, the Ill Treatment and Torture Data Collection, the Uppsala Conflict Data Program, and several other public sources.

  Variable name in original dataset is *latentmean*.

  (http://humanrightsscores.org).

- **CIRI women's political rights**: an ordinal variable from 0 – 3 that measures the extent to which women's political rights are protected, including the rights to vote, run for political office, hold elected office, join political parties, and petition government officials.

  A score of 0 indicates these rights are not guaranteed by law; a score of 1 indicates rights are guaranteed by law but severely restricted in practice; a

score of 2 indicates rights are guaranteed by law but moderately restricted in practices; and a score of 3 indicates rights are guaranteed in law and practice.

(http://www.humanrightsdata.com/p/data-documentation.html).

- **CIRI toture index**: an ordinal index that measures the extent of torture practice by government officials or by private individuals at the instigation of government officials. A score of zero indicates frequent torture practice; a score of 1 indicates occasional torture practice; and a score of 2 indicates that torture did not occur in a given year.

  (http://www.humanrightsdata.com/p/data-documentation.html).

- **Legal origins**: a cross-sectional (country) multinomial variable coded for British, French, German, Scandinavian, and Socialist legal origins. Data are from La Porta et al. [2008]. I recoded 1 for common law and 0 otherwise.

- **Ratification rules**: a cross-sectional (country) five-point ordinal variable (1, 1.5, 2, 3, 4) by [Simmons, 2009]. Its empirical maximum value, however, is only a score of 3. It measures "the institutional "hurdle" that must be overcome in order to get a treaty ratified." The coding is based on descriptions of national constitution or basic rule.

  (http://scholar.harvard.edu/files/bsimmons/files/APP_3.2_Ratification_rules.pdf).

- **Global and regional ratification rates**: continuous variables measuring the cumulative ratification rates globally and by region. Regional classification is defined using the United Nations Regional Groups of Member States, including Africa Group (AG), Asia-Pacific Group (APG), Eastern European Group (EEG), Latin American and Caribbean Group (GRULAC), and Western European and Others Group (WEOG).

  (http://www.un.org/depts/DGACM/RegionalGroups.shtml).

- **Democracy**: measured by the dummy variable *democracy* in the Democracy-Dictatorship dataset by Cheibub et al. [2010]. It is coded 1 if the regime qualifies as democratic and 0 otherwise. This measure is preferred to the Polity 4 dataset to avoid a conceptual overlap between democracy and physical integrity rights [Hill, 2016b].

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Multiple parties**: a ordinal variable coded 0 for no parties, 1 for single party, and 2 for multiple parties. Variable name in original dataset is *defacto*. I recoded 1 for multiple parties and 0 otherwise.

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Democratic transition**: a binary variable coded 1 when there is transition to or from democracy and 0 otherwise.

  Variable name in original dataset is *tt*.

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Judicial independence**: a time-series cross-sectional latent score (0 – 1) measuring judicial independence. The scores range from 0 (no judicial independence) to 1 (complete judicial independence).

  (http://polisci.emory.edu/faculty/jkstato/page3/index.html).

- **GDP per capita**: a country–year interval variable measuring gross domestic product divided by midyear population measured in current US dollars. A few country-year observations have a GDP per capita value of zero. I change that into the next smallest value of 65.

  (http://data.worldbank.org/indicator/NY.GDP.PCAP.CD).

- **Population**: a country–year interval variable measuring the total number of residents in a country regardless of their legal status.

  (http://data.worldbank.org/indicator/SP.POP.TOTL).

- **Trade**: a country–year interval variable measuring the sum of exports and imports of goods and services as a share of gross domestic product.

  (http://data.worldbank.org/indicator/NE.TRD.GNFS.ZS).

- **Net ODA received (current USD)**: data are from the World Bank Indicators database.

  (http://data.worldbank.org/indicator/DT.ODA.ODAT.CD).

- **Involvement in militarized interstate dispute**: a country–year binary variable from the Militarized Interstate Dispute Data (MIDB dataset, version 4.1). It is recoded 1 to indicate a country's involvement in any side of an militarized dispute and 0 otherwise between the start year and the end year of a dispute.

  (http://cow.dss.ucdavis.edu/data-sets/MIDs).

A.2 Summary Statistics

Table A.1.: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
COW country code	8,062	—	—	2	990
Year	8,062	—	—	1966	2013
ICCPR ratification	8,062	0.560	0.496	0	1
CEDAW ratification	8,062	0.563	0.496	0	1
CAT ratification	8,062	0.370	0.483	0	1
Human rights scores	8,062	0.345	1.420	−3.110	4.710
CIRI women's political rights	4,840	1.780	0.649	0	3
CIRI torture index	4,850	0.778	0.747	0	2
Legal origins	7,956	—	—	1	5
Ratification rules	7,796	1.800	0.640	1	3
ICCPR global rate	8,062	0.561	0.268	0	0.869
CEDAW global rate	8,062	0.564	0.379	0	0.964
CAT global rate	8,062	0.369	0.316	0	0.792
ICCPR regional rates	8,062	0.563	0.311	0	1
CEDAW regional rates	8,062	0.565	0.397	0	1
CAT regional rates	8,062	0.372	0.356	0	1
Democracy	6,886	0.442	0.497	0	1
Multiple parties	6,886	1.650	0.653	0	2
Transition	6,886	0.018	0.134	0	1
Judicial independence	7,679	0.465	0.321	0.01	0.995
Population	7,798	31,846,961	115,863,080	9,419	1,357,380,000
GDP per capita	7,055	6,907	14,088	37.5	193,648
Trade	6,536	75.7	49.3	0.021	532
Net ODA	7,490	268,622,622	619,681,691	−943,150,000	22,057,090,000
Militarized dispute	7,501	0.308	0.462	0	1

A.3 Multiple Imputation of Missing Data

Multiple imputation is used to fill in missing data and create five imputed datasets, covering 192 countries from 1965 – 2013. All variables in Table **??** are used to make the MAR assumption as plausible as possible. When modeling and estimating causal effects, however, I subset the observations by their appropriate time periods. For example, I only use observations from 1985–2013 when estimating the causal effects of predictive covariates on CAT ratification and 1982–2013 for modeling CEDAW ratification. As a result, the fractions of imputed missing data that are actually used for estimation tend to be lower. Variables with the highest missing fractions that are in use are CIRI torture index (missing fraction is 0.197) and CIRI measures of women's political rights (missing fraction is 0.196).

Table A.2.: Fractions of missing data by variables

Variables	**Missing fraction**
CIRI women's political rights	0.400
CIRI torture index	0.398
Trade participation	0.189
DD transition	0.146
DD multiple parties	0.146
DD democracy	0.146
GDP per capita	0.125
Judicial independence	0.048
Net ODA	0.071
Involvement in militarized dispute	0.070
Population size	0.033
Ratification rules	0.033
Legal origins	0.013
CAT ratification	0.000
CAT global ratification rate	0.000
CAT regional ratification rates	0.000
CEDAW ratification	0.000
CEDAW global ratification rate	0.000
CEDAW regional ratification rates	0.000
ICCPR ratification	0.000
ICCPR global ratification rate	0.000
ICCPR regional ratification rates	0.000
N of obs. after list-wise deletion	3,615
N of obs. after imputation	8,062

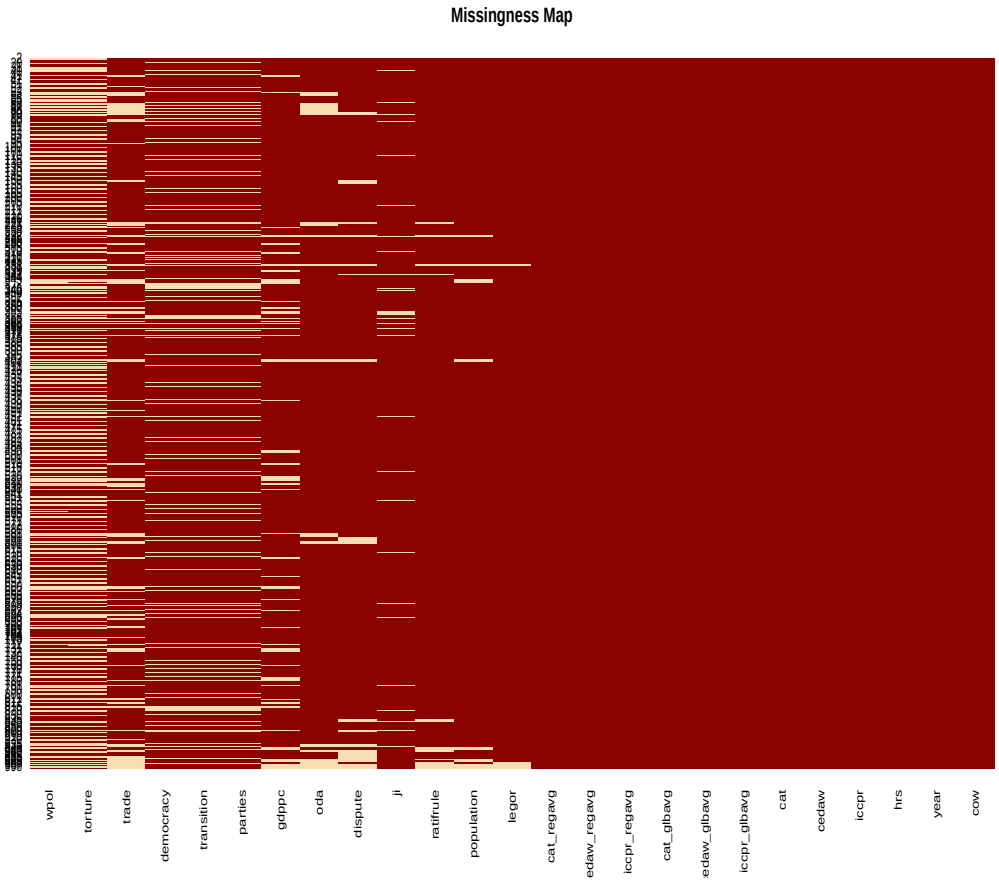


Fig. A.1.: Map of missing data

B. CHAPTER 3: APPENDIX

B.1 Variable Description

- **Ratification Status of ICCPR, CEDAW, CAT**: A country–year binary variable coded 1 for ratification and 0 otherwise. Data are coded manually from the database of the Office of the High Commissioner for Human Rights.

  (http://www.ohchr.org/EN/HRBodies/Pages/HumanRightsBodies.aspx).

- **Political Terror Scale**: a country–year five-point ordinal variable measuring levels of political murders, torture, political imprisonment, and disappearances. This variable was originally coded from 5 for worst level of abuses to 1 for least abuses. For consistency of interpretation with the other two measures of human rights outcome, we reverse-coded into 0 (worst performance) to 4 (best performance).

  (http://www.politicalterrorscale.org).

- **Women's Political Empowerment Index**: a country–year interval variable gauging women's political empowerment from 1900 to 2012 in 173 countries. The index is an aggregation of three sub-indices that range from 0 (lowest level of political empowerment) to 1 (highest level of political empowerment), including a women civil liberty index, a women civil society participation index, and a women political participation index. The overall women's political empowerment index is the average of these three indices.

  Variable name in original dataset is *v2x_gender*.

  (https://www.v-dem.net/en/data/data-version-5).

- **CIRI toture index**: an ordinal index that measures the extent of torture practice by government officials or by private individuals at the instigation of government officials. A score of zero indicates frequent torture practice; a score of 1 indicates occasional torture practice; and a score of 2 indicates that torture did not occur in a given year.

  (http://www.humanrightsdata.com/p/data-documentation.html).

- **Political Constraints Index**: an expert-coded country–year interval variable on a scale from 0 (most hazardous - no checks and balances) to 1 (most constrained–extensive checks and balances).

  Variable name in original dataset is *polconiii*

  (https://whartonmgmt.wufoo.com/forms/political-constraint-index-polcon-dataset/).

- **Judicial independence**: a time-series cross-sectional latent score, measuring judicial independence. The scores range from 0 (no judicial independence) to 1 (complete judicial independence).

  (http://polisci.emory.edu/faculty/jkstato/page3/index.html).

- **Name and shame index**: An country–year index that Cole [2015, 423] computes that "sums the standardized scores of four variables: media reporting of human rights abuses in (1) *The Economist* and (2) *Newsweek*; (3) Amnesty International press releases targeting a country's human rights blemishes; and (4) UN Commission on Human Rights resolutions condemning a country's human rights performance."

  Variable name in original dataset is *name_shame*.

- **Treaty Commitment Preference**: a country–year interval variable, ranging from -1 to 1, that Lupu [2016] computes to measure a country's commitment preference across a large number of treaties in different domains. We use

the first-dimension coordinates as a proxy of the degree to which states are internationally socialized as measured by their participation in the pool of 280 universal treaties.

Variable name in original dataset is *coord1d*.

- **Legal Origin**: a cross-sectional (country) multinomial variable coded for British, French, German, Scandinavian, and Socialist legal origins. Data are from La Porta et al. [2008]. We recoded 1 for British origin and 0 otherwise.

- **Ratification Rules**: a cross-sectional (country) five-point ordinal variable (1, 1.5, 2, 3, 4) by [Simmons, 2009]. It measures "the institutional "hurdle" that must be overcome in order to get a treaty ratified."

  Coding is based on descriptions of national constitutions or basic rule.

  (http://scholar.harvard.edu/files/bsimmons/files/APP_3.2_Ratification_rules.pdf).

- **Electoral System**: a cross-sectional (country) categorical variable coded Parliamentary (2), Assembly-elected President (1), Presidential (0). Some missing values are filled in using a relatively comparable coding system by Simmons [2009], in which the variable is coded 2 for primarily parliamentary system, 1 for hybrid system, and 0 for primarily presidential system. We recoded 0 for presidential system and 1 otherwise.

  (http://www.iadb.org/en/research-and-data/publication-details,3169.html?pub_id=IDB-DB-121).

- **Population**: a country–year interval variable measuring the total number of residents in a country regardless of their legal status.

  (http://data.worldbank.org/indicator/SP.POP.TOTL).

- **GDP per capita**: a country–year interval variable measuring gross domestic product divided by midyear population measured in current US dollars.

(http://data.worldbank.org/indicator/NY.GDP.PCAP.CD).

- **Trade**: a country–year interval variable measuring the sum of exports and imports of goods and services as a share of gross domestic product.

  (http://data.worldbank.org/indicator/NE.TRD.GNFS.ZS).

- **Democracy**: measured by the dummy variable *democracy* in the Democracy-Dictatorship dataset by Cheibub et al. [2010]. It is coded 1 if the regime qualifies as democratic and 0 otherwise. This measure is preferable to the Polity dataset since it may help avoid a conceptual overlap between democracy and physical integrity rights [Hill, 2016b].

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Multiple parties**: a ordinal variable coded 0 for no parties, 1 for single party, and 2 for multiple parties. Variable name in original dataset is *defacto*.

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Regime Durability**: a country–year interval variable measuring the number of age in years of the current regime as classified by regime.

  Variable name in original dataset is *agereg*.

  (https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited).

- **Involvement in militarized interstate dispute**: a country–year binary variable from the Militarized Interstate Dispute Data (MIDB dataset, version 4.1). It was recoded 1 to indicate a country's involvement in any side of an militarized dispute and 0 otherwise between the start year and the end year of a dispute.

  (http://cow.dss.ucdavis.edu/data-sets/MIDs).

B.2 Summary Statistics

Table B.1.: Summary Statistics

Statistic	N	Mean	SD	Min	Max
COW country code	5,460	—	—	2	990
Year	6,005	—	—	1981	2008
ICCPR ratification	5,460	0.614	0.487	0	1
CEDAW ratification	5,460	0.663	0.473	0	1
CAT ratification	5,460	0.410	0.492	0	1
Political Terror Scale	4,642	2.503	1.161	0	4
Women's empowerment index	4,077	0.630	0.205	0.107	0.965
CIRI torture index	4,285	0.787	0.748	0	2
Political constraints	4,844	0.234	0.218	0	0.726
Judicial independence	5,042	0.505	0.320	0.011	0.995
Name and shame index	3,388	−0.001	2.943	−1.499	26.310
Treaty preference (1d)	4,880	0.105	0.481	−1	0.993
Legal origins	5,320	1.958	0.978	1	5
Ratification rules	5,096	1.791	0.644	1	3
Electoral system	4,872	0.718	0.875	0	2
Population size	5,206	31,003,645	115,540,360	8,160	1,324,655,000
GDP per capita	4,646	7,324	13,849	64.810	193,648
Trade/GDP	4,217	78.710	50.470	0.021	531.700
Democracy	4,937	0.509	0.500	0	1
Multiple parties	4,937	1.720	0.612	0	2
Regime durability	4,937	26.920	27.120	1	139
Militarized disputes	4,863	0.333	0.471	0	1

B.3 Multiple Imputation of Missing Data

Multiple imputation is used to fill in missing data and create five imputed datasets. All variables in Table B.2 are used to make the MAR assumption as plausible as possible. The 1981–2008 time frame for observations was used to conduct multiple imputation.

Table B.2.: Fractions of missing data by variables

Variables	**Fraction of Missing**
Name and Shame Index	0.379
Women's Political Empowerment Index	0.253
Trade Participation	0.228
CIRI Torture Index	0.215
Political Terror Scale Score	0.150
GDP per capita	0.149
Political Constraints	0.113
Militarized dispute	0.109
Electoral Systems	0.108
Treaty Commitment Propensity (1d)	0.106
Democracy	0.096
De facto Multiple parties	0.096
Age of Regime	0.096
Judicial Independence	0.077
Ratification Rules	0.067
Population	0.047
Legal Origins	0.026
ICCPR Ratification status	0.000
CEDAW Ratification status	0.000
CAT Ratification status	0.000
N of observations after list-wise deletion	2,624
N of observations after imputation	5,460

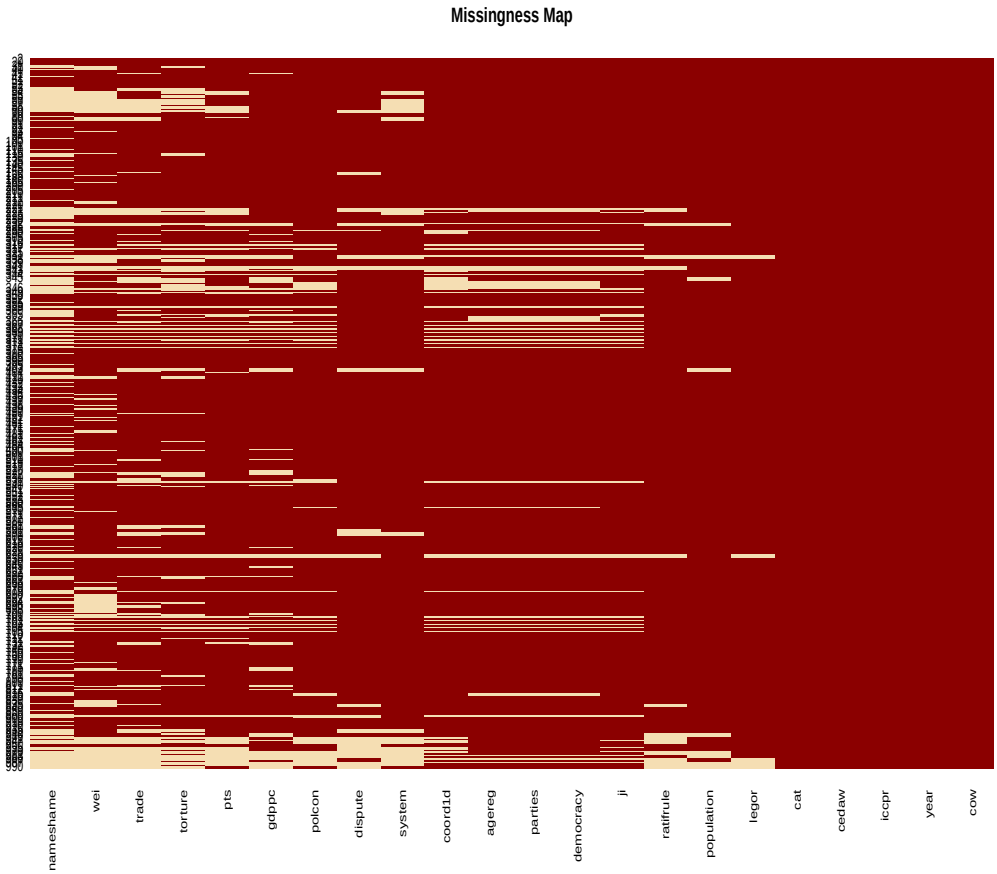


Fig. B.1.: Map of missing data

C. CHAPTER 4: APPENDIX

C.1 United Nations Human Rights Treaties

C.1.1 Status of ratification

CAT Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, 1465 UNTS 85, adopted 10 December 1984, entered into force 26 June 1987, ratified by 158 states.

CED Convention for the Protection of All Persons from Enforced Disappearance, UNTS 2715 Doc.A/61/448, adopted 20 December 2006, entered into force 23 December 2010, ratified by 50 states.

CEDAW Convention on the Elimination of All Forms of Discrimination against Women, 1249 UNTS 13, adopted 18 December 1979, entered into force 3 September 1981, ratified by 189 states.

CERD Convention on the Elimination of All Forms of Racial Discrimination, GA Res. 2106 (XX), Annex, 20 UN GAOR Supp. (No. 14) at 47, UN Doc. A/6014 (1966), 660 UNTS 195, adopted 7 March 1965, entered into force 4 January 1969, ratified by 177 states.

CMW International Convention on the Protection of the Rights of All Migrant Workers and Members of their Families, GA Res. 45/158, Annex, 45 UN GAOR Supp. (No. 49A) at 262, UN Doc. A/45/49, adopted 18 December 1990, entered

into force 1 July 2003, ratified by 48 states.

CRC Convention on the Rights of the Child, 1577 UNTS 3, adopted 20 November 1989, entered into force 2 September 1990, ratified by 194 states.

CRPD Convention on the Rights of Persons with Disabilities, UN Doc. A/61/611, adopted 13 December 2006, entered into force 3 May 2008, ratified by 157 states.

ICCPR International Covenant on Civil and Political Rights, 999 UNTS 171, adopted 16 December 16 1966, entered into force 23 March 1976, ratified by 168 states.

ICESCR International Covenant on Economic, Social and Cultural Rights, 993 UNTS 3, adopted 16 December 1966, entered into force 3 January 1976, ratified by 164 states.

C.1.2 Monitoring procedures

Table C.1.: Monitoring procedures under UN core human rights treaties

	State reporting	State communication	Individual communication	Inquiry	Country visit
CERD	✓	✓	✓(optional)	✗	✗
ICESCR	✓	✓(OP)	✓(OP)	✓(OP)	✗
ICCPR	✓	✓(optional)	✓(OP)	✗	✗
CEDAW	✓	✗	✓(OP)	✓(OP)	✗
CAT	✓	✓(optional)	✓(optional)	✓(optional)	✓(OP)
CRC	✓	✓(OP)	✓(OP)	✓(OP)	✗
CMW	✓	✓(optional)	✓(optional)	✗	✗
CRPD	✓	✗	✓(OP)	✓(OP)	✗
CED	✓	✓(optional)	✓(optional)	✓	✗

OP: Optional Protocol.

C.2 Variable Description

- **Ratification Status of Monitoring Procedures**: A country–year binary variable coded 1 for ratification and 0 otherwise. Monitoring procedures include (i) Art. 19 of the Convention against Torture (CAT), (ii) Art. 20, (iii) Art. 22, and (iv) Optional Protocol to the Convention against Torture (OPCAT). Data are from the Office of the High Commissioner for Human Rights database.

  (http://www.ohchr.org/EN/HRBodies/Pages/HumanRightsBodies.aspx).

- **Political Constraints Index**: an expert-coded country–year interval variable on a scale from 0 (most hazardous - no checks and balances) to 1 (most constrained–extensive checks and balances).

  Variable name in original dataset is *polconiii*.

  (https://whartonmgmt.wufoo.com/forms/political-constraint-index-polcon-dataset/).

- **Judicial independence**: a time-series cross-sectional interval variable ranging from 0 (no judicial independence) to 1 (complete judicial independence).

  (http://polisci.emory.edu/faculty/jkstato/page3/index.html).

- **Name and shame index**: An country–year index that Cole [2015, 423] computes that "sums the standardized scores of four variables: media reporting of human rights abuses in (1) *The Economist* and (2) *Newsweek*; (3) Amnesty International press releases targeting a country's human rights blemishes; and (4) UN Commission on Human Rights resolutions condemning a country's human rights performance."

  Variable name in original dataset is *name_shame*.

- **Treaty Commitment Propensity**: a country–year interval variable, ranging from -1 to 1, that Lupu [2016] computes to measure a country's commitment preference across a large number of treaties in different domains. I use

the first-dimension coordinates as a proxy of the degree to which states are internationally socialized as measured by their participation in the pool of 280 universal treaties.

Variable name in original dataset is *coord1d*.

- **Political Terror Scale**: a country–year five-point ordinal variable measuring levels of political murders, torture, political imprisonment, and disappearances. This variable was originally coded from 5 for worst level of abuses to 1 for least abuses. For consistency of interpretation with the other two measures of human rights outcome, we reverse coded into 0 (worst performance) to 4 (best performance).

  (http://www.politicalterrorscale.org).

- **Human Rights Scores**: a country–year interval variable that measures respect for physical integrity human rights. Rescaled to a 0–1 range from the empirical range for ease of interpretation. Low scores indicate low government's respect for physical integrity right whereas high scores indicate greater government's respect. The scores were generated using a dynamic ordinal item-response theory model that accounts for systematic change in the way human rights abuses have been monitored over time. The human rights scores model builds on data from the CIRI Human Rights Data Project, the Political Terror Scale, the Ill Treatment and Torture Data Collection, the Uppsala Conflict Data Program, and several other published sources.

  (http://humanrightsscores.org).

- **CIRI toture index**: an ordinal index that measures the extent of torture practice by government officials or by private individuals at the instigation of government officials. A score of zero indicates frequent torture practice; a score of 1 indicates occasional torture practice; and a score of 2 indicates that torture did not occur in a given year.

(http://www.humanrightsdata.com/p/data-documentation.html).

- **Legal origins**: a cross-sectional (country) multinomial variable coded for British, French, German, Scandinavian, and Socialist legal origins. Data are from La Porta et al. [2008]. We recoded 1 for common law and 0 otherwise.

- **Ratification rules**: a cross-sectional (country) five-point ordinal variable (1, 1.5, 2, 3, 4) by [Simmons, 2009]. Its empirical maximum value, however, is only a score of 3. It measures "the institutional "hurdle" that must be overcome in order to get a treaty ratified." The coding is based on descriptions of national constitution or basic rule.

  (http://scholar.harvard.edu/files/bsimmons/files/APP_3.2_Ratification_rules.pdf).

- **Electoral System**: a cross-sectional (country) categorical variable coded Parliamentary (2), Assembly-elected President (1), Presidential (0). Some missing values are filled in using a relatively comparable coding system by Simmons [2009], in which the variable is coded 2 for primarily parliamentary system, 1 for hybrid system, and 0 for primarily presidential system. We recoded 0 for presidential system and 1 otherwise.

  (http://www.iadb.org/en/research-and-data/publication-details,3169.html?pub_id=IDB-DB-121).

- **Population**: a country–year interval variable measuring the total number of residents in a country regardless of their legal status.

  (http://data.worldbank.org/indicator/SP.POP.TOTL).

- **GDP per capita**: a country–year interval variable measuring gross domestic product divided by midyear population measured in current US dollars.

  (http://data.worldbank.org/indicator/NY.GDP.PCAP.CD).

- **Trade**: a country–year interval variable measuring the sum of exports and imports of goods and services as a share of gross domestic product.

  (http://data.worldbank.org/indicator/NE.TRD.GNFS.ZS).

- **Regime Type**: measured by the Polity Score. The Polity Score is a country–year interval variable measuring regime authority spectrum on a 21-point scale ranging from -10 (hereditary monarchy) to $+10$ (consolidated democracy).

  (http://www.systemicpeace.org/polityproject.html).

- **Regime Durability**: a country–year interval variable measuring the number of years since the most recent regime change (defined by a three-point change in the POLITY score over a period of three years or less) or the end of transition period defined by the lack of stable political institutions.

  (http://www.systemicpeace.org/polityproject.html).

- **Involvement in militarized interstate dispute**: a country–year binary variable from the Militarized Interstate Dispute Data (MIDB dataset, version 4.1). It is recoded 1 to indicate a country's involvement in any side of an militarized dispute in a given year and 0 otherwise between the start year and the end year of a dispute.

  (http://cow.dss.ucdavis.edu/data-sets/MIDs).

C.3 Summary Statistics

Table C.2.: Summary Statistics

Statistic	N	Mean	SD	Min	Max
COW country code	5,609	—	—	2	990
Year	5,609	—	—	1986	2015
Reporting (Art. 19)	5,609	0.586	0.493	0	1
Inquiry (Art. 20)	5,609	0.555	0.497	0	1
Individual complaint (Art. 22)	5,609	0.250	0.433	0	1
Country visit (OPCAT)	5,609	0.107	0.309	0	1
Political Terror Scale scores	5,178	2.515	1.170	0	4
Human rights scores	5,227	0.559	1.442	−2.940	4.705
CIRI torture index	4,198	0.741	0.733	0	2
Political constraints index	5,285	0.267	0.216	0	0.726
Judicial independence	4,900	0.514	0.314	0.013	0.995
Name shame index	2,775	0.188	3.140	−1.499	26.310
Treaty commitment propensity	4,115	0.086	0.483	−1	0.993
Legal origins	5,503	1.949	0.975	1	5
Ratification rules	5,337	1.782	0.643	1	3
Electoral systems	5,081	0.715	0.876	0	2
Population size	5,386	34,533,273	126,414,302	9,419	1,371,220,000
GDP per capita	5,130	9,330	16,654	64.810	193,648
Trade/GDP proportion	4,657	82.090	49.720	0.021	531.700
Polity 4 scores	4,611	2.630	6.861	−10	10
Regime durability	4,676	24.480	30.290	0	206
Involvement in MID	4,983	0.282	0.450	0	1

C.4 Multiple Imputation of Missing Data

Multiple imputation is used instead to fill in missing data and create five imputed data sets. Estimates computed from these five datasets are then pooled according to Rubin's rules. All variables in the data summary statistics table (Table **??**) are used in the multiple imputation stage to make the missing at random (MAR) assumption as plausible as possible. The 1986–2015 time frame is used to impute missing data.

Table C.3.: Fractions of missing data by variables

Variables	**Fraction**
Name and shame index	0.505
Treaty commitment propensity	0.266
CIRI torture index	0.252
Polity 4 scores	0.178
Trade/GDP proportion	0.170
Regime durability	0.166
Judicial independence	0.126
Involvement in MID	0.112
Electoral system	0.094
GDP per capita	0.085
Political Terror Scale	0.077
Human rights protection scores	0.068
Political constraint index	0.058
Ratification rule	0.048
Population size	0.040
Legal origin	0.019
State reporting procedure	0.000
Inquiry procedure	0.000
Individual complaint procedure	0.000
Country visit procedure	0.000
Number of observations after listwise deletion	2,302
Number of observations after imputation	5,609

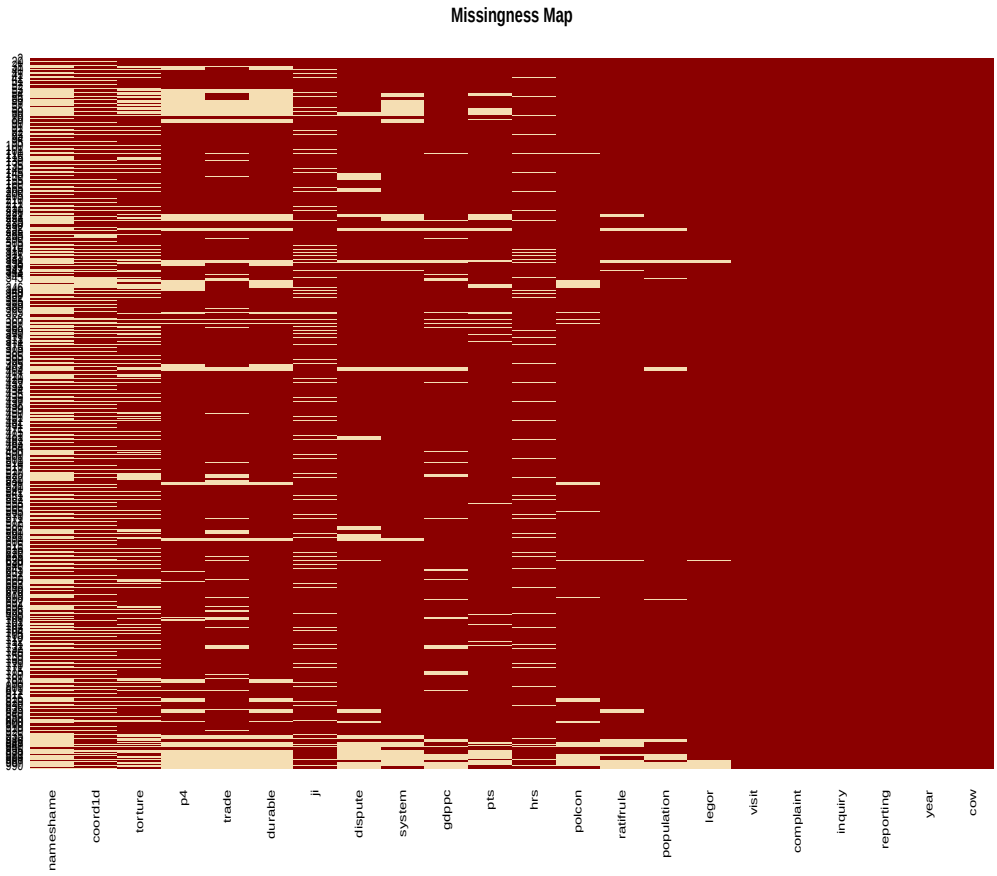


Fig. C.1.: Map of missing data for multiple imputation

D. CHAPTER 5: APPENDIX

D.1 Summary Statistics

Table D.1.: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
COW country code	3,443	—	—	2	990
Year	3,443	—	—	1981	1999
INGOs presence	3,443	565.3	650.8	0	3,523
Democracy (Polity 2)	2,181	1.834	7.646	−10	10
Competitiveness of exe. recruit.	2,181	2.063	0.922	1	3
Openness of exe. recruit	2,181	3.779	0.708	1	4
Executive constraints	2,181	4.591	2.210	1	7
Competitiveness of participation	2,181	3.067	1.495	1	5
CIRI physical integrity rights	2,625	4.846	2.400	0	8
CIRI disappearance	2,625	1.639	0.658	0	2
CIRI extrajudicial killings	2,625	1.332	0.776	0	2
CIRI political imprisonment	2,625	1.030	0.847	0	2
CIRI torture	2,625	0.845	0.765	0	2
Judicial independence	2,625	1.194	0.753	0	2
PTS score	2,412	2.740	1.137	1	5
Population (millions)	3,234	29,854	111,857	16.650	1,252,766
GDP per capita	3,234	6,213	6,559	155.1	43,138
Oil revenue per capita	3,048	447.8	2,134	0	49,588
Military regime	3,265	0.191	0.393	0	1
Left/right regime	2,935	1.370	1.282	0	3
Trade/GDP	2,934	76.950	46.720	1.064	401
Foreign direct investment	2,761	2.367	7.467	−82.890	145.2
Public trial	2,970	0.547	0.637	0	2
Fair trial	2,970	0.355	0.662	0	2
Final decision by court	2,958	0.586	0.883	0	2
Legislative approval	2,970	0.083	0.937	−1	2
WB structural adjustment	3,169	0.133	0.340	0	1
IMF structural adjustment	3,169	0.139	0.346	0	1
WB/IMF structural adjustment	3,187	0.222	0.415	0	1
British colony	3,443	0.332	0.471	0	1
Common law	3,443	0.249	0.432	0	1
PTA w/ human rights clause	2,309	0.267	0.443	0	1
CAT ratification	3,443	0.289	0.453	0	1
ICCPR ratification	3,443	0.553	0.497	0	1
Youth population	3,071	29.52	7.195	11.6	45
Latent score (Fariss 2014)	3,238	0.273	1.394	−3.134	4.311
Civil war	3,443	0.132	0.339	0	1
International war	3,443	0.008	0.091	0	1
AI press release (lagged)	2,286	0.930	2.505	0	26
AI background reports (lagged)	2,286	3.885	6.376	0	77
Wester media shaming (lagged)	2,286	0.297	1.102	0	25.5
HRO shaming (lagged)	1,257	0.168	0.857	0	13

D.2 Multiple Imputation of Missing Data: R code from Hill and Jones [2014]

- R version 3.4.1 (2017-06-30)

- Platform: x86_64-w64-mingw32/x64 (64-bit)
- Running under: Windows ¿= 8 x64 (build 9200)

```
rm(list = ls())
cat('\014')

## Use setup from original R code
df <- read.csv("rep_published.csv")
df$gdppc <- log(df$gdppc)
df$pop <- log(df$pop)
df$rentspc <- log(df$rentspc + 1)
df$trade_gdp <- log(df$trade_gdp)
df$ingo_uia <- log(df$ingo_uia + 1)
df$disap <- as.ordered(df$disap)
df$kill <- as.ordered(df$kill)
df$tort <- as.ordered(df$tort)
df$polpris <- as.ordered(df$polpris)
df$physint <- as.ordered(df$physint)
df$amnesty <- as.ordered(df$amnesty)
df$wbimfstruct <- as.integer(df$wbimfstruct)
df <- df[!is.na(df$physint) & !is.na(df$amnesty), ]

## Use MI from original R code
require(mice)
MI_ITER <- 5
methods <- c(rep("", 3), rep("ri", 5), rep("", 5), "", "", rep("ri", 3),
             "rf", rep("ri", 3), rep("rf", 7), "", "", "rf",
             "", "", "ri", "rf", rep("", 2), rep("rf", 4), rep("", 7))

mi <- mice(df, m = MI_ITER, method = methods, print = FALSE)
df.mi <- lapply(seq(1, MI_ITER), function(x) complete(mi, x))
```

```
29
30 for(i in 1:5){
31   write.csv(complete(mi, i), paste0("midata", i, ".csv" ), row.names = FALSE)
32 }
33 save.image("MI.RData")
```

Bibliography

Beth A. Simmons. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge: Cambridge University Press, 2009. 1, 8, 19, 20, 21, 37, 48, 56, 60, 61, 69, 72, 90, 95, 108, 111, 113, 117, 146, 169, 177, 187

Beth A Simmons. Reflections on mobilizing for human rights. *Journal of International Law and Politics*, 44:729–750, 2012. 1, 2, 4, 23, 105

Emilie M. Hafner-Burton. Book review - mobilizing for human rights: International law in domestic politics. *American Journal of International Law*, 104(3), 2010. 1

David Cingranelli. Book review - mobilizing for human rights: International law in domestic politics. *Human Rights Quarterly*, 32(3):761–764, 2010. 1

Steven C Poe and C Neal Tate. Repression of human rights to personal integrity in the 1980s: A global analysis. *American Political Science Review*, 88(4):853–872, 1994. 1

Steven C Poe, C Neal Tate, and Linda Camp Keith. Repression of the human right to personal integrity revisited: A global cross-national study covering the years 1976–1993. *International Studies Quarterly*, 43(2):291–313, 1999. 1

Linda Camp Keith. The united nations international covenant on civil and political rights: Does it make a difference in human rights behavior? *Journal of Peace Research*, 36(1):95–118, 1999. 1

Oona Hathaway. Do human rights treaties make a difference? *Yale Law Journal*, 111(8):1935–2042, 2002. 1, 90, 95

Todd Landman. *Protecting Human Rights: A Comparative Study*. Georgetown University Press, 2005. 1, 22, 49, 95

Emilie Hafner-Burton and James Ron. Seeing double: Human rights impact through qualitative and quantitative eyes. *World Politics*, 61(2):360–401, 2009. 1

Ann Marie Clark and Kathryn Sikkink. Information effects and human rights data: Is the good news about increased human rights information bad news for human rights measures? *Human Rights Quarterly*, 35(3):539–568, 2013. 1, 17, 91, 132

Christopher J Fariss and Geoff Dancy. Measuring the impact of human rights: Conceptual and methodological debates. *Annual Review of Law and Social Science*, 13 (1), 2017. 1, 17

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2000. 1, 109, 114

Judea Pearl. *Causality*. Cambridge University Press, 2009a. 2, 25, 27, 32, 37, 59, 68, 109, 117, 119, 126, 131, 141, 150, 153, 163

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016. 2, 27, 38, 40, 63, 74, 75, 76, 80, 110, 118, 143, 146, 163, 167

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 2

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 2005. 2

Jasjeet Sekhon. The neyman–rubin model of causal inference and estimation via matching methods. In Janet M Box-Steffensmeier, Henry E Brady, and David Collier, editors, *The Oxford handbook of Political Methodology*. Oxford University Press, 2008. 2

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009b. 2, 165

Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001. 3, 91

Judea Pearl. The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436, 2012. 3, 59, 68, 77

Ryan Goodman and Derek Jinks. International law and state socialization: Conceptual, empirical, and normative challenges. *Duke Law Journal*, 54:983–998, 2004a. 3

Ryan Goodman and Derek Jinks. How to influence states: Socialization and international human rights law. *Duke Law Journal*, 54:621–703, 2004b. 3

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The Elements of Statistical Learning*, volume 2. Springer, 2009. 5

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013. 5, 82, 134

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–231, 2001a. 5, 139

Gary King and Langche Zeng. Improving forecasts of state failure. *World Politics*, 53(4):623–658, 2001. 5

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375, 2010. 5

Kristian Skrede Gleditsch and Michael D Ward. Forecasting is difficult, especially about the future: Using contentious issues to forecast interstate disputes. *Journal of Peace Research*, 50(1):17–31, 2013. 5

Daniel W Hill and Zachary M Jones. An empirical evaluation of explanations for state repression. *American Political Science Review*, 108(3):1–27, 2014. vii, 5, 15, 25, 35, 128, 129, 130, 131, 132, 133, 134, 135, 137, 139, 140, 141, 142, 149, 153, 193

Mark S Bell. Examining explanations for nuclear proliferation. *International Studies Quarterly*, 2015. 5

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. 6, 43, 44, 82, 135

Tianqi Chen and Tong He. Higgs boson discovery with boosted trees. In *JMLR: Workshop and Conference Proceedings*, number 42, pages 69–80, 2015. 6, 44, 82, 135

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016. 6, 44, 82, 135

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. 6, 43, 83, 134

Mark J van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011. 25, 45, 46, 119

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994. 7, 86

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference*, volume 5. Cambridge University Press, 2016. 7, 53

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017. 13

Miguel A Hernán. The c-word: Scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, (0):e1–e4, 2018. 2, 14

Judea Pearl. The deductive approach to causal inference. *Journal of Causal Inference*, 2(2):115–129, 2014a. 16, 55

Judea Pearl. The eight pillars of causal wisdom. *Lecture Notes for the UCLA WCE Conference*, 2017. 16

Stephen Leeder. Special section: Causality in epidemiology. *International Journal of Epidemiology*, 45(6), 2016. 17

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002. 17

Rhian M Daniel, Michael G Kenward, Simon N Cousens, and Bianca L De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012. 18

Felix Thoemmes and Karthika Mohan. Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4):631–642, 2015. 18

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *arXiv preprint arXiv:1801.03583*, 2018. 18

Christopher J Fariss. Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review*, 108 (2):297–318, 2014. 17, 37, 90, 91, 112, 113, 132, 142, 168

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436, 2015.

Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. *Science*, 358(6370):1530–1534, 2017.

Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017. 58

Yonatan Lupu. Why do states join some universal treaties but not others? an analysis of treaty commitment preferences. *Journal of Conflict Resolution*, pages 1–32, 2014. 21, 25

Eric C Polley and Mark J van der Laan. Super learner in prediction. *Working Paper Series UC Berkeley Division of Biostatistics*, 2010. 25, 44, 83, 120, 134

Olivier De Schutter. *International Human Rights Law: Cases, Materials, Commentary*. Cambridge University Press, 2010. 19

Gudmundur Alfredsson, Jonas Grimheden, BC Ramcharan, and Alfred de Zayas. *International Human Rights Monitoring Mechanisms: Essays in Honour of Jakob Th. Möller*. Martinus Nijhoff, 2009. 19

Thomas Buergenthal. The evolving international human rights system. *American Journal of International Law*, 100:783–807, 2006. 19

Helen Keller and Geir Ulfstein. *UN Human Rights Treaty Bodies: Law and Legitimacy*, volume 1. Cambridge University Press, 2012. 19, 97

Nigel S Rodley. The role and impact of treaty bodies. In Dinah Shelton, editor, *The Oxford Handbook of International Human Rights Law*, pages 621–648. Oxford University Press, 2013. 19, 62, 97, 104

Martha Finnemore and Kathryn Sikkink. International norm dynamics and political change. *International Organization*, 52(4):887–917, 1998. 19, 72, 117

Jay Goodliffe and Darren G. Hawkins. Explaining commitment: States and the convention against torture. *Journal of Politics*, 68(2):358–371, 2006. 8, 19, 21, 37, 49, 52, 69, 113

Oona A Hathaway. Why do countries commit to human rights treaties? *Journal of Conflict Resolution*, 51(4):588–621, 2007. ix, 19, 21, 28, 29, 37, 40, 50, 69, 113

James H Lebovic and Erik Voeten. The politics of shame: The condemnation of country human rights practices in the unhcr. *Internationl Studies Quarterly*, 50 (4):861–888, 2006. 20

Douglas Hamilton Spence. Foreign aid and human rights treaty ratification: Moving beyond the rewards thesis. *The International Journal of Human Rights*, 18(4-5): 414–432, 2014. 20

Shannon Lindsey Blanton and Robert G Blanton. What attracts foreign investors? an examination of human rights and foreign direct investment. *Journal of Politics*, 69(1):143–155, 2007. 20

Emilie M Hafner-Burton. Trading human rights: How preferential trade agreements influence government repression. *International Organization*, 59(3):593–629, 2005. 21

Heather Smith-Cannoy. *Insincere Commitments: Human Rights Treaties, Abusive States, and Citizen Activism*. Georgetown University Press, 2012. 21, 56, 61

Richard A Nielsen and Beth A Simmons. Rewards for ratification: Payoffs for participating in the international human rights regime? *International Studies Quarterly*, 59(2):197–208, 2015. 21, 37, 49

Andrew Moravcsik. The origins of human rights regimes: Democratic delegation in postwar europe. *International Organization*, 54(2):217–252, 2000. 21, 37

Eric Neumayer. Qualified ratification: Explaining reservations to international human rights treaties. *The Journal of Legal Studies*, 36(2):397–429, 2007. 21, 37, 69, 113

Daniel W Hill. Avoiding obligation: Reservations to human rights treaties. *Journal of Conflict Resolution*, 60(6):1–30, 2016a. 21, 72, 117

Jana von Stein. Making promises, keeping promises: Democracy, ratification and compliance in international human rights law. *British Journal of Political Science*, 46(3):655–679, 2016. 22, 27, 49, 52

George .W. Downs, David M. Rocke, and Peter N. Barsoom. Is the good news about compliance good news about cooperation? *International Organization*, 50:379–406, 1996. 22, 35, 70

Jana von Stein. Do treaties constrain or screen? selection bias and treaty compliance. *American Political Science Review*, 99(4):611–622, 2005. 22, 35, 70, 116

Beth A Simmons and Daniel J Hopkins. The constraining power of international treaties: Theory and methods. *American Political Science Review*, 99(04):623–631, 2005. 22, 116

Emilie M Hafner-Burton, Edward D Mansfield, and Jon CW Pevehouse. Human rights institutions, sovereignty costs and democratization. *British Journal of Political Science*, 45(1):1–27, 2015a. 22, 36

James R. Vreeland. Political institutions and human rights: Why dictatorships enter into the united nations convention against torture. *International Organization*, 62 (1):65, 2008. ix, 22, 23, 29, 30, 31, 37, 41, 51, 69, 72

James Hollyer and B. Peter Rosendorff. Why do authoritarian regimes sign the convention against torture? signaling, domestic politics and non-compliance. *Quarterly Journal of Political Science*, 6(3-4):275–327, 2011. 22, 23, 37, 41, 52, 69, 72, 105, 106

Emilie M. Hafner-Burton. International regimes for human rights. *Annual Review of Political Science*, 15:265–286, 2012. 23

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA, 2017. 27, 65, 70, 131, 141, 143, 146, 163

Stephen Chaudoin, Jude Hays, and Raymond Hicks. Do we really know the wto cures cancer? false positives and the effects of international institutions. *British Journal of Political Science*, pages 1–26, 2016. 28

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2410–2416, 2014. 30

Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009. 32

Felix Elwert. Graphical causal models. In *Handbook of Causal Analysis for Social Research*, pages 245–273. Springer, 2013. 32, 68, 141

Iván Díaz, Alan Hubbard, Anna Decker, and Mitchell Cohen. Variable importance and prediction methods for longitudinal problems with missing variables. *PloS One*, 10(3):1–17, 2015. 33, 53, 144

Yogesh Tyagi. The denunciation of human rights treaties. *British Yearbook of International Law*, 79(1):86–193, 2009. 36, 64

Wade M. Cole. Sovereignty relinquished? explaining commitment to the international human rights covenants, 1966-1999. *American Sociological Review*, 70(3): 472–495, 2005. 8, 36

Yonatan Lupu. The informative power of treaty commitment: Using the spatial model to address selection effects. *American Journal of Political Science*, 57(4), 2013a. 36, 91, 123

Xinyuan Dai. Why comply? the domestic constituency mechanism. *International Organization*, 59(02):363–398, 2005. 36

Geoffrey PR Wallace. International law and public attitudes toward torture: An experimental study. *International Organization*, 67(01):105–140, 2013. 36

Yonatan Lupu. Legislative veto players and the effects of international human rights agreements. *American Journal of Political Science*, 59(3):578–594, 2015. 36, 56, 60, 69, 108, 111, 113

Charles D Crabtree and Christopher J Fariss. Uncovering patterns among latent variables: Human rights and de facto judicial independence. *Research & Politics*, 2(3):2053168015605343, 2015. 36, 61

Emilia J. Powell and Jeffrey K. Staton. Domestic judicial institutions and human rights treaty violation. *International Studies Quarterly*, 53(1):149–174, 2009. 36, 37, 61, 69, 111, 113, 149

Sara McLaughlin Mitchell, Jonathan J Ring, and Mary K Spellman. Domestic legal traditions and states' human rights practices. *Journal of Peace Research*, 50(2): 189–202, 2013. 37, 69, 113

Rafael La Porta, Florencio Lopez-de Silanes, and Andrei Shleifer. The economic consequences of legal origins. *Journal of Economic Literature*, 46(2):285–332, 2008. 37, 69, 113, 169, 177, 187

Terrence L Chapman and Stephen Chaudoin. Ratification patterns and the international criminal court1. *International Studies Quarterly*, 57(2):400–409, 2013. 37, 69, 113

José Antonio Cheibub, Jennifer Gandhi, and James Raymond Vreeland. Democracy and dictatorship revisited. *Public choice*, 143(1-2):67–101, 2010. 37, 49, 69, 113, 170, 178

Erik Melander, Therése Pettersson, and Lotta Themnér. Organized violence, 1989–2015. *Journal of Peace Research*, 53(5):727–742, 2016. 37

Nils Petter Gleditsch, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. Armed conflict 1946-2001: A new dataset. *Journal of Peace Research*, 39(5):615–637, 2002. 37

Drew A Linzer and Jeffrey K Staton. A global measure of judicial independence, 1948–2012. *Journal of Law and Courts*, 3(2):223–256, 2015. 37, 66, 69, 111, 113

Emilie Hafner-Burton and Kiyoteru Tsutsui. Justice lost! the failure of international human rights law to matter where needed most. *Journal of Peace Research*, 44 (4):407–425, 2007. 37, 69, 90, 95, 113

Emilie M Hafner-Burton. *Forced to Be Good: Why Trade Agreements Boost Human Rights*. Cornell University Press, 2013. 37, 69, 113

David L. Cingranelli, David L. Richards, and K. Chad Clay. The cingranelli-richards (ciri) human rights dataset. 2013. 37, 69, 112, 113, 132

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986. 42, 86, 147

James M Robins, Sander Greenland, and Fu-Chang Hu. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999. 42, 86, 147

Rhian M Daniel, Bianca L De Stavola, Simon N Cousens, et al. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal*, 11(4):479, 2011. 42, 63, 86, 133

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science & Business Media, 2007. 42, 86, 133

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014. 42

Sandra E Sinisi, Eric C Polley, Maya L Petersen, Soo-Yon Rhee, and Mark J van der Laan. Super learning: An application to the prediction of hiv-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007. 43

Noémi Kreif, Richard Grieve, Iván Díaz, and David Harrison. Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Economics*, 24(9):1213–1228, 2015. 43, 86

Cyrus Samii, Laura Paler, and Sarah Daly. Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. *Political Analysis*, Forthcoming, 2016. 43, 86

Romain Neugebauer, Bruce Fireman, Jason A Roy, Marsha A Raebel, Gregory A Nichols, and Patrick J O'Connor. Super learning to hedge against incorrect inference from arbitrary parametric assumptions in marginal structural modeling. *Journal of Clinical Epidemiology*, 66(8):S99–S109, 2013. 43

Romain Pirracchio, Maya L Petersen, and Mark van der Laan. Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2):108–119, 2015. 43, 86

Trevor J Hastie and Robert J Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990. 43, 82

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 43, 82

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001b. 43, 82

Robert E Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT press, 2012. 44

Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 2013. 44, 82, 135

James Honaker, Gary King, Matthew Blackwell, et al. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. 47, 81, 121

Daniel W Hill. Democracy and the concept of personal integrity rights. *Journal of Politics*, 78(3):822–835, 2016b. 49, 141, 170, 178

Alan Hubbard, Ivan Diaz Munoz, Anna Decker, John B Holcomb, Martin A Schreiber, Eileen M Bulger, Karen J Brasel, Erin E Fox, Deborah J Del Junco, Charles E Wade, et al. Time-dependent prediction and evaluation of variable importance using superlearning in high dimensional clinical data. *The Journal of Trauma and Acute Care Surgery*, 75(1):S53–S60, 2013. 53

Romain Pirracchio, John K Yue, Geoffrey T Manley, Mark J van der Laan, Alan E Hubbard, et al. Collaborative targeted maximum likelihood estimation for variable importance measure: Illustration for functional outcome prediction in mild traumatic brain injuries. *Statistical Methods in Medical Research*, pages 1–15, 2016. 53

Jennifer Ahern, K Ellicott Colson, Claire Margerson-Zilko, Alan Hubbard, and Sandro Galea. Predicting the population health impacts of community interventions: The case of alcohol outlets and binge drinking. *American Journal of Public Health*, 106(11):1938–1943, 2016. 53

Geoff Dancy and Kathryn Sikkink. Ratification and human rights prosecutions: Toward a transnational theory of treaty compliance. *NYU Journal of International Law and Policy*, 44:751, 2012. 56, 61

M Rodwan Abouharb, Laura P Moyer, and Megan Schmidt. De facto judicial independence and physical integrity rights. *Journal of Human Rights*, 12(4):367–396, 2013. 56

Margaret E. Keck and Kathryn Sikkink. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca: Cornell University Press, 1998. 56, 62, 111

Ryan Goodman and Derek Jinks. *Socializing States: Promoting Human Rights Through International Law*. Oxford University Press, 2013. 56, 62, 69, 111, 113

Ann Marie Clark. The normative context of human rights criticism: Treaty ratification and un mechanisms. In Thomas Risse, Stephen C. Ropp, and Kathryn Sikkink, editors, *From Commitment to Compliance: The Persistent Power of Human Rights*. Cambridge University Press, Cambridge, 2013. 56, 62, 69, 91, 95, 111, 113

Charles M Judd and David A Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981. 57

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986. 57

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–1035, 2013. 57

Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013. 57, 63, 79, 81, 92

Judea Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459, 2014b. 58, 65, 77, 78, 94

Martin et al. Scheinin. Final report on the impact of findings of the united nations human rights treaty bodies. In *Report of the 71st Conference of the International Law Association*, 2004. 61

Amanda M. Murdie and David R. Davis. Shaming and blaming: Using events data to assess the impact of human rights ingos. *International Studies Quarterly*, 56 (1):1–16, 2012. 61, 69, 113

Thomas Risse and Kathryn Sikkink. *The Persistent Power of Human Rights: From Commitment to Compliance*, volume 126. Cambridge University Press, 2013. 62

Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015. 63, 81

Kosuke Imai, Luke Keele, Teppei Yamamoto, et al. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010. 63

Matthew Blackwell. A framework for dynamic causal inference in political science. *American Journal of Political Science*, 57(2):504–520, 2013. 65

Valerio Bacak and Edward H Kennedy. Marginal structural models: An application to incarceration and marriage during young adulthood. *Journal of Marriage and Family*, 77(1):112–125, 2015. 65

Tyler J VanderWeele and Eric J Tchetgen Tchetgen. Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017. 65

Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2013. 65, 83

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology (Cambridge, Mass.)*, 25(2):282, 2014. 65, 79

Tyler J VanderWeele, Stijn Vansteelandt, and James M Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306, 2014. 65, 73, 79

Witold J Henisz. The political constraint index (polcon) dataset. *Website: https://whartonmgmt.wufoo.com/forms/political-constraint-index-polcon-dataset*, 2002. 66, 69, 111, 113

Wade Cole. Mind the gap: State capacity and the implementation of human rights treaties. *International Organization*, 69(02):405–441, 2015. 67, 69, 111, 113, 176, 185

Yonatan Lupu. Why do states join some universal treaties but not others? an analysis of treaty commitment preferences. *Journal of Conflict Resolution*, 60(7):1219–1250, 2016. 67, 69, 111, 113, 176, 185

Courtenay R. Conrad. Divergent incentives for dictators: Domestic institutions and (international promises not to) torture. *Journal of Conflict Resolution*, 2013. 69, 72, 91, 111, 113

Mark Gibney, Linda Cornett, Reed Wood, and Peter Haschke. Political terror scale 1976-2015. 2016. 69, 112, 113

Aksel Sundström, Pamela Paxton, Yi-ting Wang, and Staffan I Lindberg. Women's political empowerment: A new global index, 1900-2012. *World Development*, Forthcoming, 2017. 69

Lotta Themnér. Ucdp/prio armed conflict dataset codebook. *Uppsala Conflict Data Program (UCDP)*, 2014. 69, 113

David Cingranelli and Mikhail Filippov. Electoral rules and incentives to protect human rights. *Journal of Politics*, 72(1):243–257, 2010. 69, 113

Philip Keefer Cruz, Cesi and Carlos Scartascini. Database of political institutions codebook, 2015 update (dpi2015). 2016. 69, 113

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009. 68, 114

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, pages 3.1–3.29, 2017. 68

Maya L Petersen and Mark J van der Laan. Direct effect models. *International Journal of Biostatistics*, 4(1):1–27, 2008. 74

Avidit Acharya, Matthew Blackwell, and Maya Sen. Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, Forthcoming, 2016. 74, 80, 81, 83

Tyler J VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009. 81

Stijn Vansteelandt. Estimating direct effects in cohort and case–control studies. *Epidemiology*, 20(6):851–860, 2009. 81, 83

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(04):765–789, 2011. 81, 92

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008. 87

Daniel W. Hill. Estimating the effects of human rights treaties on state behavior. *Journal of Politics*, 72(4):1161–1174, 2010. 90, 95, 123

Courtenay R. Conrad and Emily H. Ritter. Tenure, treaties, and torture: The conflicting domestic effects of international law. *Journal of Politics*, 2013. 90

Robert et al. Kolb. *Academic Platform Project on the 2020 Review: Strengthening Human Rights Protection by Enhancing the Effective Functioning of the Human Rights Treaty Body System*, 2016. https://www.geneva-academy.ch/our-projects/our-projects/un-human-rights-mechanisms/detail/16-academic-platform-on-treaty-body-review-2020. 95

Eric Neumayer. Do international human rights treaties improve respect for human rights? *Journal of Conflict Resolution*, 49(6):925–953, 2005. 95

Yonatan Lupu. Best evidence: The role of information in domestic judicial enforcement of international human rights agreements. *International Organization*, 67 (3):469–503, 2013b. 95, 107

Geir Ulfstein and Helen Keller. *UN Human Rights Treaty Bodies*. Cambridge: Cambridge University Press, 2012. 95, 96, 102, 103

Miloon Kothari. From commission to council: Evolution of un charter bodies. In Dinah Shelton, editor, *The Oxford Handbook of International Human Rights Law*, pages 587–620. Oxford University Press, 2013. 97

Suzanne Egan. *The United Nations Human Rights Treaty System: Law and Procedure*. Bloomsbury Professional, 2011. 97

Cherif M. Bassiouni and William A. Schabas. *New Challanges for the UN Human Rights Machinery*. Intersentia, 2011. 97

Kenneth W Abbott, Robert O Keohane, Andrew Moravcsik, Anne-Marie Slaughter, and Duncan Snidal. The concept of legalization. *International Organization*, 54 (03):401–419, 2000. 101

Emilie Hafner-Burton, Edward Mansfield, and Jon Pevehouse. Human rights institutions, sovereignty costs and democratization. *British Journal of Political Science*, 45:1–27, 2015b. 101, 105, 106

Navanethem Pillay. Strengthening the united nations human rights treaty body system. *The Office of the High Commissioner for Human Rights*, 2012. 101

Elina Steinerte, Malcolm David Evans, and Antenor Hallo de Wolf. *The Optional Protocol to the UN Convention Against Torture*. Oxford University Press, 2011. 103

SPT. *SPT visits*, 2018. http://tbinternet.ohchr.org/_layouts/TreatyBodyExternal/CountryVisits.aspx?SortOrder=Alphabetical. 103

Daniel A. Farber. Rights as signals. *Journal of Legal Studies*, 31:83–94, 2002. 105

David S Jonas and Thomas N Saunders. The object and purpose of a treaty: Three interpretive methods. *Vanderbilt Journal of Transnational Law*, 43(3):565–609, 2010. 106

Ryan Goodman and Derek Jinks. Measuring the effects of human rights treaties. *European Journal of International Law*, 14(1):171–183, 2003. 106

Wade M Cole. Hard and soft commitments to human rights treaties, 1966–2001. *Sociological Forum*, 24(3):563–588, 2009. 8, 106

Ann Marie Clark. Human rights ngos at the united nations: Developing an optional protocol to the convention against torture. In Jutta Joachim and Birgit Locher, editors, *Transnational Activism in the UN and EU: A Comparative Study*. Routledge, 2009. 106

Malcolm Evans. The opcat at 50. In Geoff Gilbert and Clara Sandoval Villalba, editors, *The Delivery of Human Rights: Essays in Honour of Professor Sir Nigel Rodley*. Taylor & Francis, 2011. 106

Rachel Brewster. Reputation in international relations and international law theory. In Jeffrey L Dunoff and Mark A Pollack, editors, *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, pages 524–543. Cambridge University Press, 2013. 107

Cosette D Creamer and Beth A Simmons. Ratification, reporting, and rights: Quality of participation in the convention against torture. *Human Rights Quarterly*, 37(3): 579–608, 2015. 110

Brian Greenhill. *Transmitting Rights: International Organizations and the Diffusion of Human Rights Practices*. Oxford University Press, 2016. 112

G Marshall Monty, Jaggers Keith, and Gurr Ted Robert. Polity iv project: Political regime characteristics and transitions, 1800-2015. *Dataset Users' Manual. Center for International Development and Conflict Management, University of Maryland*, 2016. 113

George W. Downs, David M. Rocke, and Peter N. Barsoom. Managing the evolution of multilateralism. *International Organization*, 52(2):397–419, 1998. 125

Barbara Koremenos. Contracting around international uncertainty. *American Political Science Review*, 99(4):549, 2005. 125

Michael J Gilligan and Leslie Johns. Formal models of international institutions. *Annual Review of Political Science*, 15:221–243, 2012. 125

Kenneth W Abbott and Duncan Snidal. Law, legalization, and politics: An agenda for the next generation of il/ir scholars. In Jeffrey L Dunoff and Mark A Pollack, editors, *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, pages 33–57. Cambridge University Press, 2013. 125

Valen E Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013. 128

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, Eric-Jan Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2017. 128

Blakeley B McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett. Abandon statistical significance. *arXiv preprint arXiv:1709.07588*, 2017. 128

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. 131, 150, 151

Mark Gibney, Linda Cornett, Reed Wood, and Peter Haschke. Political terror scale 1976-2014. 2015. 132

Courtenay R Conrad, Daniel W Hill Jr, and Will H Moore. Torture and the limits of democratic institutions. *Journal of Peace Research*, page 0022343317711240, 2018. 149

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *arXiv preprint arXiv:1706.08576*, 2017. 151

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *arXiv preprint arXiv:1706.08058*, 2017. 151

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995. 153

Peter Spirtes and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. In *Applied Informatics*, volume 3, page 3. Springer, 2016. 153

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(1):1103–1204, 2016. 153

Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, 2017. 153

Thomas S Richardson and J Robins. *Single World Intervention Graphs (swigs)*. Now Publishers Incorporated, 2014. 165