ONLINE APPENDIX

Robustness Checks and Supplementary Analyses for
"Presidential Policymaking, 1877—2020"

Contents

A Source Record Groups **1**

A.1 Executive Orders . . . 1

A.2 Memoranda . . . 1

A.3 Proclamations . . . 2

A.4 Distribution of significance estimates . . . 4

B Training Data, Modeling Validity, and Robustness Checks **5**

B.1 Training Data Drawn from Chiou & Rothenberg . . . 5

B.2 Robustness and Validity . . . 6

B.3 AUC and Multi-class classification . . . 8

B.4 Document Transcription . . . 8

B.5 Language Heterogeneity & Unrepresentative Training Sets . . . 9

B.6 Heteroskedastic Predictive Accuracy . . . 10

B.7 Temporal Variation in Modeling Accuracy . . . 10

B.8 Feature Importance . . . 12

B.9 Robustness to Threshold Choices . . . 16

B.10 Face Validity . . . 18

C Alternative Polynomial Specifications **21**

D Additional Results: Time Series Analysis **23**

D.1 Factor analysis . . . 23

D.2 Nonparametric estimation and Disaggregated Changepoints . . . 23

E Additional details related to replications **25**

E.1 Additional tests related to replication of Christenson and Kriner (2019) . . . 25

E.2 Full results for Djourelova and Durante (2022, Table 3) replication . . . 27

A Source Record Groups

In this section we indicate which unilateral action "Source Record Groups" we group into each larger category of unilateral action.

A.1 Executive Orders

This category contains documents which are numbered and unnumbered executive orders.

EO - Executive Orders 1862-present

03 - Public Land Orders 1942-present

06 - Secretary of Interior Orders 1920-1950

22 - Executive Orders Relating to the Panama Canal 1902-1934

33 - Executive Orders Relating to Public Lands 1841-1935

41 - Executive Orders Relating to Public Lands 1820-1913

A.2 Memoranda

This category contains Executive Memoranda or other such memoranda from collections of presidential documents.

04 - Presidential Documents 1936-present

05 - White House Records 1869-present

08 - Manuscript collections 1790-1929

12 - Treasury and Justice Dept Records 1789-1908

17 - Navy and War Dept Records 1789-1884

20 - Messages and Papers of the President 1789-1899

21 - Public Papers of the Presidents 1789-present

37 - Abandoned Military Lands 1826-1905

52 - Miscellaneous Printed Sources 1789-1936

53 - Weekly Compilation of Presidential Documents 1965-present

56 - Presidential Policy Directives & National Security Decision Memoranda

A.3 Proclamations

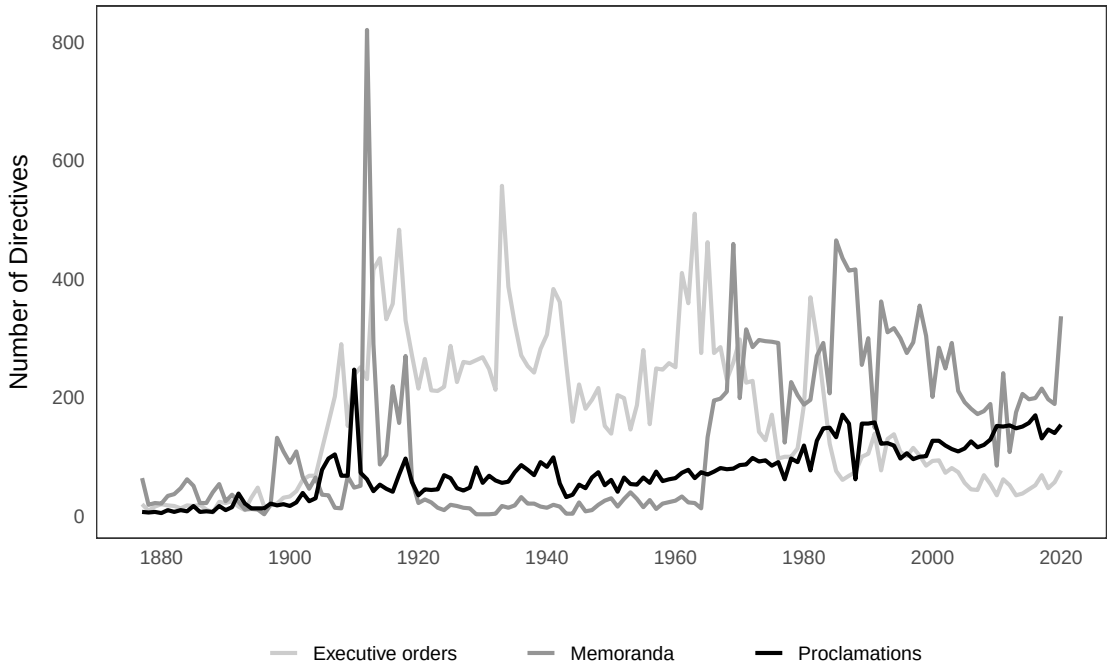
This category includes only documents clearly noted as proclamations.

PR - Proclamations 1789-present

29 - Treaty Proclamations 1789-present

35 - Proclamations Relating to Public Lands 1834-1907

Figure A.1: Annual Directives by Category, 1877 to 2020



A.4 Distribution of significance estimates

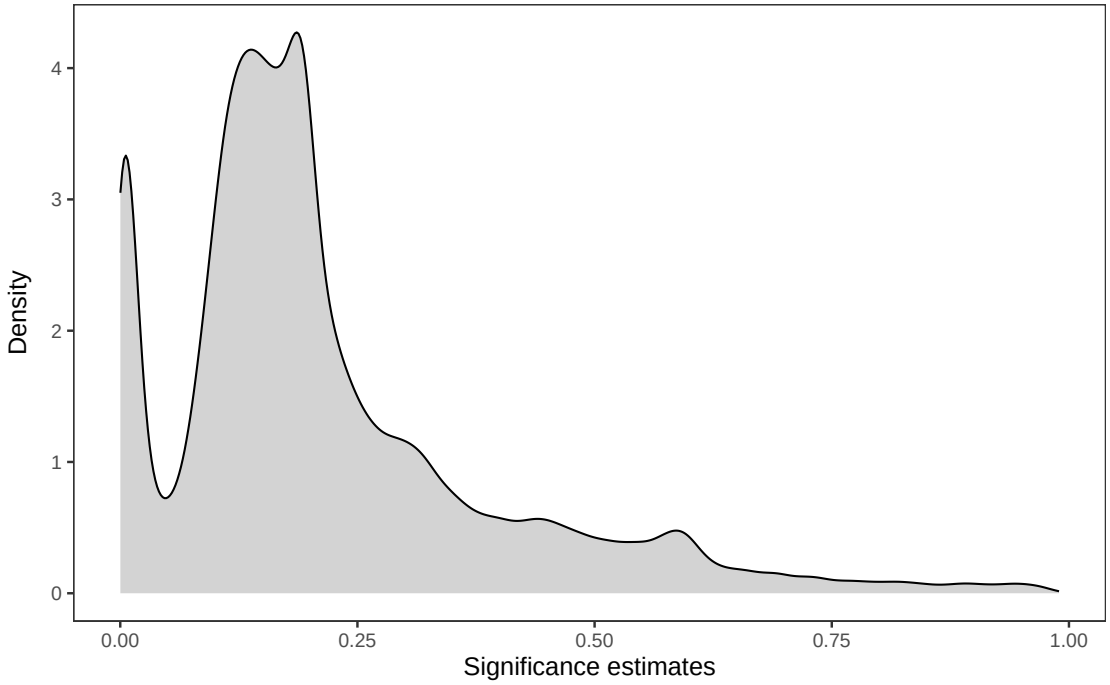


Figure A.2: Distribution of significance estimates across documents.

B Training Data, Modeling Validity, and Robustness Checks

This Appendix provides additional details about our training data construction, results from our modeling validity checks, and the robustness checks we produce.

B.1 Training Data Drawn from Chiou & Rothenberg

Our training scores rely heavily on measures of significance drawn from Chiou and Rothenberg (2017). As this is a cornerstone of our paper, we discuss some aspects of that paper and its measure here.

Chiou and Rothenberg (CR) collect 3,513 numbered executive orders from 1947 to 2003. First they hand-code 19 measures of significance drawn from contemporary and retrospective accountings: law journal articles about significant orders, newspaper mentions of executive activity, etc. Then they fit an item-response model to that data and produce a continuous score from roughly -1 to 3.5 for each executive order indicating its significance. The authors acknowledge some concerns about their scores: different raters might have preferences for different *subject matter*, such as the Wall Street Journal preferring to cover economic orders rather than social policy orders even if they are very important. To account for some of this variation CR also incorporate some other structured covariates related to both the raters themselves and other features of the orders such as the order's topic area and the political context during which it was signed.

There are several highly desirable properties of these scores. One is that they are continuous, whereas all prior measures were dichotomous (and often contradictory). This allows for finer-grained distinctions between orders' significance. Second, the CR measure correlates highly with other standard measures such as Howell (2003).

However, we also argue that CR scores are artificially precise. They are the result of a complex IRT model and are measured with error, but without attached uncertainty estimates. As such, there is a meaningful chance that an order with a score of 0.1 is more significant than an order

with a score of 0.2 due to noise in the modeling procedure.

As well, CR scores do not scale well. To produce new CR scores requires collecting extensive additional data, both rater data and associated covariates, and the rater-level covariates may change over time as well.

B.2 Robustness and Validity

In interpreting the results of our significance model, we note that it all documents are scored on the same scale: an estimated significance of 1 means the same thing for executive orders, proclamations, and memoranda. However, we expect that our measurement strategy induces *heterogeneous measurement error*. Some of our documents have their significance scores hand-coded or assumed; these have minimal measurement error if any. Some documents, on the other hand, have measurement error derived from any number of challenges: poor document transcription, language heterogeneity, the rarity of highly significant documents, and other problems generally associated with transfer learning.

We assess model accuracy through two means. The first is through cross-validated AUC, or the area under the precision-recall curve – we do this for both our significance model and our policy classifier. The second method is through comparisons to human coders, which we perform for our significance model only. Cross-validated AUC measures how well a model measures the relationship between covariates and outcomes in the training data. This is a difficult task: text-as-data methods are best suited to measuring concrete and measurement error-free concepts, while unilateral action significance is anything but concrete. Regardless, we achieve notable success in cross-validation accuracy.

When examining binary outcomes, as is the case in our significance modeling approach, the most common accuracy measures are precision and recall (Ling et al. 2003; Huang and Ling 2005). Precision is the number of correct positive identifications divided by the total number of identifications; recall is the number of correct positive identifications divided by the total number of

true positive cases. Taken together, these measures produce a Receiver operating characteristic (ROC).[28] The area under the ROC curve, called AUC, is the gold standard standard measure of predictive accuracy for binary classification tasks. In Figure B.1, we present the ROC and AUC for our model with a significance threshold of 0.5. The AUC is 0.904, on par with many of the best results in the application of machine learning to the social sciences.

This result is difficult to interpret without a relevant benchmark. Ideally, that benchmark would be the best alternative to using a machine learning model. To establish that benchmark, we trained three undergraduate research assistants to manually code unilateral actions as significant (1) or ceremonial (0) and compared those human coders' accuracy to that of our significance model.[29] We presented the research assistants with 100 executive orders that have significance scores from Chiou and Rothenberg (2017), as well as 100 other unilateral actions from our data set that did not have Chiou and Rothenberg (2017) estimates, and asked the students to code the significance of those documents. We then performed two analyses on these hand-coded significance scores. The first measures inter-coder reliability. An important advantage of machine learning models for coding documents is consistency: the model will yield a similar or identical result every time it is queried. Human coders, however, are often inconsistent. The research assistants' hand-coded executive order significance scores were not highly correlated with each

[28]Consider a model which produces probabilities that an observation is either a 1 or a 0. To measure the accuracy of this model, we must first specify a predictive cutoff. Perhaps we determine that any observation predicted to be 1 with $p > 0.5$ is a 1, and otherwise a 0, then we can measure precision and recall. However, as we vary our predictive cutoff, precision and recall change. The ROC curve captures precision and recall for all values of the predictive cutoff from 0 to 1.

[29]The undergraduate coders were asked to research the unilateral actions and assess their policy significance using their own best judgment and knowledge of the relevant historical/political context.

other. Taking the undergraduates in pairs and measuring their percent agreement at coding unilateral actions as significant or ceremonial, the three undergraduates agree with each others at rates of 65%, 71%, and 63%.

In the second analysis, we calculate AUC scores for the three sets of hand-coded documents compared to with Chiou and Rothenberg's scores for the same documents, thresholded at 0.5. If the research assistants' AUC scores, individually or aggregated, are lower than the machine learning model, then we can be confident that the machine learning model is an improvement over the current state of the art. We find that the research assistants' codings produce AUCs of 0.68, 0.67, and 0.65, each of which is substantially lower than the AUC of 0.90 produced by the machine learning model. In practice, when using research assistants to hand code noisy data, it is common to average hand codes to produce a more reliable measure. We take the elementwise average of the three hand-coded significance codes and calculate that the AUC for that aggregated coding is 0.71. These exercises suggest that our machine learning model is substantially more accurate than trained undergraduate research assistants, and provides a dramatic improvement as a consistent and scalable approach for measuring document significance.

B.3 AUC and Multi-class classification

A variant of AUC is applicable to our policy classification model as well: for each of 20 policy categories, we can calculate an AUC by dichotomizing the outcome as related to Policy X or not; by averaging each of these 20 AUC curves we can construct an average AUC. Our policy classifier receives an average AUC of 0.86, a strong score for a 20-class classification problem.

B.4 Document Transcription

Many of the documents we analyze are simply scanned images of printed pages in PDF format. We extracted text from these PDFs using Google's Tesseract 4 optical character recognition

(OCR) system. For documents with typed text, this OCR procedure produces high quality text. However, for many earlier and hand-written documents, the OCR-derived text is of poor quality. To improve the data quality in these cases, as well as in cases where more than 10% of the words are not found in a dictionary, we transcribed these documents by hand. Together, these two samples account for 5% of our total corpus. As a validity check, we transcribed 20% of this sample twice; concordance between the doubly-transcribed documents ensures us that our transcriptions are satisfactory.

B.5 Language Heterogeneity & Unrepresentative Training Sets

A critical assumption for our analysis is that the language and word choice indicative of significant executive orders is sufficiently similar to that of other types of significant unilateral directives. For example, the tone and style of significant executive orders and proclamations may be very legalistic, while important memoranda may be more rhetorical; if this is the case, then many of the textual features which contribute to a document's significance may be legalistic, biasing downward a memorandum's estimated significance. This problem may be especially severe in cases where the training set comes from a more limited set of years than the test set.

To fortify both of our model against this weakness, we expand our training set to include more representative documents. However, since we do not have significance scores or policy labels for documents other than executive orders, we infer them using a manual matching procedure. We first select a random 500 executive orders from our training set. Then, using the ProQuest Executive Actions database, we manually search for documents which reference one and only one executive order in our random sample. If we find a document which is substantively related to a single executive order, we assign that document the same significance, either 0 or 1, as the executive order it mentions. By assigning equal significance and identical policy labels to those two documents, we teach our model to recognize the significance and policy classifications of a wider variety of rhetorical styles. We find matches for 86 of the 500 executive orders in our

random sample. Many of those executive orders have multiple matching documents; as a result, our matching procedure adds 287 observations to our training data for both models.

B.6 Heteroskedastic Predictive Accuracy

A third challenge is ensuring that our models' predictions, which we aggregate into dependent variables for regressions, are not systematically biased. If the predictions are unbiased but measured with error, that measurement error will force our regression coefficients toward zero. If, however, the predictions are biased, then the regression coefficents may be artificially extreme. We can observe whether our predictions are systematically biased by examining closely our cross-validation accuracy for heteroskedasticity.

Importantly, there is little observed heteroskedasticity: our significance model's residuals are only very weakly correlated with the true significance labels ($\rho = 0.017$). However, insofar as there is heteroskedasticity, it is among the predicted high-significance documents. Documents which Chiou and Rothenberg estimate to be of high significance our model often overestimate as being highly significant, further justifying our coarsening. This is critical for performing additional analysis, as any systematic bias in our model's accuracy would subsequently bias any regression results for which we use our model's predictions.

B.7 Temporal Variation in Modeling Accuracy

A potential criticism of this significance modeling approach is that our model may underestimate the significance of documents whose text is unlike the text of numbered Executive Orders or proclamations in our training set. Consider, for example, a model trained only on data from the 1940s, used to evaluate the significance of documents from the 2010s. Due to changes in language over time, that model may be unlikely to perform well. The same result holds, though, if there are *more* documents from the 1940s than there are from the 2010s.

To test this, we perform a similar cross-validation procedure as in our main results, except instead of each fold consisting of random subsamples, each subsample is a decade of text. This allows us to test whether our model fails to accurately estimate the significance of documents from time periods outside the training set. As Figure B.1 shows, the temporal cross-validation accuracy is not substantially lower than the randomly partitioned cross-validation accuracy, providing evidence that the lexical cues indicating document significance do change over time. This gives us confidence that our model is robust to the relatively mild changes in language usage we observe, though we still acknowledge that much earlier documents pose a significant estimation challenge. However, as we discuss in the Results section, any measurement error induced by this estimation challenge should serve only to reduce the absolute magnitude of our regression coefficients.

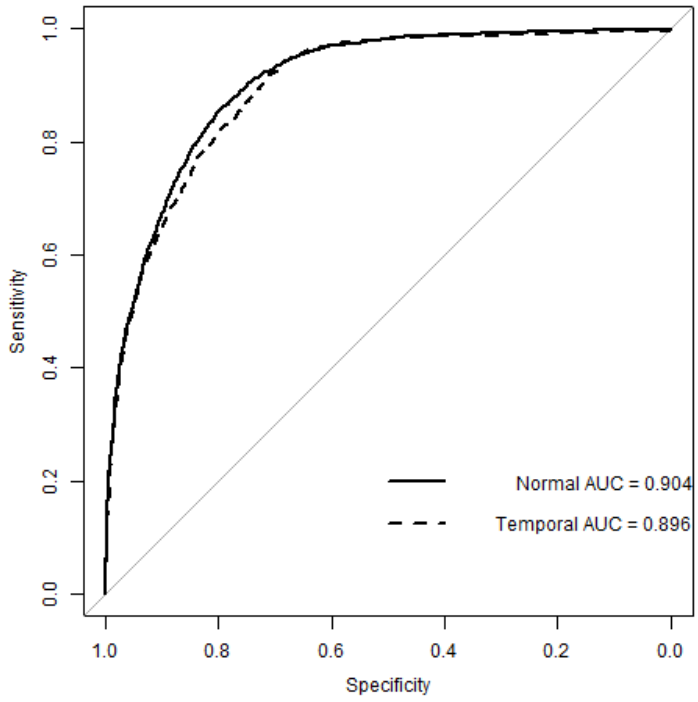


Figure B.1: The cross-validated AUC for the randomly partitioned analysis is not substantially lower than the temporally partitioned analysis, indicating that the model has significant capacity to estimate document significance outside its temporal support.

B.8 Feature Importance

As a face validity check on our significance model, we perform a descriptive feature importance analysis. Since random forest models are largely black boxes where features enter and predictions are returned, determining which covariates contribute most to the model's success can be difficult. One commonly used method to extract feature importances from tree-based models involves "feature depth" (Archer and Kimes 2008). Since random forests consist of decision trees that are ordered variable splits, features that systematically appear earlier in the decision tree are more important to the model. A covariate's feature importance, then, is proportional to the average number of times that feature appears in the decision tree, weighted by how early in the tree it appears; more simply, higher values indicate more strongly predictive features.

Note that this is purely a descriptive exercise; if our model performs as we intend, we expect that the most important features it identifies relating to policy significance will be ones which intuitively discriminate significant orders, which discuss tariffs, military conflict, and industry, from ceremonial ones, which memorialize the dead or declare National Ice Cream Day. We calculate feature importance for random forests model and present the 30 most important terms below in columns 1 and 2 of Table B.1, then 20 largely insignificant terms in columns 3 and 4. Among the most significant words are section, persuant, provision, necessary, and articles, which generally indicate appeals to either constitutional or statutory authority. Low-significance terms include liberty, anniversary, Thanksgiving, bravery, and victims. A feature importance table derived only from executive orders follows in Table B.1.

An initial inspection of our approach to coding presidential documents, therefore, suggests that we have uncovered a meaningful dimension that distinguishes actions based on whether they address consequential policy issues or more are ceremonial in nature.

Similarly, we expect a different but related set of important features for our policy classifier: terms related to the policies in question. We find that housing, education, land, labor, defense, agricultural, and discrimination are all important policy-related key words; advancing, stronger,

undermine, desired, and everywhere are policy-unrelated ones (Table B.2).

Term	Importance	Term	Importance
section	6.456	billion	-0.067
pursuant	5.692	liberty	-0.068
provisions	4.422	working	-0.068
necessary	3.946	remember	-0.069
proclamation	3.616	anniversary	-0.069
including	3.561	sacrifice	-0.07
withdrawn	3.342	choices	-0.07
entered	3.302	countless	-0.07
warehouse	2.976	college	-0.07
consumption	2.85	greatest	-0.07
follows	2.771	patients	-0.07
purposes	2.545	victims	-0.07
modified	2.545	mothers	-0.071
respect	2.196	productive	-0.072
effective	2.176	courage	-0.073
americans	2.081	generations	-0.074
secretary	2.043	generation	-0.075
authority	1.999	substances	-0.075
provided	1.961	disabled	-0.077
determined	1.955	helping	-0.077
actions	1.955	challenges	-0.078
articles	1.915	thanksgiving	-0.079
certain	1.862	safeguard	-0.08
register	1.832	equality	-0.082
executive	1.823	millions	-0.084
modifications	1.811	inspiration	-0.084
limited	1.757	success	-0.085
schedule	1.739	transition	-0.086
amended	1.729	bravery	-0.086
imported	1.709	teachers	-0.088

Table B.1: Random Forest feature importances for the significance model.

Term	Importance	Term	Importance
housing	100.944	preclude	0.113
recreational	47.765	attended	0.113
education	46.919	discipline	0.112
revokes	42.538	realized	0.112
reserved	41.118	presente	0.112
refugees	35.577	pittsburgh	0.11
interior	34.484	everywhere	0.109
executive	33.678	advancing	0.109
agriculture	33.131	ciation	0.108
educational	31.274	package	0.107
defense	31.223	articte	0.107
section	30.958	disclosed	0.106
agricultural	30.232	benjamin	0.105
president	29.819	complaint	0.104
appointment	29.419	managing	0.102
discrimination	27.994	proximately	0.102
federal	27.564	madison	0.102
secretary	27.555	christopher	0.102
meridian	26.809	providers	0.101
announced	26.03	conveyed	0.101
approved	24.888	shipped	0.1
roosevelt	24.791	stronger	0.1
franklin	23.716	desiring	0.099
ordered	23.6	empower	0.098
dispute	21.912	specification	0.098
national	21.504	remarkable	0.097
manager	21.491	attempted	0.096
amended	21.414	sanction	0.096
sections	21.35	undermine	0.096
service	21.096	nominate	0.091

Table B.2: Random Forest feature importances for the policy classification model.

B.9 Robustness to Threshold Choices

In two places in our measurement strategy we identify thresholds for indicating significant documents. First, we select a threshold of 0.50 in coarsening Chiou & Rothenberg's significance scores into a dichotomous indicator. Second, after generating predictions from our random forest model, we identify documents with a probability of ≥ 0.355 as being significant.

The first threshold of 0.50 is the result of our qualitative reading of a large number of these documents. As we argue in the Section "Measuring the Significance and Policy Domains of Directives," our decision to threshold here reduces measurement error vis-á-vis Chiou and Rothenberg (2017). In general, coarsening removes valuable information. But when coarsening removes inappropriate precision, it can be beneficial. In this case, we do not believe that Chiou & Rothenberg's scores' continuous values from -1 to 3 are appropriately precise, and the measurement error in those scores is not quantified. Therefore if we were to feed the raw Chiou & Rothenberg scores to our supervised learner, it would overfit to the noise in those scores scores and result in worse modeling accuracy. We still believe there is a very strong signal in the their data, but we do not believe that there is a meaningful difference between an order that they rate as 0.1 and an order they rate as 0.2.

Secondly, coarsening turns our modeling problem from a regression problem on a scale from -1 to 3 with difficult interpretability to a much easier and more reliable classification problem. The outcome of our model can be interpreted as the probability that a document would receive a Chiou and Rothenberg score of greater or less than our threshold of 0.50, which is straightforward. The choice of the threshold itself at 0.50, rather than at 0 or at 1, is the result of extensive qualitative readings of these documents.

The second threshold is of the modeling results. Our model produces probabilities that a document would receive a Chiou and Rothenberg score of greater or less than 0.50. To convert these probabilities into a prediction of whether a document is significant, we identify a second

threshold.[30] We identify 0.355 as the estimated probability at which we declare a document "significant." While we could reasonably choose 0.50, or 0.95, or many other thresholds, we select 0.355 since it is the value which equalizes the false-positive rate and the false-negative rate. This is a desirable property since it mitigates bias and heteroskedasticity in our regression analyses.

While we perform robustness checks to show that our substantive results are not sensitive to our choices of these thresholds, we acknowledge that these choices induce "researcher degrees of freedom" (Simmons, Nelson and Simonsohn 2011). In the interests of transparency and reproducibility, our complete replication file will be made public on the Harvard Dataverse upon publication. Importantly, the replication code will include clearly demarcated points where we perform our coarsening allowing interested researchers to experiment with alternative thresholding decisions.

[30]Note that measuring AUC as above does not require selecting a threshold here.

B.10 Face Validity

Table B.3: Presidential Proclamations in 2019 (* denotes ceremonial)

#	Title	Score	Issue Area
9836	Religious Freedom Day*	0.04	Govt operations
9837	National School Choice Week*	0.01	Govt operations
9838	National Sanctity of Human Life Day*	0.01	Govt operations
9839	Martin Luther King, Jr., Federal Holiday*	0.02	Govt operations
9840	American Heart Month*	0.01	Govt operations
9841	National African American History Month*	0.01	Govt operations
9842	Addressing Mass Migration Through the Southern Border of the United States	0.71	Immigration
9843	Death of John David Dingell, Jr.*	0.02	Govt operations
9844	Declaring a National Emergency Concerning the Southern Border of the United States	0.53	Defense
9845	American Red Cross Month*	0.01	Govt operations
9846	Irish-American Heritage Month*	0.02	Govt operations
9847	Women's History Month*	0.03	Govt operations
9848	National Consumer Protection Week*	0.01	Govt operations
9849	National Agriculture Day*	0.01	Govt operations
9850	National Poison Prevention Week*	0.01	Govt operations
9851	Greek Independence Day: A National Day of Celebration of Greek and American Democracy*	0.01	Govt operations
9852	Recognizing the Golan Heights as Part of the State of Israel	0.06	Govt operations
9853	Cancer Control Month*	0.01	Govt operations
9854	National Child Abuse Prevention Month*	0	Govt operations
9855	National Donate Life Month*	0	Govt operations
9856	National Sexual Assault Awareness and Prevention Month*	0	Govt operations
9857	Second Chance Month*	0.01	Govt operations
9858	World Autism Awareness Day*	0	Govt operations
9859	National Crime Victims' Rights Week*	0.02	Govt operations
9860	National Volunteer Week*	0	Govt operations
9861	National Former Prisoner of War Recognition Day*	0.01	Govt operations
9862	Pan American Day and Pan American Week*	0.02	Govt operations
9863	Education and Sharing Day, U.S.A.*	0	Govt operations
9864	National Park Week*	0.01	Govt operations
9865	World Intellectual Property Day*	0.01	Govt operations
9866	Days of Remembrance of Victims of the Holocaust*	0.01	Govt operations
9867	Asian American and Pacific Islander Heritage Month*	0.01	Govt operations
9868	Jewish American Heritage Month*	0.02	Govt operations
9869	National Foster Care Month*	0.01	Govt operations
9870	National Physical Fitness and Sports Month*	0.01	Govt operations
9871	Older Americans Month*	0	Govt operations
9872	Law Day, U.S.A.*	0.02	Govt operations
9873	Loyalty Day*	0.02	Govt operations
9874	National Day of Prayer*	0	Govt operations
9875	National Mental Health Awareness Month*	0.03	Govt operations
9876	National Hurricane Preparedness Week*	0.03	Govt operations
9877	National Small Business Week*	0.01	Govt operations
9878	Public Service Recognition Week*	0.01	Govt operations
9879	Missing and Murdered American Indians and Alaska Natives Awareness Day*	0.03	Govt operations
9880	Addressing Mass Migration Through the Southern Border of the United States	0.7	Immigration
9881	Military Spouse Day*	0.01	Govt operations
9882	National Charter Schools Week*	0	Govt operations
9883	National Defense Transportation Day and National Transportation Week*	0.01	Govt operations
9884	Peace Officers Memorial Day and Police Week*	0	Govt operations
9885	Mother's Day*	0.01	Govt operations

Table B.3 (continued): Presidential Proclamations in 2019 (* denotes ceremonial)

#	Title	Score	Issue Area
9886	Adjusting Imports of Steel Into the United States	0.8	Trade
9887	To Modify the List of Beneficiary Developing Countries Under the Trade Act of 1974	0.88	Intl affairs
9888	Adjusting Imports of Automobiles and Automobile Parts Into the United States	0.62	Trade
9889	National Safe Boating Week*	0.01	Govt operations
9890	Emergency Medical Services Week*	0.01	Govt operations
9891	World Trade Week*	0.03	Govt operations
9892	Armed Forces Day*	0.01	Govt operations
9893	Adjusting Imports of Aluminum Into the United States	0.74	Trade
9894	Adjusting Imports of Steel Into the United States	0.8	Trade
9895	National Maritime Day*	0.01	Govt operations
9896	Prayer for Peace, Memorial Day*	0.01	Govt operations
9897	African-American Music Appreciation Month*	0	Govt operations
9898	Great Outdoors Month*	0.02	Govt operations
9899	National Caribbean-American Heritage Month*	0	Govt operations
9900	National Homeownership Month*	0.01	Govt operations
9901	National Ocean Month*	0.02	Govt operations
9902	To Modify the List of Beneficiary Developing Countries Under the Trade Act of 1974	0.89	Trade
9903	Honoring the Victims of the Tragedy in Virginia Beach, Virginia*	0.01	Govt operations
9904	National Day of Remembrance of the 75th Anniversary of D-Day*	0.03	Govt operations
9905	Flag Day and National Flag Week*	0	Govt operations
9906	Father's Day*	0.01	Govt operations
9907	Pledge to America's Workers Month*	0.01	Govt operations
9908	Made in America Day and Made in America Week*	0.01	Govt operations
9909	Death of John Paul Stevens*	0.02	Govt operations
9910	Captive Nations Week*	0.01	Govt operations
9911	50th Anniversary Observance of the Apollo 11 Lunar Landing*	0.03	Govt operations
9912	Anniversary of the Americans with Disabilities Act*	0.01	Govt operations
9913	National Korean War Veterans Armistice Day*	0.03	Govt operations
9914	Honoring the Victims of the Tragedies in El Paso, Texas, and Dayton, Ohio*	0.02	Govt operations
9915	National Employer Support of the Guard and Reserve Week*	0.01	Govt operations
9916	Women's Equality Day*	0	Govt operations
9917	National Alcohol and Drug Addiction Recovery Month*	0.01	Govt operations
9918	National Childhood Cancer Awareness Month*	0.01	Govt operations
9919	National Preparedness Month*	0.01	Govt operations
9920	Labor Day*	0.02	Govt operations
9921	National Days of Prayer and Remembrance*	0.01	Govt operations
9922	National Historically Black Colleges and Universities Week*	0.01	Govt operations
9923	Opioid Crisis Awareness Week*	0.04	Govt operations
9924	Minority Enterprise Development Week*	0.02	Govt operations
9925	Patriot Day*	0.01	Govt operations
9926	National Farm Safety and Health Week*	0.01	Govt operations
9927	National Hispanic Heritage Month*	0.01	Govt operations
9928	National Gang Violence Prevention Week*	0.03	Govt operations
9929	Constitution Day, Citizenship Day, and Constitution Week*	0.03	Govt operations
9930	National POW/MIA Recognition Day*	0.01	Govt operations
9931	Suspension of Entry as Immigrants and Nonimmigrants of Persons Responsible for Policies or Actions That Threaten Venezuela's Democratic Institutions	0.73	Immigration
9932	Suspension of Entry as Immigrants and Nonimmigrants of Senior Officials of the Government of Iran	0.73	Immigration
9933	National Domestic Violence Awareness Month*	0	Govt operations
9934	Gold Star Mother's and Family's Day*	0	Govt operations
9935	National Hunting and Fishing Day*	0	Govt operations
9936	National Breast Cancer Awareness Month*	0	Govt operations

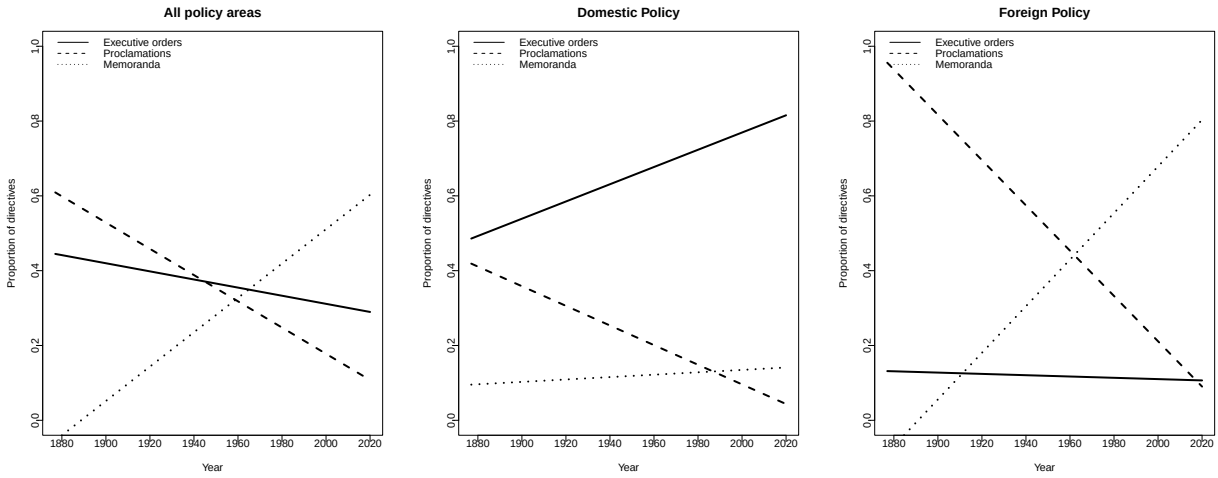
Table B.3 (continued): Presidential Proclamations in 2019 (* denotes ceremonial)

#	Title	Score	Issue Area
9937	National Cybersecurity Awareness Month*	0.01	Govt operations
9938	National Disability Employment Awareness Month*	0.02	Govt operations
9939	National Energy Awareness Month*	0.01	Govt operations
9940	National Substance Abuse Prevention Month*	0	Govt operations
9941	National Manufacturing Day*	0.02	Govt operations
9942	Fire Prevention Week*	0.02	Govt operations
9943	German-American Day*	0.01	Govt operations
9944	Child Health Day*	0.01	Govt operations
9945	Suspension of Entry of Immigrants Who Will Financially Burden the United States Healthcare System, in Order To Protect the Availability of Healthcare Benefits for Americans	0.72	Immigration
9946	Leif Erikson Day*	0	Govt operations
9947	General Pulaski Memorial Day*	0.01	Govt operations
9948	National School Lunch Week*	0	Govt operations
9949	Columbus Day*	0.01	Govt operations
9950	Blind Americans Equality Day*	0	Govt operations
9951	Death of Elijah E. Cummings*	0.02	Govt operations
9952	National Character Counts Week*	0	Govt operations
9953	National Forest Products Week*	0.01	Govt operations
9954	United Nations Day*	0.02	Govt operations
9955	To Modify Duty-Free Treatment Under the Generalized System of Preferences and for Other Purposes	0.85	Trade
9956	Critical Infrastructure Security and Resilience Month*	0.05	Govt operations
9957	National Adoption Month*	0	Govt operations
9958	National American History and Founders Month*	0	Govt operations
9959	National Entrepreneurship Month*	0.02	Govt operations
9960	National Family Caregivers Month*	0.02	Govt operations
9961	National Native American Heritage Month*	0.01	Govt operations
9962	National Veterans and Military Families Month*	0.01	Govt operations
9963	Veterans Day*	0.02	Govt operations
9964	National Apprenticeship Week*	0.01	Govt operations
9965	World Freedom Day*	0.01	Govt operations
9966	American Education Week*	0.01	Govt operations
9967	National Family Week*	0.01	Govt operations
9968	Thanksgiving Day*	0.01	Govt operations
9969	National Impaired Driving Prevention Month*	0.01	Govt operations
9970	World AIDS Day*	0.02	Govt operations
9971	National Pearl Harbor Remembrance Day*	0.03	Govt operations
9972	Human Rights Day, Bill of Rights Day, and Human Rights Week*	0.02	Govt operations
9973	Wright Brothers Day*	0.01	Govt operations
9974	To Take Certain Actions Under the African Growth and Opportunity Act and for Other Purposes	0.87	Trade
9975	National Slavery and Human Trafficking Prevention Month*	0.03	Govt operations
	Average significance, nonceremonial proclamations: 0.55 Average significance, ceremonial proclamations: 0.03		

C Alternative Polynomial Specifications

Figure C.1: Directive Substitution across Policy Areas

(a) Linear specification



(b) Quadratic specification

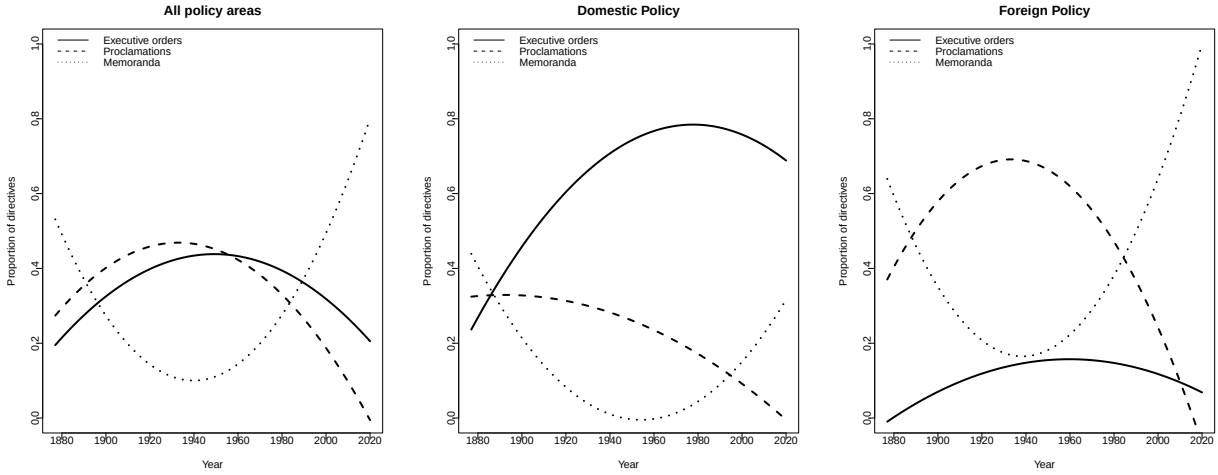
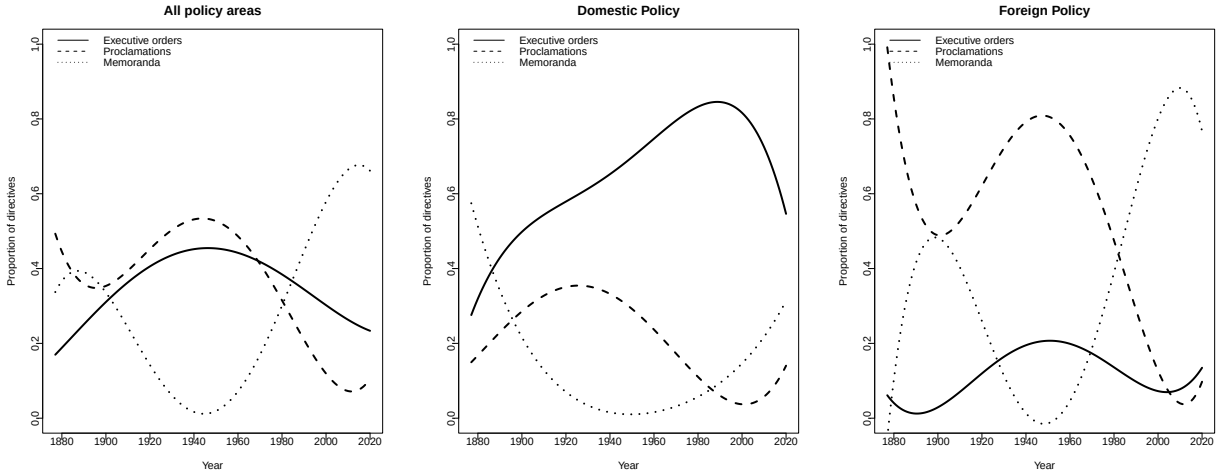
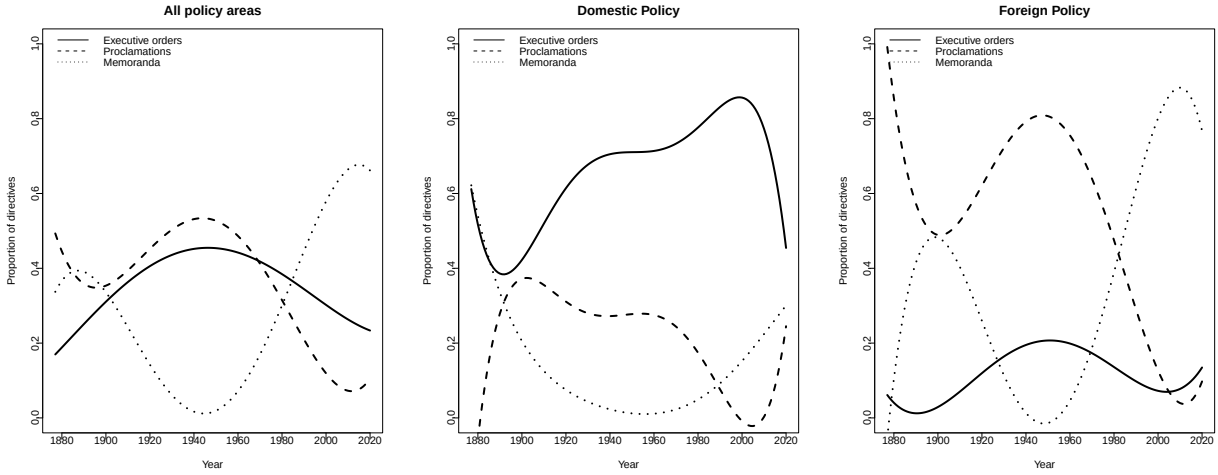


Figure C.2: Directive Substitution across Policy Areas

(a) Fourth-degree polynomial



(b) Fifth-degree polynomial



D Additional Results: Time Series Analysis

This section includes additional results related to our time series analyses of unilateral directives over time.

D.1 Factor analysis

A central question in our analysis of patterns of unilateral action relates to aggregating different types of unilateral directives into a pooled count of total orders, and possible hazards that accompany that aggregation (Bartels 1996). To provide evidence that these directives are meaningfully similar and can be aggregated safely, we perform a principal components analysis.

We decompose a T by k matrix where each row is a year from 1877 to 2021 and the three columns each contain the count of yearly significant directives categorized as executive orders, memoranda, and proclamations respectively. We find that the first principal component explains 66% of the variation in directive counts over time, the second dimension explains an additional 22%, and the final dimension explains 12%.

Interestingly, when we split this matrix before 1933, when Franklin Roosevelt was inaugurated, we find starkly different results. Prior to 1933, the first dimension explains approximately 53% of the variation, whereas it explains 87% afterwards, suggesting that the modern presidents (who issued by far the bulk of directives) used the different tools as substitutes more than their predecessors.

D.2 Nonparametric estimation and Disaggregated Changepoints

Our primary method for estimating structural change points in the yearly counts of significant unilateral directives is with a linear model via the `strucchange` library in R (Zeileis et al. 2002). However, since our data set contains count data with significant changes in magnitude and variance over time, some of the OLS assumptions may not hold. To show that our change points

are robust to model specification, we also implement a fully non-parametric method of identifying structural change points from James and Matteson (2013) that is designed to make "as few assumptions as possible" (see also Matteson and James 2014).To parallel our parametric procedure, we allow the model to identify both the number and location of change points, specifying only the minimum length of period, which is four years.

Under this specification, the `ecp` method identifies two change points at 1904 and 1992. In contrast, the linear model finds six change points at 1903, 1909, 1913, 1919, 1991, and 2016.

A second advantage of the nonparametric estimation method is that it is *multinomial*: it can detect change points in multiple time series outcomes simultaneously. To use this advantage we disaggregate our count of significant unilateral orders to significant executive orders, significant memoranda, and significant proclamations and ask the model to detect change points in those three jointly. The results of this analysis are confirming: it again identifies two change points at 1904 and 1992.

Finally, we experiment with estimating structural change points for each directive type *separately*. For executive orders, the model identifies 1904, 1915, 1920, 1927, and 1988 as change points; 1909, 1913, 1919, and 1992 for memoranda; and 1905, 1911, and 1917 for proclamations.

E Additional details related to replications

E.1 Additional tests related to replication of Christenson and Kriner (2019)

Table E.1: Testing for Stationarity

	Augmented Dickey-Fuller		Phillips-Perron	
	Test statistic	p	Test statistic	p
Approval	-3.05	.031	-4.17	<.001
Monthly count, all	-16.70	<.001	-18.39	<.001
Summed significance, all	-17.51	<.001	-19.09	<.001
Monthly count, EOs	-22.03	<.001	-22.71	<.001
Summed significance, EOs	-23.39	<.001	-23.92	<.001
Monthly count, MMs	-14.99	<.001	-16.35	<.001
Summed significance, MMs	-14.58	<.001	-15.75	<.001
Monthly count, PRs	-24.41	<.001	-24.60	<.001
Summed significance, PRs	-25.48	<.001	-25.62	<.001

Entries show test statistics and p-values from Dickey-Fuller (left columns) and Phillips-Perron (right columns) tests. In all cases we can reject the null hypothesis that the monthly approval rating, count of significant directives, and summed directive significance have a unit root. Instead, the tests reveal the data to be stationary.

Table E.2: Granger-Causality Tests: Presidential Approval and Unilateral Directives, 1953-2018

	Number of significant directives			Aggregated significance		
	χ^2	df	p	χ^2	df	p
Executive orders only, from our data						
Orders → approval	1.234	2	.539	2.013	2	.365
Approval → orders	10.373	2	.006	12.468	2	.002
Memoranda only, from our data						
Orders → approval	1.302	2	.521	5.223	2	.073
Approval → orders	0.005	2	.998	0.757	2	.685
Proclamations only, from our data						
Orders → approval	2.944	2	.230	2.435	2	.296
Approval → orders	0.741	2	.691	0.740	2	.691

Entries show results from Granger causality tests conducted with coefficients from vector autoregression models. The rows labeled Orders → approval examine the hypothesis that unilateral action Granger causes presidential approval by testing whether the coefficients on two lags of unilateral directives are jointly zero. The rows labeled Approval → orders examine the hypothesis that presidential approval Granger causes unilateral action by testing whether the coefficients on two lags of unilateral directives are jointly zero. The left side of the table shows results when using the monthly count of significant directives and the right side shows results when using the summed significance of directives issued in a month. Entries show results using directives in our data for the same time period and model specifications as in Christenson and Kriner (2019).

E.2 Full results for Djourelova and Durante (2022, Table 3) replication

Table E.3: News Pressure and the Timing of Unilateral Action: Divided versus Unified Government (1979–2016)

	Full Sample				Divided Government			Unified Government		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Any directive										
NP	0.019	0.005	0.003	-0.011	0.021	0.007	0.007	0.016	0.004	-0.003
	(0.018)	(0.019)	(0.020)	(0.041)	(0.021)	(0.022)	(0.022)	(0.038)	(0.041)	(0.041)
NP (t+1)	0.019	0.015	0.010	0.020	0.011	0.008	0.006	0.043	0.037	0.023
	(0.018)	(0.018)	(0.019)	(0.042)	(0.020)	(0.020)	(0.022)	(0.038)	(0.040)	(0.042)
NP (t-1)		0.043*	0.042*	0.007		0.052*	0.051*		0.012	0.007
		(0.019)	(0.019)	(0.038)		(0.022)	(0.022)		(0.037)	(0.037)
NP × divided				0.020						
				(0.046)						
NP (t+1) × divided				-0.014						
				(0.046)						
NP (t-1) × divided				0.046						
				(0.044)						
Panel B: Any significant directives										
NP	-0.014	-0.015	-0.016	-0.020	-0.021	-0.015	-0.014	0.008	-0.015	-0.019
	(0.013)	(0.014)	(0.014)	(0.027)	(0.016)	(0.017)	(0.017)	(0.025)	(0.027)	(0.027)
NP (t+1)	0.021	0.020	0.018	-0.041	0.032	0.036*	0.036*	-0.006	-0.023	-0.036
	(0.014)	(0.015)	(0.015)	(0.028)	(0.016)	(0.017)	(0.018)	(0.028)	(0.028)	(0.029)
NP (t-1)		0.009	0.009	0.021		0.002	0.003		0.027	0.024
		(0.016)	(0.016)	(0.030)		(0.018)	(0.018)		(0.030)	(0.031)
NP × divided				0.006						
				(0.032)						
NP (t+1) × divided				0.078*						
				(0.033)						
NP (t-1) × divided				-0.017						
				(0.035)						
N	13875	13854	13836	13836	10133	10126	10114	3742	3728	3722
7 lags of NP	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes
7 leads of NP	No	No	Yes	Yes	No	No	Yes	No	No	Yes
Weeks in office	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year, month, DoW FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
7 leads and lags of NP × divided	No	No	No	Yes	No	No	No	No	No	No

NP = news pressure. DoW = day of week. FEs = fixed effects.

Full sample in columns (1)–(4), divided government in columns (5)–(7), unified government in columns (8)–(10).

The dependent variable is an indicator for the signing of any directive (Panel A) or a significant directive (Panel B).

OLS regressions in all columns. Standard errors clustered by month × year. * $p < 0.05$.

Appendix References

Archer, Kellie J. and Ryan V. Kimes. 2008. "Empirical Characterization of Random Forest Variable Importance Measures." *Computational Statistics & Data Analysis* 52:2249–2260.

Bartels, Larry M. 1996. "Pooling Disparate Observations." *American Journal of Political Science* 40(3):905–942.

Chiou, Fang-Yi and Lawrence S. Rothenberg. 2017. *The Enigma of Presidential Power: Parties, Policies and Strategic Uses of Unilateral Action.* New York: Cambridge University Press.

Huang, Jin and Charles X Ling. 2005. "Using AUC and accuracy in evaluating learning algorithms." *IEEE Transactions on knowledge and Data Engineering* 17(3):299–310.

James, Nicholas A and David S Matteson. 2013. "ecp: An R package for nonparametric multiple change point analysis of multivariate data." *arXiv preprint arXiv:1309.3295* .

Ling, Charles X, Jin Huang, Harry Zhang et al. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In *IJCAI.* Vol. 3 pp. 519–524.

Matteson, David S. and Nicholas A. James. 2014. "A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data." *Journal of the American Statistical Association* 109(505):334–345.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22(11):1359–1366.

Zeileis, Achim, Friedrich Leisch, Kurt Hornik and Christian Kleiber. 2002. "strucchange: An R package for testing for structural change in linear regression models." *Journal of statistical software* 7:1–38.