Using Deep and Active Learning Classifiers to Identify Congressional Delegation to Administrative Agencies*

Austin Bussing[1], Joshua Y. Lerner[2], and Gregory P. Spell[3]

[1]Department of Political Science, Sam Houston State University
[2]Data Scientist, NORC at the University of Chicago
[3]Department of Electrical and Computer Engineering, Duke University

June 21, 2022

Abstract

Congressional oversight of the federal bureaucracy remains key to understanding implementation of the law. Essential to this are theories of how and why Congress delegates powers to administrative agencies. Using an active learning convolutional neural network on bill text, we classify bill sections by their role in delegating to administrative agencies, applying an iteratively improving coding scheme that enhances existing supervised learning approaches. We systematically study the statutory scope of administrative agencies and develop a first-of-its-kind dataset to study how executive agency characteristics and changes in presidential administrations affect congressional decisions about delegation. First, we benchmark our measure against existing proxies for delegation. We then find evidence that variation in delegatory activity is driven more by cross-agency differences in ideology and structural independence than by shifts between unified and divided government. Our findings call for a reevaluation of the ally principle, and a reexamination of interbranch relations that takes into account the internal complexity of the executive branch. We conclude with a discussion of delegatory scope and other extensions of our method and data.

Word Count: 9875

*We would like to thank the following people for their helpful comments on earlier versions: Mathew McCubbins, John Aldrich, Kristen Renberg, Kiran Auerbach, Hannah Ridge, Robert Shaffer, Justin Grimmer, Michael Crespin, Sarah Bouchat, and attendees at PolMeth 2018 and APSA 2018.

Delegation of powers from political actors to agents or agencies tasked with enacting policies remains at the heart of the challenges of modern governance. Even beyond the formal substance of the powers, any delegation has two essential elements that a coherent measurement approach must deal with: the identity of the agents who hold policy-making authority due to the delegation. For example, Congress may designate that the Environmental Protection Agency shall promulgate rules and regulations to protect endangered species, but only after issuing the rules and allowing for public comment. As important as it is, delegating legislation is not the only kind of law enacted by governments, and researchers have applied selection criteria to yield a small sample on which to apply a labor-intensive coding framework.

While this conceptualization is well known, applications of this concept have always run into a problem of measurement. Delegation is fundamentally a textual act: the legislature restricts or enables an agency in their actions through written law and is codified subsequently in oversight hearings, mark-up procedures, court cases, and more. Consequently, even the most rudimentary measurement of delegation has to grapple with the written word, either directly or indirectly, to capture what in particular is occurring. In any individual piece of legislation, this task is straightforward; identifying the agency and the change to their authority requires only a close reading of the bill. However, this approach has severe limitations for large-scale applications and empirical studies – namely the time and effort such an undertaking would require.

Innovations in machine learning and natural language processing simplify scaling up these connections. In particular, these methods are capable of predicting qualitative labels from quantifiable features. After being trained on a set of hand-labeled examples, the learning algorithm is deployed to predict labels for a much larger data set, significantly reducing the up-front costs of setting up a human-run labeling project. However, most machine learning operations separate the hand-coding from the evaluation process and treat it merely as a single attempt at attaining accuracy. Complex concepts will be hard to translate into a simple coding scheme, and thus machine coding done this way can have additional problems.

In this paper, we argue that researchers can effectively combine their expertise through an interactive machine-learning framework that will, in essence, learn on the go. More commonly referred to as an "active learning" approach to classification, we argue that this framework best combines the portability and flexibility of machine learning with the expert evaluation usually used in more modestly defined approaches, such as case studies.

To demonstrate the utility of this framework, we tackle a canonical problem in political science: how congressional delegation decisions are affected by switches between unified and divided government. We create a new dataset on legislative delegation during the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congresses—a time period encompassing a change from divided government to unified government. Additionally, we use measures of agency ideology and agency independence to determine whether variation in congressional delegation decisions is better explained by these stable agency-level characteristics or by changes in presidential administration.

We address congressional delegation for two main reasons. First, it is substantively important, as delegation facilitates the implementation of laws written by Congress and the bureaucracy represents the most direct way any average citizen interacts with the federal government. Second, theories of delegation in Congress abound in political science, economics, and public administration, but there have been few large-n empirical tests of these theories. There are numerous competing, highly comprehensive theories about how, why, and when Congress delegates authority, but up until now, most studies either focused only on a single bit of policy (Huber & Shipan 2002), a handful of "significant" bills (Epstein & O'Halloran 1999), or forewent empirical analysis altogether and focused on model building (see Gailmard & Patty 2012a, for a comprehensive overview of such models). Having a measure that is both broad and readily comparable should allow us to evaluate parts of these theories as well as further research into the relationship between Congress and the bureaucracy.

This paper proceeds as follows: first, we discuss the delegation of authority from

Congress to administrative agencies and explain why it has always posed a unique measurement challenge. Second, we address this measurement problem by using a deep and active learning model, with discussions of what makes this different from other machine learning approaches. We then assess classification accuracy and discuss the performance of our active learning model, how it improves upon existing models, and what can be done with our newly labeled dataset. We conclude by using our measure of delegation to gain new leverage on the ally principle—the theoretical proposition that variation in delegation is explained by the degree of ideological disagreement between principal and agent.

Congressional Delegation to Administrative Agencies

Delegation is a necessity for the operation of any large-scale enterprise, least of all governments. Political scientists generally view congressional delegation as a trade-off between efficiency and accountability (Kiewiet & McCubbins 1991; Epstein & O'Halloran 1999). Hypothetically, congressional delegation can have enormous productivity gains for both individual members of Congress and administrative agencies. However, in the course of designing delegation, Congress faces numerous internal coordination problems. Further, Congress must account for how delegation can create opportunities for agencies to act against its interests — standard in all principal-agent models. For these reasons — the institutional, partisan, and policy-making nature — congressional delegation has been a focal point for the study of political institutions. Scholars have formalized the intuition behind delegation through versions of the "ally principle" (Epstein & O'Halloran 1999; Huber & Shipan 2002; Moe 2012; Farhang & Yaver 2016), which argues that when the executive's interests are aligned with those of the legislature, legislators are more willing to pass legislation that delegates significant authority to agencies. By contrast, when legislative and executive policy interests diverge, legislators favor institutional structures that provide increased oversight opportunities.

Beyond institutional factors, some have posited that the design of agency authority is affected by characteristics of the issues and policy areas addressed (McCubbins 1985; Epstein & O'Halloran 1999). While the role of Congressional process and policy areas has been discussed extensively in the theoretical literature on allocating authority to agencies, these concepts are understudied in empirically oriented scholarship. This limitation results from a measurement problem, even for motivated researchers: reading and interpreting legal texts is labor-intensive. Thus, most empirical work on the allocation of authority has been restricted to single policy areas or to small sets of "significant" legislation. Given the limitations of previous empirical studies on delegation, our work to create a generalized dataset measuring delegation should fill an essential hole in the literature and provide fertile ground for the testing of new and old theories.

Difficulty of Measuring Delegation

While delegation has always been an important topic in political science, how it has been measured has attracted some controversy. Through most of the 1980s, congressional delegation to the executive branch mainly was evidenced through the use of case studies; close textual readings of bills, but nothing directly lending itself to a measurement strategy (McCubbins *et al.* 1987, 1989; Kiewiet & McCubbins 1991).

Epstein & O'Halloran (1996) pushed the literature to measure administrative discretion in individual acts of delegation by Congress by comparing the number of provisions that grant and constrain the president's authority. A limitation with the Epstein & O'Halloran (1999) approach is that it did not directly scale: they were only able to look at a small subset of legislation that became law. For as comprehensive as their book is, they only analyze 262 bills.

The next major development of delegation as a measurement comes from Huber & Shipan (2002), who use bill length — the total number of words in a bill — as a proxy for how much detail the legislature leaves for the administrative agency. They assert that this measure captures how much discretion is left for agencies to interpret the bill, and how much freedom they have to act on those interpretations. Huber & Shipan (2002) claim that longer bills are more likely to require specific returns from agencies, and this generally entails more fine-grained oversight and control. Shorter bills, ones that discuss delegation and discretion more broadly, are less likely to have the same requirements and instead reflect less deference from the legislature. It has been accepted by the literature (see Clinton *et al.* (2012) as an example) that longer bills mean added delegation.

In some ways, our approach marries the generalizability of Epstein & O'Halloran (1999) and Huber & Shipan (2002) with the explicitly close reading based approaches of earlier studies, in particular McCubbins *et al.* (1987) and Kiewiet & McCubbins (1991). Because of the textual nature of the task, we expect the classification of delegation in legislation to work well in a deep and active learning environment. As will be discussed later, active learning is most often utilized in machine learning settings where getting additional training labels is costly. In this case, we can see precisely why hand-coding delegation in bills would be an apropos application of this approach. First, it is a task that requires both training and familiarity with how Congress writes bills: the use of statutory language is always done deliberately and with the mind that the courts, the Executive Branch, and future Congresses will have to interpret specifically what was written (McNollgast 1994). Furthermore, it is in Congress's best interest to standardize the writing of bills such that little is open to interpretation (McCubbins *et al.* 1987).

Secondly, although relying on the uniformity of statutory language, this task still requires a fair amount of careful reading, given the agencies' idiosyncrasies. This uncertainty, combined with the simple classification scheme, gives us a situation where coding any given section as delegatory is labor-intensive but straightforward to scale up the coding. Thus, the need for active learning: if we achieve reasonable accuracy in this classification task without hand-coding too many sections, identifying which sections improve accuracy efficiently is critical. We believe that this exact issue is not unique to coding agency delegation in Congress but would apply to many classification tasks measuring latent concepts. This approach reduces the need for additional computational bottlenecks for such tasks and could make classifying abstract features in large documents much more tractable.

An Active Learning Convolutional Neural Network for Classifying Text

Classification is one of the most popular objectives in all text-as-data work. Generally falling under the umbrella of supervised learning, classification tasks take a set of inputs, along with their corresponding outputs (also called labels), with the goal of discovering some underlying map between input and output. Supervised methods often require thousands of training examples, rendering them a non-starter for many researchers and projects. However, there are often creative ways to reduce the effort required. We see an example related to our own in Anastasopoulos & Bertelli (2020), who use existing classifications of agency delegation in EU legislation to perform supervised labeling of other years not covered in the dataset.

Active learning — also called "query learning" in computer science — is a machine learning approach to bolster classification performance by selecting (or "querying") further examples to use in model training. This querying allows an algorithm to obtain higher classification accuracies as training data is added than if new examples were instead randomly chosen. An active learner typically poses its queries over unlabeled data, which are then labeled by a human annotator and subsequently added to the dataset. Active learning is well-motivated in many modern machine learning problems, where unlabeled data may be abundant, but labels are complicated, time-consuming, or expensive to obtain (Settles 2009; Miller *et al.* 2019).

Researchers may find that they need to frequently retrain or update their models, particularly given new information they discover in the course of their exploration. Active learning offers a judicious way to update with new examples. We provide a motivating example in identifying delegation to administrative agencies. The general framework for delegating authority is relatively straightforward: some agency is given a task — often told it "shall" or "must" do something. A standard learner would identify the agency named and the verbs specifying the task and then use that information to identify most instances of delegation. However, there are hundreds of currently active agencies and programs, some of which have unusual names (the "Corporation for National and Community Service" as an example), that would reduce the likelihood of delegation being properly identified.[1]

We encountered this problem early on when discussions of the "Attorney General" were frequently mislabeled because, up until that point, there were no observations in our training data with a cabinet-level secretary referred to as anything but "Secretary." Because the specific words "Attorney General" have few analogous positions in other departments, it would have to be hand-coded explicitly for the model to learn what that is. Generally, this is not a complicated fix: hand label some of these aberrant observations, and it should solve the problem. However, what we could not know *a priori* was what exact issues were going to appear: moving to an interactive labeling and machine-learning framework alleviated these concerns because the classifier would be able to tell us where it was having difficulties discriminating between classes. Learning on the go was the preferred option because, though our classification scheme is simple, there are enough moving parts that it is hard to know precisely what problems would have arisen before we started the coding.[2]

In a typical supervised machine learning environment, a model is trained on a training partition of the data and evaluated on a testing partition. Given the performance, applied researchers may then decide to "deploy" the model for some practical purpose: for instance, the automated labeling of new examples. They may also determine that the model perfor-

[1]Unless that agency also showed up in the training data which, given that there are hundreds of agencies, is likely to miss many.

[2]This is a limitation of grammar-based approaches to similar problems. Since our method adapts and learns as we hand-code more and more uncertain observations, it begins to cover the realm of possible mismatches more broadly. Given the complexity of Congressional language and the continuing evolution of the legislative agenda, we are skeptical of the long-term returns to structure-based approaches seen specifically in Vannoni *et al.* (2021)

mance, as evaluated on the test data, was insufficient, and so they may choose to label more examples for model training. For our case, we argue that active learning should be used when selecting new examples to label, with the criterion for selection being the model's uncertainty in labeling new examples. After querying these uncertainly labeled documents from the model, they may be manually labeled by the researchers and appended to the training dataset. The process of training, evaluation, and querying then repeats until model evaluation suggests the model is robust enough to be deployed to label all remaining examples automatically.

As mentioned above, our criterion for actively querying new data examples is the model's uncertainty in label prediction — "uncertainty sampling." In the case of classification, this strategy amounts to identifying the observations closest to the classification boundary, such as examples with the smallest margin in max-margin models (e.g., SVM) or examples with logits closest to 0.5 for logistic regression. Other common querying strategies are ensemble approaches — wherein queries are made by identifying observations for which there is classification disagreement between multiple models — and expected-model-change approaches, in which examples are selected that would most significantly change the current model (Settles 2009). For our problem, we have implemented uncertainty sampling, as it is the most widely used, computationally straightforward, and flexible for comparison between various methods (Tong & Koller 2001; Settles 2009).

This paper uses a convolutional neural network (CNN) with a multi-layer perceptron (MLP) as our primary classifier. We note that a variety of other supervised learning methods can accommodate active learning, and the choice of model is, in general, the researcher's preference. We use uncertainty sampling, querying those examples with logits closest to 0.5.

Convolutional Neural Networks for Text Classification

We follow the example of Kim (2014) who defined a straightforward CNN for text classification as well as Zhang & Wallace (2015), who provide practical guidance on using such models. The primary advantages of using deep learning for our text analysis — rather than "bags of words" approaches (see those discussed in Grimmer & Stewart 2013) — are to model the similarity between words and to account for word order in text sequences. Both of these advantages involve considering words in context.

Underpinning the use of CNNs for text analysis is the distributed representation of words, wherein each word in a vocabulary is associated with a real-valued feature vector (Bengio *et al.* 2003; Mikolov *et al.* 2013; Pennington *et al.* 2014). These expressive vector representations encode many linguistic regularities and patterns, such as the relationships between synonymous words, and their use has been shown to improve the accuracy of supervised NLP tasks (Turian *et al.* 2010). Unsupervised training of word embeddings is typically accomplished by predicting the incidence of words given local context words, but practitioners may wish to instead randomly initialize word embeddings to be fully learned as parameters of their specific task (Rodriguez & Spirling 2021).

Concurrently with the maturation of distributed representations of words, convolutional neural networks have been shown to leverage word vectors for text classification effectively Zhang & Wallace (2015). The architecture adapted in this paper is that of Kim (2014), with practical guidelines for use outlined by Zhang & Wallace (2015). Prior to the convolutional network, the word tokens of the text to be classified are transformed to their real vector space word embeddings via a lookup table. Let d be the dimension of the word embeddings, and let $\mathbf{w}_i \in \mathbb{R}^d$ be the word embedding of the i-th word in the text. Each bill section is padded to be the same length n, and a single section is then represented as the concatenation (stacking) of the word embeddings that comprise it. We will denote the document (i.e., bill section) matrix as $\mathbf{W} \in \mathbb{R}^{n \times d}$ - the stacking of the word embeddings - with $\mathbf{W}_{i:j}$ denoting the sub-matrix of $\mathbf{W}$ from row i to row j.

A convolutional filter is parameterized by a matrix $\mathbf{q} \in \mathbb{R}^{h \times d}$, where h is the region size of the filter. For text applications, h indicates how many words the filter operates upon at once (e.g., $h = 1$ corresponds to a single word, $h = 2$ to a bigram, and $h = n$ to an n-gram).

Convolution is performed by applying the filter $\mathbf{q}$ to a window of words $\mathbf{W}_{i:i+h-1}$, which is accomplished via the summation of element-wise multiplication of the matrices, which shall be denoted by the $\cdot$ operator. The feature, c_i extracted by the operation is obtained by adding a bias term $b \in \mathbb{R}$ and applying a non-linear activation function f:

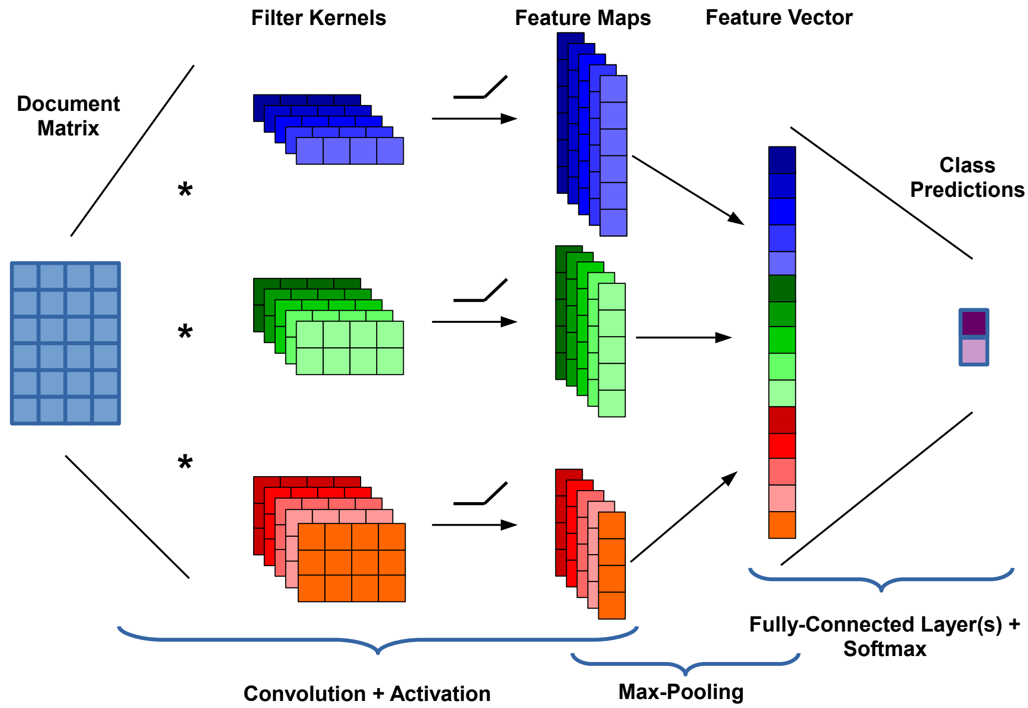
$$c_i = f(\mathbf{q} \cdot \mathbf{W}_{i:i+h-1} + b), \tag{1}$$

For our purposes, the non-linear activation function chosen is the *rectified linear unit* (ReLU - Glorot *et al.* (2011)), which is defined as: $f(x) = \max(0, x)$. The filter is applied to each possible contiguous window of h words in the matrix $\mathbf{W}$ to produce a feature map $\mathbf{c}$, a vector of features extracted by the operation described in equation (1). To increase the model's ability to capture relevant information from the text, multiple filters of the same size are used, with the idea being that they will extract complementary features from the same regions of text. Additionally, multiple filter sizes may be used within the same model.

Given our use of filters of different sizes and the varying lengths of text, a pooling scheme is used over the acquired feature maps to assemble a fixed-length feature vector for the text. Following Zhang & Wallace (2015), we choose 1-max-pooling, in which only the maximum activation from each feature map is retained and all such scalar values are concatenated to form the final feature vector. This strategy adheres to the intuition of choosing the "most important" feature from each map, and is furthermore computationally efficient (Zhang & Wallace 2015).

The above describes the process whereby a fixed-dimensional feature vector $\mathbf{c} \in \mathbb{R}^F$ may be obtained from text using a convolutional neural network, where F is the total number of convolutional filters in the network. This vector may be used for many tasks that can leverage a compressed representation of the text. For text classification, we use a *multi-layer perceptron* (MLP) to map the text feature vector $\mathbf{c}$ to a vector of scores for each possible class $\mathbf{s} \in \mathbb{R}^C$, where C is the number of possible categories. Our MLP comprises two fully-connected layers with ReLU nonlinearities. Each fully-connected layer applies a

Figure 1: Illustration of CNN for text classification



For this example, $d = 4$, there are 5 filters of lengths {1,2,3}, and there are 2 classes.

weight matrix $\mathbf{M}_{FC}$ to its input and adds a bias b_{FC} before applying the nonlinear activation function:

$$\mathbf{y} = f(\mathbf{M}_{FC}\ \mathbf{c} + b_{FC}) \tag{2}$$

Finally, we use the softmax function to compute the model probabilities for each class from the preceding scores. The model is trained by minimizing the binary cross-entropy loss between model predictions and true class labels with respect to authority delegation. Figure 1 depicts an illustration of the described architecture for binary classification of text. The architecture can accommodate arbitrarily many classes. [3]

[3]For the results presented in this paper, we use the CNN architecture described here with 64 filters each of sizes $\{1, 2, 3, 4, 5\}$. We choose a word embedding dimension of $d = 300$, and our dataset yields a vocabulary size of 5775 words. Our first MLP hidden layer has 64 neurons, while the second has 32. To train our model, we optimize the parameters using the Adam optimizer (Kingma & Ba 2014) with a learning rate of 0.0001

Data

We utilize data for all versions of all bills (both successful and unsuccessful) from the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congresses. We separate each version of each bill into titles and analyze them at the bill section level. We do this for three reasons: first, because each title in each bill deals with a particular agency or activity and likely contains an entire delegatory phrase, keeping the task more straightforward. Second, any given bill could delegate authority to multiple agencies in multiple titles, so to avoid missing any additional delegations, we wanted to reduce it to units that are about a single delegation. The final reason we chose bill titles is that bill titles are the smallest comparable distinct units of a bill: titles are more comparable to one another than either a sentence or an entire bill would be.[4]

Our complete dataset has several components. The division of our dataset is into labeled and unlabeled sub-datasets. The labeled component comprises 2098 bill sections from the 110$^{\text{th}}$ that were read by human annotators and assigned a binary label with respect to delegation; the unlabeled component contains the remaining 137,616 bill sections from the 110$^{\text{th}}$ Congress as well as all sections from the 111$^{\text{th}}$ Congress. As is standard practice within machine learning, we divided our labeled data into subsets for training, validation, and test. In performing this splitting, we ensured that all titles/versions from the same bill were apportioned to the same subset. We randomly divided the bills in our labeled data into training, validation, and test using proportions of 65%, 15%, and 20%, respectively. The remaining bills from the 110$^{\text{th}}$ Congress were also apportioned into training, validation, and test using the same proportions. The numbers of bills (by bill number) and examples (e.g.,

and batch size of 64 bill sections. The loss to minimize is the binary cross-entropy loss, with bill sections labeled according to whether they delegate authority or not. We allow the model to train for 13 epochs and regularize the model using Dropout (Srivastava *et al.* 2014) in the MLP with a drop-rate of 15%.

[4]While it might seem advantageous to go even more fine-grained than the bill section, there are fundamental limitations drawn from how bills are written. For example, the most common unit of analysis in natural language processing tasks is the sentence. The construction of any given sentence within a bill is, however, is contingent primarily on the remaining information presented at the title level. Furthermore, in titles where authority is delegated to an agency, sentences may be less precise than the section, which provides much-needed context. Studying bills at the title level will ultimately let us make better inferences about the agencies or programs to which Congress has delegated.

bill section) for each subset are presented in the Appendix.

Delegation Coding

Essential to our project is a consistent definition of delegation to administrative agencies. An act of delegation is a mandate or permission for a federal agency or program (including the president) to exercise public authority in some way (see McCubbins *et al.* 1987; Kiewiet & McCubbins 1991; Huber & Shipan 2002; Gailmard & Patty 2012a, for a discussion of this point). For our task, we stated that allocating money for federal agencies to spend, instructing agencies to promulgate rules, granting agencies the ability to exempt themselves from preexisting rules, requiring agencies to compile reports or commission pilot studies, and charging agencies with the enforcement of specific policies are all delegation (Kiewiet & McCubbins 1991).

For the hand-coding, we gave straightforward instructions as to how we identify delegation. First, is Congress acting upon an administrative agency? This will include all references to both the agency itself and the person in charge of that agency. We operated with a list of administrative agencies and matched each instance of delegation to one of those agencies.[5]

Second, what is the title asking the agency to do? In general, Congress delegates authority by asking an agency to perform a specific task, collect information, write new regulations, hire people, write a report to Congress on their activities, delegate authority to sub-agencies or outside of government, and make or distribute an award, among many other things. A bill is not delegating authority if it only appropriates money, if they are referencing actions already taken, or if Congress is writing new rules or regulations. Keeping these actions separate allows us to track statutorily derived authority for the agencies, not merely what funds they have been allotted.

[5] Most often, if Congress is referring to a governmental entity (except for organizations already within Congress, which they make apparent), it is an administrative agency. We make an exception for delegating to the courts or states and local governments because those tasks are defined differently.

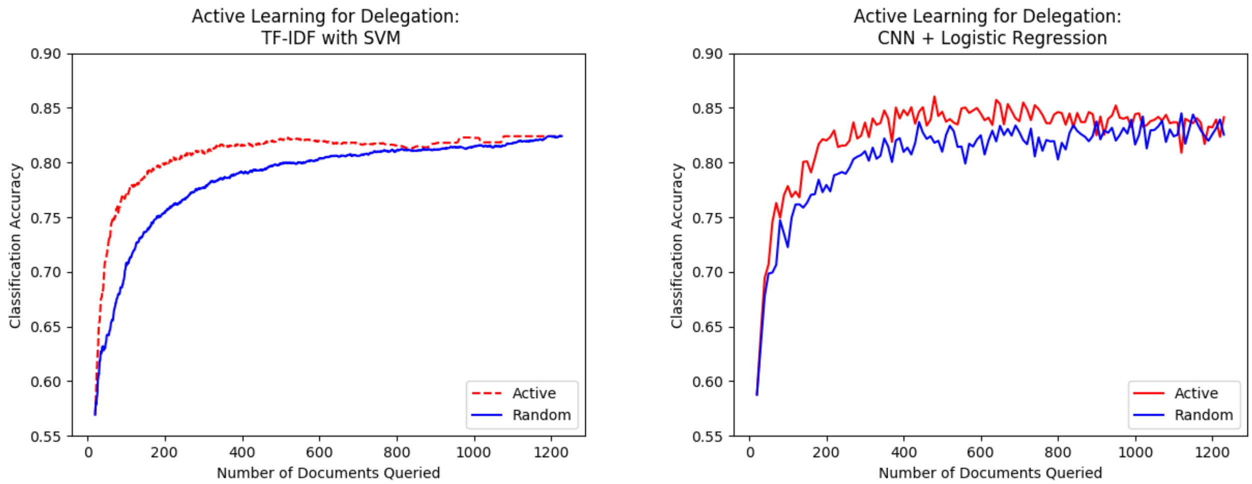
Below are example bill titles that the active learner selected in early runs as uncertain and how they were coded.

- `Section.2402. energy conservation projects. using amounts appropriated pursuant to the authorization of appropriations in section.2403.a.6, the secretary of defense may carry out energy conservation projects under chapter 173 of title 10, united states code, in the amount of 800000` (delegates authority to an administrative agency)

- `Section.2. reemployment of foreign service annuitants... the authority of the secretary to waive the application of subsections a through d for an annuitant pursuant to subparagraph c of paragraph 1 shall terminate on September 30, 2008. the authority of the secretary to waive the application of subsections a through d for an annuitant pursuant to subparagraph c ii of paragraph 1 shall terminate on September 30, 2009` (does NOT delegate authority to an administrative agency)

In the above examples, it is clear why the algorithm would have selected them as ambiguous classifications and why human readers would classify them correctly. Take the top section, dealing with "Energy Conservation Projects." A quick read of the section makes it clear that Congress is delegating authority to Defense (through the Secretary of Defense) to spend $800,000 on energy conservation projects. The classifier may have had issues with the added verbiage of the task (may carry out) and the addition of US code language in between. In the second section, where Congress is setting up "foreign service annuitants", it is clear that the agency is not being given an extra task or authority but only describes how applications must be processed (and when they terminate). This section is an example where the language (context-free) would indicate the possibility of delegating authority, but the additional context (plus a close reading) makes it clear that this is not occurring. These are only two examples pulled from an early run of the active learning module, set up to illustrate the nature of the classification task.

Figure 2: Active learning querying versus random sampling for SVM and CNN models.



Note that the active learning models generally outperforms the random sampled models and that the CNN generally outperforms the SVM.

Model Performance

We propose two aspects regarding our convolutional neural network and active learning model performance. First, we suspect the CNN will outperform a traditional word-document-matrix vector-space method of text classification. For comparison, we use baselines of a term-frequency, inverse-document-frequency (tf-idf) text representation with linear support vector machine (SVM), L1-penalized logistic regression (LASSO), and random forest classifiers.[6] Second, we expect that incorporating active learning into the classifiers will outperform random sampling as additional documents are appended to the training data.

To evaluate active learning performance, we employed a standard demonstration. Generally, our scheme involves artificially restricting the training set to a small number of examples and iteratively evaluating performance on the validation set while augmenting the training set. When adding to the training set, we use either active learning or random sampling for selecting new examples before retraining the model and then re-evaluating per-

[6]See Appendix Section 3 for a formal description of the baseline models and Miller *et al.* (2019) for work using tf-idf with active learning.

formance. As the training set grows, we expect validation performance to increase, but more rapidly for the active learner, as it has judiciously queried examples to strengthen performance. We begin by randomly selecting ten documents from our training set. We then query ten additional documents from training, either actively or randomly, and retrain the model before evaluating again. This process continues until the whole of the training set has been queried.

Figure 2 shows this active learning demonstration for our neural model and for the SVM baseline. In both cases, the active learner outperforms the random sampler in that classification accuracy increases more dramatically as training documents are added. The trend is particularly apparent for training sizes of 20-800 examples for the SVM and 200-600 examples for the CNN. As expected, when the entire training set is used, performance converges between active and random sampling. We note that the CNN's active/random sampling curves exhibit significantly more variability than for the SVM due to the inherent stochasticity in neural architectures, which are more subject to the randomness of parameter initializations. To mitigate this and the randomness of sampling training examples, we obtained these active learning curves by averaging the results over five trials.

Also apparent in Figure 2 is that the CNN outperforms the SVM on classification performance. In fact, the CNN outperforms all three of our tf-idf baselines: SVM, LASSO, and Random Forest. In Table 1, we present the classification accuracies on both the validation and test dataset splits for all baseline classifiers and the neural model. We additionally used the fully-trained CNN to actively query an additional 200 examples from the unlabeled training partition. After incorporating those new examples into our dataset and retraining, we obtain the "Post-Query" accuracies in Table 1, where modest improvement is evident.

Table 1: Delegation Classification Accuracy

Features	Classifier	Pre-Query Val.	Pre-Query Test	Post-Query Val.	Post-Query Test
TF-IDF	SVM	82.4	87.1	82.6	86.3
	Random Forest	81.2	87.3	82.4	88.4
	LASSO	83.9	87.9	85.9	87.1
CNN	MLP	**86.5**	**90.2**	**87.6**	**90.4**

Classification accuracy for tf-idf baseline models and our neural network – convolutional neural network with a multilayer perceptron (CNN and MLP). We provide accuracies on both the validation and test splits, as well as before and after the addition of new documents queried by the active learner.

Classification and Delegation Results

With delegation predictions for each version of each bill section from the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congress, the remainder of the paper will examine the consequences of these delegation classifications. First, we examine how closely our classification of delegatory sections concurs with existing measures of discretion (in particular, we validate the textual measure popularized by Huber & Shipan (2002)). Then, we use our delegation data to empirically explore the ally principle. Here, we are particularly interested in testing whether congressional delegation decisions are more responsive to stable agency characteristics—like ideology and structural independence—or to shifts from divided to unified government. Our results suggest that congressional decisions about delegation are shaped by the stable contours of the administrative state rather than the top-level political leadership that changes with presidential administrations.

What does Delegation Look Like?

First, we compare our measure of delegation to the most widely used proxy–bill length–which was first introduced in Huber & Shipan (2002). The logic of the measure is straightforward; Huber & Shipan (2002) claim that longer bills are more likely to require specific returns from agencies, and this generally entails more fine-grained oversight and control. Shorter bills, ones that discuss delegation and discretion more broadly, are less likely to have the same requirements and instead reflect less deference from the legislature. It has been accepted by the literature (see Clinton *et al.* (2012) as an example) that longer bills mean more delegation.

If Huber & Shipan (2002)'s measure of delegation is accurate, we would expect the length of a given section to predict delegation, which we can compare directly to the predictions from our classification model. We can also examine discretion for the aggregate of a bill over all of its sections using the delegation ratio (Epstein & O'Halloran 1999; Anastasopoulos & Bertelli 2020). Another widely used general measure of discretion, the delegation ratio, is defined as the total number of sections delegating authority divided by the total number of sections; essentially, how much of the bill delegates authority. This penalizes omnibus legislating and other forms of massive legislation that attempt to do many things at once since raw counts would overweight these bills.

Table 2: Comparing Total Words to Delegation Measures for 110$^{\text{th}}$ Congress

	Delegation		**Delegation ratio**	
	Logit	**Logit w/ Mixed-Effects**	**Beta**	**Beta w/ Mixed-Effects**
	Model 1	**Model 2**	**Model 3**	**Model 4**
Words/1000	0.829*	0.942*	−0.002*	−0.001*
	(0.008)	(0.006)	(0.0003)	(0.0003)
Constant	−1.087*	−1.584*	0.002	0.132*
	(0.007)	(0.009)	(0.010)	(0.021)
N	139714	139714	8847	8847
Log Likelihood	−82265.290	−77319.430	754.258	876.266

$^{*} = p < .05$

Table 2 compares delegation measures for the 110$^{\text{th}}$ Congress. First, we observe that the number of words corresponds with the likelihood a given bill section delegates authority (Models 1 & 2). However, total discretion in a bill — as measured by the delegation ratio — is not connected with the total number of words (Models 3 & 4). Importantly, the delegation ratio is negatively associated with total bill length, a relationship that directly challenges the efficacy of bill length as a proxy for discretion.[7] The relationship between the length of bills is

[7] It is important to note that the measure developed by Huber & Shipan (2002) is meant to be a proxy

only weakly associated with how much agency discretion there is and that indirect measures of agency discretion may have been confounding, even though this measure is pretty widely used (see Huber & Shipan 2002; Clinton *et al.* 2012, as examples though there are many more as well). This demonstrates the value of using a text-based classification scheme of delegation section-by-section.

We see this further illustrated in Table 3, which shows the confusion matrix for Model 1 from Table 2 (which used only bill section length) against our model predictions. Overall, predictive accuracy is middling, around 71.2%, which is lower than our test set predictive rate of 90% from Table 1. Considering the dramatic difference in model sophistication, the lack of agreement between the models is unsurprising.

Table 3: Confusion Matrix for Word Length Measure to Machine Labeled Sections

	Label from Deep Learning:	
Label from Section Length	Does Not Delegate	Delegates
Does Not Delegate	84194	36312
Delegates	4892	14316

We see that relying only on word count, a logit model predicts non-delegatory sections in agreement with our model at 94.1%. However, for sections that delegate authority, the logit model performs significantly worse: 14,316/50,628 correct, only 28.3% agreement with our predictions. In terms of predictive modeling, knowing the total number of words helps eliminate the non-delegatory sections, but is *worse than random* for longer bill sections. This is consistent with total word length for a section being a poor proxy for overall delegation and deference to an agency. Similarly, if we use median or mean words selection criteria (no model, all sections above the mean/median are coded as delegatory), we find that predictive accuracy stays about the same, at 72.1% and 69.4% accuracy, respectively. This suggests

for the extent to which an executive agency is constrained in carrying out delegated authority, whereas our method is designed to identify instances of delegatory language. Therefore, the delegation ratio we create using our method does not allow us to speak definitively to questions of constraints on delegated authority. However, our finding that bill length is negatively associated with delegation ratio is problematic for Huber & Shipan's (2002) measure because it suggests that much of the text in long bills, rather than being devoted to constraining agency authority, may not be related to delegation at all.

that using a word count/volume proxy for delegation or discretion provides only a weak signal: the performance of the logistic regression suggests even more strongly that these designations are almost certainly missing entire types of delegatory action, in most likely systematic ways.

Combining this with the discussion of the delegation ratio models from Table 2, we see that though bill length provides some information consistent with discretion, there are many issues with using it directly. Longer bills almost certainly delegate more frequently than shorter bills but is that a function of the type of discretion they are providing, or is it a tautology that *longer bills tend to do more*? With the rise of omnibus bills and the pervasiveness of hitchhiker legislating (see Krutz 2001; Casas *et al.* 2020, for more), simply identifying that these longer bills delegate more is insufficient.[8]

Theoretical Expectations for Variation in Delegation

Having described our method and compared our measure of delegation to a prominent existing measure, we move on to using our measure to test an important theoretical proposition from the literature. Perhaps the most prominent finding from the extensive literature on delegation is the ally principle, which posits that interbranch ideological conflict drives variation in legislative delegation decisions. Specifically, the ally principle yields a directional hypothesis that the scope and amount of delegation decreases as the ideological distance between the legislature and the recipient of delegated authority—usually an executive agency—increases.

The ally principle rests on a straightforward assumption that legislative principals are averse to losing policy utility at the hands of unfaithful agents. While some literature has articulated compelling theoretical reasons the ally principle may be violated in practice (Gailmard & Patty 2012b; Miller & Whitford 2016), the body of empirical literature confirms that, on average, the principle holds (Epstein & O'Halloran 1999; Huber & Shipan 2002;

[8]Each of the following results in this paper are presented with bill length as a robustness check in Appendix Section 8, with only one result differing.

Clinton *et al.* 2012; Farhang & Yaver 2016). Our new measure of delegation allows us to add to this literature in two specific ways that both shed light on previously unappreciated facets of the ally principle.

First, our methodology enables us to identify delegatory language in large datasets of legislation spanning multiple Congresses. This allows us to expand our focus beyond a small set of laws to the universe of all legislation within a certain time period. This broader scope gives us the ability to test the generalizability of the ally principle across different legislative contexts. For example, we are able to determine whether findings supporting the ally principle are artifacts of a narrow focus on a small subset of important laws, or if these trends are more broadly applicable. There is a growing body of literature that shows that Congress treats "significant" legislation differently than ordinary legislation, and most of the work on the ally principle has focused on high-profile or otherwise seemingly important legislation (Epstein & O'Halloran 1994; Farhang & Yaver 2016; Shaffer 2020).

Secondly, the dataset we create—which identifies every instance of delegation, as well as the recipient of that delegation, in all bills introduced in the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congresses—allows us to gain leverage on fundamental questions about how legislators view the executive branch. Any test of the ally principle requires ideological measures for both the principal and the agent—and a basic prerequisite for this measurement strategy is to be able to specifically identify an agent in each bill. However, this identification is not necessarily a straightforward task. If the president is able to set the ideal point of each executive agency—an assumption common in much of the delegation literature—a simple indicator for divided government may be sufficient for testing the ally principle. If, on the other hand, agency ideology is relatively stable across presidential administrations but varies substantially across agencies, agency-specific ideology measures may be necessary.

Our data uniquely equip us to differentiate between two potential legislative views of the executive branch—one as a presidentially-controlled unitary institution, and the other as a complex and internally-differentiated conglomeration. Because our data span the 110$^{\text{th}}$ and

111$^{\text{th}}$ Congresses, we are able to observe variation in regime type (from divided to unified government) while keeping the partisanship of the congressional majority constant. Using agency ideology and independence measures developed by other scholars, we are also able to measure the cross-agency variation that exists within the executive branch at any given point in time.Importantly, our measurement strategy here varies from Clinton *et al.* (2012), who assign a primary recipient of delegated authority for each bill. As described in more detail below, we create bill-level measures that make use of information on each agency receiving delegated authority – an advantage that we have because of our delegation data.

Models

Our intent in this section is to construct a set of empirical tests to adjudicate between two theoretical frameworks of interbranch relations. The first framework, which is prominent in much of the literature on delegation, views the executive branch as a unified institution under the direction of the president (Epstein & O'Halloran 1996, 1999; Huber & Shipan 2002; Moe 2012; Farhang & Yaver 2016).[9] Following this framework, the ally principle holds that delegating activity will increase during unified government and decrease during divided government—since the majority party in Congress will be more willing to delegate to an executive branch controlled by a copartisan president.

The second framework views the executive branch as a pluralistic and internally-complex institution (McCubbins *et al.* 1987; Kiewiet & McCubbins 1991; McNollgast 1994; Volden 2002; McGrath 2013; Clinton *et al.* 2014; Lowande 2018a; Rudalevige 2021). Following this framework, the partisanship of the president—and the unified/divided government dichotomy—should be much less important in explaining variation in delegation activity compared to agency characteristics that remain relatively stable across presidential administra-

[9]For example, in their formal model, Epstein & O'Halloran (1999, pg. 60) allow the president to set the ideal point of the agency—and demonstrate that, in equilibrium, presidents set agency ideal points equal to their own. While they do later analyze how varying degrees of agency insulation from presidential control affect delegation decisions (Epstein & O'Halloran 1999, pgs. 151-161), they expect the effects of agency characteristics on delegation to be conditional on unified/divided government.

tions. We examine two such agency characteristics in the models below—agency ideology and agency independence.

Agency ideology refers to the aggregated policy preferences of each executive branch agency. We use a measure developed by Clinton *et al.* (2012), who administered a survey to agency appointees and careerists asking them to take positions on salient congressional roll call votes. This methodology yields agency-level ideological scores that are on roughly the same scale as the ideological scores for members of Congress.

Agency independence generally refers to the extent to which executive branch agencies are insulated from political pressure exerted by Congress or the president. To operationalize this concept, we use measures developed by Selin (2015), who estimated agency ideology on two dimensions—one tapping limits on the appointment of key agency officials and the other tapping the extent to which agency decisions are subject to political review.

The structure of our data—which are at the bill level and include variables measuring the number of delegations of authority to each agency—necessitates some transformation of these agency-level ideology and independence measures. For the ideology measure and both independence measures, we create bill-level weighted averages using information on all agencies receiving delegated authority in each bill. Each agency has an ideology score from Clinton & Lewis (2008); Clinton *et al.* (2012) and two independence scores from Selin (2015), and for each bill, these scores are multiplied by the number of individual delegations of authority to that agency in that bill. The resulting products, which are ideology and independence scores weighted by the prevalence of each agency in each bill, are summed across all agencies and then divided by the total number of delegations in each bill to create bill-level agency ideology and agency independence "profiles." The following equation is applied to each bill to arrive at these profile scores:

$$ideo_j = \frac{\sum_{i=1}^{n} \theta_i d_{ij}}{\sum_{i=1}^{n} d_{ij}}$$

$$ind_j = \frac{\sum_{i=1}^{n} a_i d_{ij}}{\sum_{i=1}^{n} d_{ij}},$$

in which the subscript j indexes bills, the subscript i indexes agencies, θ_i is an agency-specific ideal point, and d_{ij} is an agency-bill-specific count of delegations. For independence, we use the same formulation, but with agency specific independence scores (a_i) rather than agency specific ideal points θ_i.

To test the viability of an ally-principle conceptualization of congressional-agency relations and compare it to one that prioritizes the permanence of the administrative state, we fit a series of regressions. For our dependent variable, we use a bill's delegation ratio as defined by Epstein & O'Halloran (1999). The delegation ratio Δ_j for bill i represents the number of bill sections delegating authority d_j to administrative agencies divided by the total number of sections in the bill p_j:

$$\Delta_j = \frac{\sum_{j=1}^{n} d_j}{\sum_{j=1}^{n} p_j}$$

We fit a series of regressions predicting a bill's delegation ratio Δ_j based on the distance between bill-level ideology and the relative majority party medians, two dimensions of agency independence, and a dummy variable for unified government.[10] Because the delegation ratio is constrained between zero and one, we use a beta regression to fit the model.[11]

Results

The models presented in Table 4 below are fit on different subsets of data—a subset of bills that passed the House, a subset of bills that passed the Senate, a subset of bills that were enacted into law, and finally a subset of important bills by Mayhew's classification (Mayhew

[10] Some models also include fixed effects for policy area for bill version type. More detail on these measures is found in the subsequent section.

[11] Similar results are found if we use OLS as our underlying model. See Appendix XXX for details.

1991).[12] Following the logic of Cox & McCubbins (2005), we assume all bills that pass a legislative chamber have the support of the majority party in that chamber, which for our time period (110th and 111th Congresses) is the Democratic party in both the House and the Senate. Therefore, the coefficients on the agency ideological profile variable can be interpreted as the expected change in delegation to an agency as that agency becomes increasingly ideologically distant from the Democratic congressional majority.[13] Similarly, the coefficients on both agency independence profile variables capture the expected effect of increasing agency independence on delegation.

The dichotomous indicator for unified government straightforwardly tests the main expectation of the ally principle—that delegation activity in a separation of powers system will systematically vary with the governing regime. In our specific case, coefficients on this variable can be interpreted as the change in delegating activity from a Democratic Congress attributable to the change in the presidency from George W. Bush to Barack Obama, the same change exploited by Lowande (2018b) to test Congressional relations with the changing executive branch.

Because delegation has been shown to vary across issue area, all models presented below, with the exception of the model fit on the Mayhew bills subset in Table 4, include fixed effects for the Policy Agendas Project topic area of each bill, as coded by the the Congressional Bills Project (Adler & Wilkerson 2011). Because the unit of analysis for the data we use is bill-version, we also include a fixed effect for bill version (introduced, reported from committee, passed House, etc.) to account for any systematic variation in delegatory scope that may occur as a bills advance through the legislative process. These fixed effects are included in all

[12]The models are fit on the subset of bills that delegate authority to actors in the executive branch—bills that do not include any delegation are excluded. Therefore, what is being analyzed is variation in the *amount* of delegation in each bill (measured by the delegation ratio), as opposed to the initial decision of whether or not to delegate.

[13]The ideology variables are constructed as the absolute difference between the ideological score of the median Democrat in the relevant chamber and the bill-level agency ideological profile score (which is the weighted average of the ideological score of each agency delegated to in each bill). So, for our models fit on bills that passed the House, the score is constructed using the median House Democrat in each Congress (and vice versa for the Senate). For the models fit on bills that became law, we take the median ideological score of House and Senate Democrats pooled together by Congress.

models below, with the exception of models that focus exclusively on enacted bills—because these models are only fit on data containing the enacted versions of bills that became law.

The weight of the evidence presented in Table 4 suggests that individual agency characteristics—specifically agency ideology and independence—are more important in explaining variation in delegation activity than the existence of unified or divided government. In the model fit on bills that passed the Senate—the only model in which the coefficient on the unified government indicator is statistically significant—the coefficient is in the positive direction predicted by the ally principle. However, all other models have a null result for unified government.

The findings on agency ideology and independence are consistent across all models besides the one fit on the subset of bills classified as important by the Mayhew coding system (Mayhew 1991), which is under-powered given the small number of observations. The overall trend demonstrates that the Democratic majorities of the 110$^{\text{th}}$ and 111$^{\text{th}}$ Congresses wrote less delegation into bills that delegated to more conservative and more independent agencies, regardless of whether these bills were passed during unified or divided government. We do not interpret these findings as dispositive with respect to the ally principle, but they do suggest a different framing of how congressional majorities identify allies in the executive branch. The results presented in Table 4 suggest that, when making decisions about the scope and extent of delegation, congressional majorities are more attuned to the relatively stable traits of agency ideology and independence than they are to changes in political leadership in the White House and at the top levels of agency hierarchies.[14]

The models presented in Table 5 test our expectation that the effect of a change in presidential administration on congressional delegation activities should be conditional on the level of agency independence. Our expectation is that delegation to agencies that are

[14] It is worth noting here that the agency-level ideology and independence measures we use to construct our variables are time-invariant, which assumes that these constructs are not meaningfully affected by changes between unified and divided government. This would be more problematic if we studied a longer period of time, as these values are likely to change gradually, but this is not a concern given the more circumscribed time-frame we cover.

Table 4: Estimating Delegation Ratio as a Function of Regime Type and Agency Independence and Ideology

	Dependent variable:			
	Delegation Ratio			
	Passed House	Passed Senate	Became Law	Mayhew
Unified Government	-0.027	0.063^{*}	-0.074	-0.308
	(0.030)	(0.038)	(0.080)	(0.332)
$\lvert$Agency Ideo. $-$ H.Maj. Median$\rvert$	-0.173^{***}			
	(0.035)			
$\lvert$Agency Ideo. $-$ S.Maj. Median$\rvert$		-0.246^{***}		
		(0.045)		
$\lvert$Agency Ideo. $-$ C.Maj. Median$\rvert$			-0.439^{***}	-0.545^{*}
			(0.101)	(0.320)
Agency Independence Profile D1	-0.318^{***}	-0.184^{***}	-0.224^{***}	-0.973^{*}
	(0.030)	(0.040)	(0.081)	(0.555)
Agency Independence Profile D2	-0.187^{***}	-0.411^{***}	-0.338^{***}	0.257
	(0.032)	(0.047)	(0.095)	(0.496)
Constant	-0.296	-0.195	-0.450	-0.352
	(0.493)	(0.461)	(0.287)	(0.372)
Observations	4,636	2,662	488	24
Major Topic FEs	✓	✓	✓	
Bill Version FEs	✓	✓		
R^2	0.489	0.537	0.580	0.403
Log Likelihood	2,614.208	2,397.127	585.709	11.649

*p<0.1; **p<0.05; ***p<0.01

Note: The three ideological distance variables in the table are bill-level variables measuring the absolute difference of the DW-NOMINATE score of the median majority party member in each chamber (or in both chambers combined, as is the case for the model fit on bills that became law) and the weighted average of the Clinton *et al.* (2012) agency ideology scores for agencies receiving delegated authority in each bill.

designed by statute to be particularly susceptible to political influence (so, agencies that would score low on Selin's independence measures) will be more affected by changes in presidential administration. Specifically, we include an interaction between the second dimension of the independence measure—which taps the degree of independence from policy review by political principals—with the unified government indicator. Here, we are using the independence measure as a proxy for the responsiveness of agency policymaking to the president and politically-appointed leadership, and seeking to test whether the effect of that political responsiveness on delegation is conditional on unified or divided government.

Table 5: Interaction Models

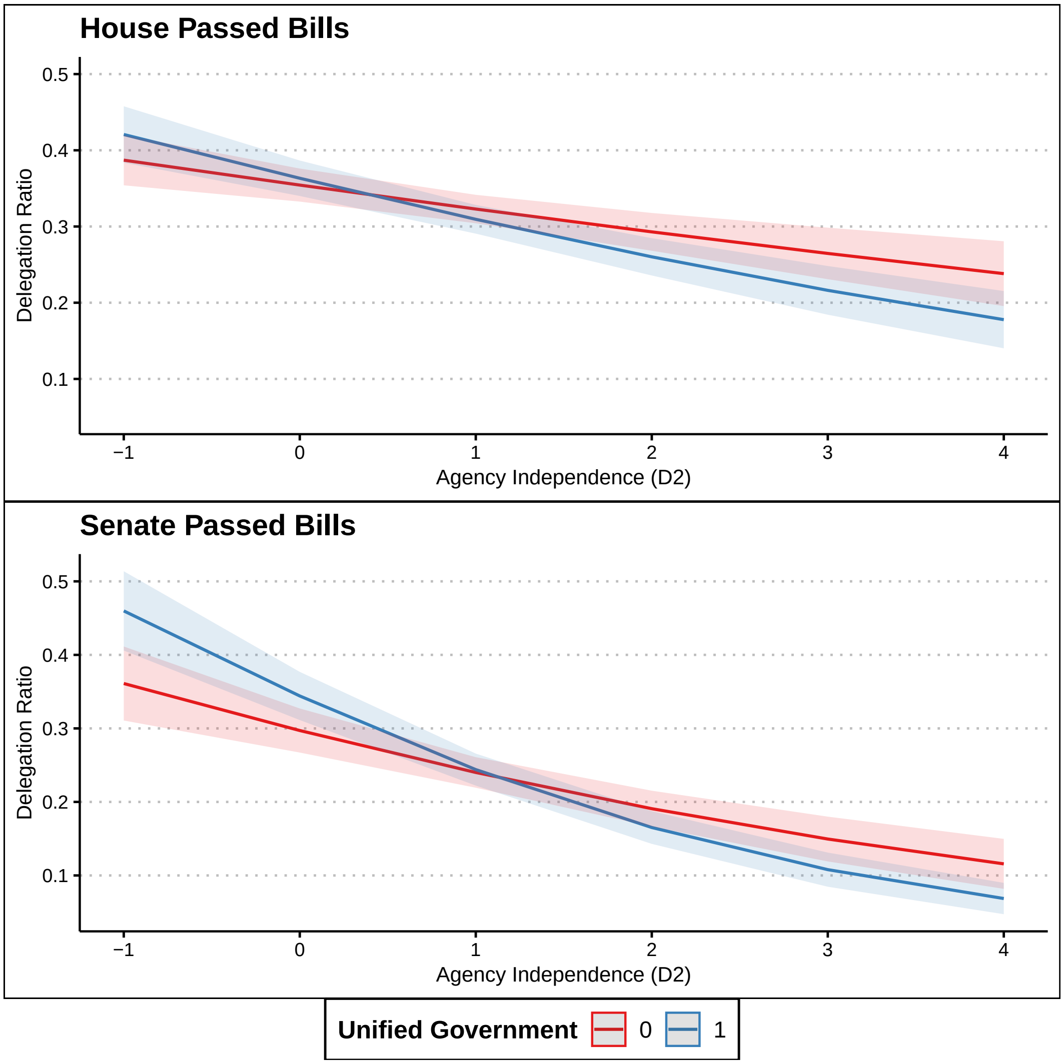
	Dependent variable:		
	Delegation Ratio		
	Passed House	Passed Senate	Became Law
Unified Government	0.040	0.224^{***}	0.025
	(0.037)	(0.052)	(0.111)
|Agency Ideo. − H.Maj. Median|	-0.179^{***}		
	(0.035)		
|Agency Ideo. − S.Maj. Median|		-0.253^{***}	
		(0.045)	
|Agency Ideo. − C.Maj. Median|			-0.446^{***}
			(0.101)
Agency Independence Profile D1	-0.317^{***}	-0.195^{***}	-0.236^{***}
	(0.030)	(0.040)	(0.081)
Agency Independence Profile D2	-0.144^{***}	-0.301^{***}	-0.268^{**}
	(0.035)	(0.052)	(0.106)
Unified Gov. X Agency Ind. Profile D2	-0.104^{***}	-0.202^{***}	−0.123
	(0.036)	(0.045)	(0.095)
Constant	−0.338	−0.340	−0.467
	(0.493)	(0.461)	(0.287)
Observations	4,636	2,662	488
Major Topic FEs	✓	✓	✓
Bill Version FEs	✓	✓	
R^2	0.491	0.544	0.584
Log Likelihood	2,618.388	2,406.569	586.501

*p<0.1; **p<0.05; ***p<0.01

Note: The three ideological distance variables in the table are bill-level variables measuring the absolute difference of the DW-NOMINATE score of the median majority party member in each chamber (or in both chambers combined, as is the case for the model fit on bills that became law) and the weighted average of the Clinton *et al.* (2012) agency ideology scores for agencies receiving delegated authority in each bill.

The coefficient on this interaction is negative and statistically significant in two out of three models. We explore this finding more directly in Figure 3, which plots the marginal mean effects from Models 1 and 2 respectively from Table 5. In both, the X-axis is the 2^{nd} dimension from Selin's independence measure, with larger values corresponding to more

Figure 3: Conditional Effect of Agency Independence from Political Review on Delegation



independent agenencies, and the Y-axis is the estimated marginal Delegation Ratio for the bill. We differentiate between unified and divided government. Consistent with Table 5, we find greater total delegation to politically responsive agencies (on the left hand side of the figures) during unified government, and relatively greater delegation to politically independent agencies during times of divided government, though estimated delegations are less for both regime types as agencies become more independent. This effect is more pronounced in the Senate figure, with a steeper slope and a lower ending points for both regime types.

The interaction here is consistent with the observations from Vannoni *et al.* (2021) (for the states) and shows that Congress is more willing to delegate more to politicized agencies during unified government than during divided government. This also fits with findings by Epstein & O'Halloran (1999) and Volden (2002) that Congress gives independent agencies more discretion during divided government. Overall, our findings demonstrate that Congress conditionally applies the ally principle in regards to agency independence, but is consistent in preferring ideologically proximate agencies.

Discussion

In this paper, we have demonstrated how an active learning convolutional neural network for classifying text can be used to study a complex problem in political science: agency delegation from Congress. We hope that the methods we have proposed are clear and usable for other applications and that gains in classification accuracy, while reducing the need for extra documents to be hand-labeled, helps researchers tackle challenging classification problems. This information is part of an ongoing endeavor to learn how Congress uses statutory language to enact its agenda and how the modern legislative process provides oversight and guidance to implement policies. These classifications will help provide the means through which we can test our theories of delegation and congressional oversight on a larger scale basis and provide the nexus for increased research into the implications of statutory language.

After classifying bill sections, we compared our measure of delegation with a well-known proxy. We find that bill length—used by Huber & Shipan (2002) and others to measure constraints on agency discretion—is actually negatively associated with the proportion of a bill devoted to delegation. This finding calls for caution in the use of bill length as a proxy for the amount of discretion delegated to executive agencies, and demonstrates the value of using a text-based classification scheme of delegation section-by-section.

We also use our measure of delegation to gain new leverage on an important theoretical proposition and empirical finding throughout the delegation literature—the ally principle. Because our convolutional neural network classification model allows us to measure delegation in a much larger and more inclusive set of legislation than has been used in previous work, we are able to move beyond a focus on landmark legislation and speak to the generalizability of previous findings (Epstein & O'Halloran 1999, pg. 89). Additionally, our use of agency-specific independence and ideology measures (Selin 2015; Clinton *et al.* 2012), along with our focus on a time period that saw a transition from divided to unified government, allows us to better isolate the effect of presidential administration on congressional delegation decisions. We find little support for the ally principle as it is traditionally operationalized—with divided government associated with a decrease in delegation and unified government associated with an increase. Instead, we find that the majority party in Congress typically prefers to delegate more to ideologically-proximate and politically-responsive agencies, regardless of whether the president is a copartisan or not. Our findings suggest that congressional delegation decisions are driven by a legislative view of the executive branch as a pluralistic internally-complex institution, the various parts of which are not perfectly under presidential control.

We expect that deep and active learning models of text classification can be used in various ways beyond agency delegation. In particular, we believe that problems involving the classification of complex texts, particularly those requiring expertise for labeling, can benefit from schemes similar to our method. We envision that this is only the first of many applications of this model to legislative texts. We are now interested in performing the same task on legislative constraints and regulations next to get a fuller sense of delegation in context, which opens up the scope of legislative outcomes dramatically, and encourages a clearer thinking about constructs.

References

Adler, E. Scott, & Wilkerson, John. 2011. *Congressional Bills Project*.

Anastasopoulos, L Jason, & Bertelli, Anthony M. 2020. Understanding delegation through machine learning: A method and application to the European Union. *American Political Science Review*, **114**(1), 291–301.

Bengio, Yoshua, Ducharme, Rejean, Vincent, Pascal, & Janvin, Christian. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*.

Casas, Andreu, Denny, Matthew J, & Wilkerson, John. 2020. More effective than we thought: Accounting for legislative hitchhikers reveals a more inclusive and productive lawmaking process. *American Journal of Political Science*, **64**(1), 5–18.

Clinton, Joshua D, & Lewis, David E. 2008. Expert opinion, agency characteristics, and agency preferences. *Political Analysis*, 3–20.

Clinton, Joshua D, Bertelli, Anthony, Grose, Christian R, Lewis, David E, & Nixon, David C. 2012. Separated powers in the United States: The ideology of agencies, presidents, and congress. *American Journal of Political Science*, **56**(2), 341–354.

Clinton, Joshua D, Lewis, David E, & Selin, Jennifer L. 2014. Influencing the bureaucracy: The irony of congressional oversight. *American Journal of Political Science*, **58**(2), 387–401.

Cox, Gary W, & McCubbins, Mathew D. 2005. *Setting the agenda: Responsible party government in the US House of Representatives*. Cambridge University Press.

Epstein, David, & O'Halloran, Sharyn. 1994. Administrative procedures, information, and agency discretion. *American Journal of Political Science*, 697–722.

Epstein, David, & O'Halloran, Sharyn. 1996. Divided government and the design of administrative procedures: A formal model and empirical test. *The Journal of Politics*, **58**(2), 373–397.

Epstein, David, & O'Halloran, Sharyn. 1999. *Delegating powers: A transaction cost politics approach to policy making under separate powers*. Cambridge University Press.

Farhang, Sean, & Yaver, Miranda. 2016. Divided government and the fragmentation of American law. *American Journal of Political Science*, **60**(2), 401–417.

Gailmard, Sean, & Patty, John W. 2012a. Formal models of bureaucracy. *Annual Review of Political Science*, **15**, 353–377.

Gailmard, Sean, & Patty, John W. 2012b. *Learning While Governing: Expertise and Accountability in the Executive Branch*. University of Chicago Press.

Glorot, Xavier, Bordes, Antoine, & Bengio, Yoshua. 2011. Deep Sparse Rectifier Neural Networks. *In: 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Grimmer, Justin, & Stewart, Brandon M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, **21**(3), 267–297.

Huber, John D, & Shipan, Charles R. 2002. *Deliberate discretion?: The institutional foundations of bureaucratic autonomy*. Cambridge University Press.

Kiewiet, D Roderick, & McCubbins, Mathew D. 1991. *The logic of delegation*. University of Chicago Press.

Kim, Yoon. 2014. Convolutional Neural Networks for Sentence Classification. *Pages 1746–1751 of: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kingma, Diederik P., & Ba, Jimmy. 2014. *Adam: A Method for Stochastic Optimization*.

Krutz, Glen S. 2001. *Hitching a ride: Omnibus legislating in the US Congress*. Ohio State University Press.

Lowande, Kenneth. 2018a. Delegation or Unilateral Action? *Journal of Law, Economics, and Organization*, **34**(1), 54–78.

Lowande, Kenneth. 2018b. Who Polices the Administrative State? *American Political Science Review*, **112**(4), 874–890.

Mayhew, David R. 1991. *Divided we govern*. Yale University New Haven, CT.

McCubbins, Mathew D. 1985. The legislative design of regulatory structure. *American Journal of Political Science*, 721–748.

McCubbins, Mathew D, Noll, Roger G, & Weingast, Barry R. 1987. Administrative procedures as instruments of political control. *Journal of Law, Economics, & Organization*, **3**(2), 243–277.

McCubbins, Matthew D, Noll, Roger G, & Weingast, Barry R. 1989. Structure and process, politics and policy: Administrative arrangements and the political control of agencies. *Virginia Law Review*, 431–482.

McGrath, Robert J. 2013. Congressional oversight hearings and policy control. *Legislative Studies Quarterly*, **38**(3), 349–376.

McNollgast. 1994. Legislative intent: The use of positive political theory in statutory interpretation. *Law & Contemp. Probs.*, **57**, 3.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg, & Dean, Jeffrey. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*.

Miller, Blake, Linder, Fridolin, & Mebane, Walter R. 2019. Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*, 1–20.

Miller, Gary J., & Whitford, Andrew B. 2016. *Above Politics: Bureaucratic Discretion and Credible Commitment*. Cambridge University Press.

Moe, Terry M. 2012. Delegation, control, and the study of public bureaucracy. *In: The Forum*, vol. 10. De Gruyter.

Pennington, Jeffrey, Socher, Richard, & Manning, Christopher D. 2014. GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing*.

Rodriguez, Pedro, & Spirling, Arthur. 2021. Word Embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*.

Rudalevige, Andrew. 2021. *By Executive Order: Bureaucratic Management and the Limits of Presidential Power*. Princeton University Press.

Selin, Jennifer L. 2015. What Makes an Agency Independence? *American Journal of Political Science*, **59**(4), 971–987.

Settles, Burr. 2009. *Active learning literature survey*. Tech. rept. University of Wisconsin-Madison Department of Computer Sciences.

Shaffer, Robert. 2020. Power in Text: Implementing Networks and Institutional Complexity in American Law. *Journal of Politics*.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, & Salakhutdinov, Ruslan. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**(56), 1929–1958.

Tong, Simon, & Koller, Daphne. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, **2**(Nov), 45–66.

Turian, Joseph, Ratinov, Lev, & Bengio, Yoshua. 2010. Word Representations: A Simple and General Method for Semi-Supervised Learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, July.

Vannoni, Matia, Ash, Elliott, & Morelli, Massimo. 2021. Measuring Discretion and Delegation in Legislative Texts: Methods and Application to US States. *Political Analysis*, **29**(1), 43–57.

Volden, Craig. 2002. A Formal Model of the Politics of Delegation in a Separation of Powers System. *American Journal of Political Science*, **46**(1), 111–133.

Zhang, Ye, & Wallace, Byron C. 2015. A Sensitivity Analysis of (and Practioners' Guide to) Convolutional Neural Networks for Sentence Classification. *In: arXiv preprint arXiv:1510.03820*.