No Public, No Power?

Analyzing the Importance of Public Support for Constitutional Review with Novel Data and Machine Learning Methods

Sebastian Sternberg

Mannheim, 2019

No Public, No Power?

Analyzing the Importance of Public Support for Constitutional Review with Novel Data and Machine Learning Methods

Sebastian Sternberg

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences
in the Graduate School of Economic and Social Sciences
at the University of Mannheim

written by
Sebastian Sternberg

Mannheim, May 2019

Dean: Prof. Dr. Michael Diehl
First Examiner: Prof. Thomas Gschwend, Ph.D.
Second Examiner: Prof. Dr. Christoph Hönnige
Third Examiner: Prof. Dr. Marc Debus

Date of oral examination: 17th July, 2019

Summary

Constitutional review is a central feature of liberal democracy. However, with neither the power of the purse nor the sword, the mere presence of constitutional courts does not automatically imply the effective exercise of judicial authority. Courts must rely on elected officials for the implementation of their rulings. The ability of a court to ensure that government officials faithfully comply with judicial decisions critically depends on the existence of sufficient public support for the court and the public's ability to monitor legislative responses to judicial decisions.

In this dissertation, I study the importance of public support for the relationship between *court-government* and *court-public*. I draw on the judicial politics literature on separation of powers, public support and legislative noncompliance and extend existing theory in two regards. First, I argue that not all courts possess the sufficient level of public support necessary to ensure legislative compliance. Varying degrees of public support strongly affect the leverage that courts possess in judicial-legislative and judicial-public interactions. Second, I argue that courts actively take measures in the form of the institutional tools at their disposal when they expect legislative noncompliance. One such tool is decision language, whose strategic usage allows judges to pressure the government or hide likely noncompliance from public view, if necessary.

I test these arguments empirically by combining classical inferential methods such as survey experiments with novel data on court decision-making and methodologies from the field of machine learning and computational linguistic. Throughout all chapters, I employ a comparative perspective and test my arguments using data on the German Federal Constitutional Court, a court with strong and robust levels of public support, and the less popular French Conseil Constitutionnel.

My empirical evidence shows that considering varying degrees of public support and the institutional tools of judges indeed helps to generate a more accurate picture of how

judges behave in judicial-legislative and judicial-public interactions. Three conclusions are drawn. First, court decisions can legitimize public policies, albeit only if the court itself is perceived as a legitimate institution. Second, courts are more attentive to the political environment of a decision than previously thought: depending on their degree of public support, they actively adapt the language of their decisions as a function of the risk of noncompliance and their institutional support. Third, public support and other political context factors are important for judicial decision-making not only from an inferential but also from a predictive perspective.

The results of my analyses confirm that public support plays a crucial role for courts' ability to effectively exercise constitutional review, as well as highlighting the benefits of increased differentiation of constitutional courts institutional tools and their diffuse support from a comparative view. Therefore, my results have implications for the growing literature on strategic courts using their institutional tools to address potential noncompliance and the general awareness of judges for their institutional reputation. Overall, this project offers new perspectives on the most important resource of judges – their public support – and has important implications not only for research on judicial politics but also for the efficacy of constitutional review in a constitutional state, and thus the sustainability of liberal democracy.

Acknowledgments

A dissertation is rarely the product of a single mind. Now that this project comes to an end, I would like to express my gratitude and appreciation to all the great people who provided me with critical discussions, technical assistance and encouragement during the past years. First, I would like to thank my dissertation committee: Thomas Gschwend, Christoph Hönnige and Marc Debus. My deepest thanks go to Thomas Gschwend. You have been the most supportive, most enthusiastic and also most demanding (in a strict positive sense) supervisor I could wish for, from the very beginning of my academic career, when I first worked as a student assistant at your chair until today. No matter how half-baked and unusual my ideas have been, you always listened to thoughts, you always gave valuable feedback and, most importantly, you took all my research ideas serious even when I was an undergraduate. Christoph Hönnige's research on constitutional courts as political actors was pivotal for my decision to do empirical research in this area in order to better understand these fascinating institutions. Thank you for your advice, especially about comparative judicial politics and the peculiarities of the French political system. I also wish to thank Marc Debus for his support by serving as the third examiner of this dissertation and his guidance as an undergraduate when writing my bachelor thesis.

My research would also not have been possible without the generous support from the Graduate School of Economic and Social Sciences at the University of Mannheim. The comments of my CDSS colleagues were very helpful for my work. In particular, I wish to thank Sebastian Juhl (although he supports the HSV), Verena Kunz, Sonja Pohle, Guido Ropers, Richard Traunmueller and Tilko Swalve.

Special thanks go to the current and former colleagues of the Chair of Quantitative Method, Anna Adendorf, Oke Bahnsen, Lukas Stötzer, Viktoriia Semenova, Lion Behrens and Sean Carey (from whom I learned a lot about passionate teaching and

local football). Five names deserve special mention. Christian Arnold, who gave me valuable life advice in many academic and non-academic questions. Benjamin G. Engst and Caroline E. Wittig shared the passion for judicial politics with me and supported me from my long journey as a student assistant to the final stages of this dissertation. I am indebted to Christel Selzer, who continuously covered my back in many regards. And of course, I wish to thank my awesome office-mate Marcel Neunhoeffer, with whom I spent most of my time during while developing this dissertation and to whom I owe a great friendship. Throughout all the last years, we suffered together from the sauna-like climate (aka tank top o'clock) in the summer, developed a similar enthusiasm for new methodologies and shared the same data- and prediction driven mindset (Fifa world cup 2018 prediction kings). Thanks a lot not only for your constant advice in all different theoretical and methodological questions, but also for always covering my back during tough times. C221 will never be forgotten.

Every project of this size requires amazing friends who help you to clear your mind when caught in unproductive thoughts. I am very grateful to have such true friends like Jan Crocoll, Dominik Scherer and Elias Schönthal and the Karlsruhe "it's escalating anyways" crew, who mastered this role with flying colors. Moreover, I owe a lot to my family. Above all, to my beloved mother Marit who always supports me in every possible way. It means the world to me that she is able to read these words today. I am thankful to my always caring father Klaus, who motivates me whenever necessary and always encouraged me to continue this project, even when I was in doubt. I owe a lot to my grandparents Else and Bertold, who always taught me that hard work will eventually pay off. I owe a lot to my amazing sisters Carolin and Teresa. Growing up with both of them has prepared me for all that is yet to come in life. Above all, my deepest thanks go to Lisa for bearing all my moods without complaint. Her love, her support, her understanding, her positivity and most of all her patience over the past years were indispensable for the completion of this project. I dedicate this book to her.

Contents

*

List of Figures

*

List of Tables

*

CHAPTER 1

Introduction

1.1 Motivation

On July 5, 2018, a headline in the New York Times claimed that the "Polish Crisis Deepens as Judges Condemn Their Own Court".[1] What has happened in this country which was once considered a role model for democratic transition following the constitutional revolutions in post-communist Europe? Beginning in 2015, Poland's ruling Law and Justice party (PiS) adopted several legal amendments to reshape the judicial frame work in Poland. These amendments sent 40% of the Supreme Court judges in retirement overnight and allowed the PiS to appoint their own judges instead. The Polish Constitutional Tribunal declared these developments and the amendments on which they were based as unconstitutional, although nothing happened and the PiS continued with their judicial reforms. The result of these events is a serious, persisting constitutional crisis in the heart of Central Europe, where judges condemn their own court and its output as politicized and unfree.

Similar developments are also observable in other countries. Viktor Orban's government in Hungary eludes court decisions by incorporating laws that have been declared as unconstitutional directly into the constitution. In addition, the Hungarian Government continues to curtail the rights and competences of the constitutional court, which is now only allowed to examine the formal but not substantive constitutionality of a law. Moreover, court crises are not unique to Europe; in the United States, the

[1] `https://www.nytimes.com/2018/07/05/world/europe/poland-court-crisis-constitutional-tribunal.html` (Santora, 2018), accessed 24.04.2019.

controversial appointment of the Supreme Court Justice Brett Kavanaugh caused an erosion of trust in the Supreme Court according to polls from Gallup.[2]

Although such open noncompliance as shown by the Polish and Hungarian governments is rare, and the Supreme Court's popularity typically recovers after some time, these developments shed light on a general problem: what are judges supposed to do in times of a crisis of the constitutional state where governments can openly ignore judicial rulings, and the legitimacy of courts begins to erode? Constitutional review is a central feature of Western-style democracy. However, with neither the power of the purse nor the sword, the mere presence of constitutional courts does not automatically imply the effective exercise of judicial authority. Courts must rely on elected officials for the implementation of their rulings. The ability of a court to ensure that government officials faithfully comply with judicial decisions decisively depends on the existence of sufficient public support for the court, and thus its institutional legitimacy, as well as the public's ability to monitor legislative responses to judicial decisions.

These events demonstrate that the question of how courts can effectively exercise constitutional review holds crucial relevance. I therefore take the opportunity to *study the relationship between the court, the government and the public in this dissertation*. In particular, I am interested in studying the importance of public support for the relationship between *court-government* and *court-public*. I draw on the comparative judicial politics literature on separation of powers, public awareness and legislative noncompliance and extend existing theory in two regards. First, I argue that *not all courts possess the sufficient level of public support* necessary to ensure legislative compliance. The degree of public support strongly affects the leverage that courts possess in judicial-legislative and judicial-public relations, where less popular courts have a much lower institutional legitimacy and thus leverage. Second, I argue that *courts actively take measures in the form of the institutional tools at their disposal to counter legislative noncompliance*. One such tool is decision language, whose strategic usage allows judges to pressure the government or hide likely noncompliance from public view, if necessary.

I test these arguments empirically by combining classical inferential methods such as survey experiments with novel data on court decision-making and methodologies from the field of machine learning and computational linguistics. Throughout the majority of the chapters, I employ a comparative perspective and test my arguments using data on the German Federal Constitutional Court, a court with strong and robust levels of public support, and the less popular French Conseil Constitutionnel.

My empirical evidence shows that considering varying degrees of public support and the institutional tools of judges is beneficial for obtaining a more accurate picture of how judges behave in judicial-legislative and judicial-public interactions. I draw

[2] `https://fivethirtyeight.com/features/is-the-supreme-court-facing-a-legitimacy-crisis/` (De-Veaux and Roeder, 2018), accessed 03.05.2019.

three main conclusions. First, court decisions can legitimize public policies, albeit only if the court itself is perceived as a legitimate institution. Second, courts are more attentive to the political environment of a decision than previously thought: depending on their degree of public support, they actively adapt the language of their decisions as a function of the risk of noncompliance and their institutional support. Third, public support and other political context factors are important for judicial decision-making not only from an inferential, but also from a predictive perspective.

The results of my analyses confirm that public support plays a crucial role for courts' ability to effectively exercise constitutional review, as well as highlighting the benefits of a greater differentiation of constitutional courts' institutional tools and public support from a comparative view. My findings therefore hold implications for the growing literature on strategic courts that use their available institutional tools to address potential noncompliance as well as the general awareness of judges for their institutional reputation. Overall, this project offers new perspectives on the most important resource of judges – their public support – and is relevant not only for research on judicial politics, but also for the efficacy of constitutional review in a constitutional state, and thus the sustainability of liberal democracy.

1.2 General Framework: Constitutional Review, Legislative Noncompliance and the Importance of the Public

The general framework of this dissertation is the relationship between the court, the government and the public. I place a particular emphasis on the *court-government* (judicial-legislative) and *court-public* (judicial-public) interaction. In what follows, I provide an overview concerning the meaning of constitutional review for liberal democracy, the problem of legislative noncompliance of judicial decisions and the importance of the public as the most important resource of courts in this regard. I conclude that although formally constitutional review is a powerful mean for courts to control legislative majorities, in reality this power is often limited and ultimately depends on whether the public supports a court or not.

1.2.1 The Power of Constitutional Review

The power of constitutional review is a central feature of Western democracies. Defined as "the formal power of a judicial body to set aside or strike legislative or administrative acts for incompatibility with the national constitution" (Vanberg, 2005, 1), many highest courts are empowered to protect constitutional rights and to oversee the political process. Constitutional review is widely spread: by 2011, 83% of the world's constitutions empowered the judicial branch with the authority of constitutional review (Ginsburg and Versteeg, 2014, 2).

These courts play a pivotal role in the political system and democratic politics. They solve conflicts between the legislative majority and the opposition, protect the basic rights and liberties of citizens, and monitor whether legislative actions are consistent with the constitution. With their ability to review and invalidate legislation, courts can serve to constrain governing majorities and constitute an important part of the system of checks and balances in Western democratic political systems. While these "guardians of the constitution" thus play a pivotal role in democratic politics, the (lack of) formal empowerment of constitutional courts does not guarantee that they can effectively exercise their authority and constrain the legislature.

1.2.2 Legislative Noncompliance

"The decisions of the Federal Constitutional Court shall be binding upon the constitutional organs of the Federation and of the Laender, as well as on all courts and those with public authority"[3] (§ 31,1 Act on the Federal Constitutional Court). This sentence of the Act on the Federal Constitutional Court illustrates that other institutions and political actors are formally constrained and tied by German constitutional court's decisions. In reality, however, the potency of constitutional review is limited.

With the words of Alexander Hamilton's often-quoted phrase, constitutional courts are the "least dangerous branch of government" because in contrast to the legislature and the executive, they neither control the "purse" nor the "sword" (Hamilton 1788, Federalist No. 78) . Therefore, in contrast to government officials or other policy-makers, the judiciary possess little formal power to enforce or implement their rulings. Instead, how and whether a judicial decision is implemented depends on the willingness of the branches of the government to faithfully comply with a ruling. In cases where a court decision corresponds to the interests of the government, the institutional weakness of courts is not problematic, since governments have no reason to not comply. However, this changes if the outcome of a ruling is not in line with the government's preferences. In such a case, court decisions are often an unwelcome constraint to a government's power and a threat to the legislative agenda of the elected officials. In such a case, the implementation of a court ruling is then in the hand of actors who have a strong interest that the court's ruling will not be implemented. This non-implementation of court decisions is generally referred to as *legislative noncompliance*.

The possibility of legislative noncompliance is not simply academic. In order to illustrate that even powerful courts face a real threat of noncompliance, I provide three short case studies of occasions where legislative noncompliance occurred in the context of decisions of the German Federal Constitutional Court (hereafter referred to as the GFCC).

[3] German original: "Die Entscheidungen des Bundesverfassungsgerichts binden die Verfassungsorgane des Bundes und der Länder sowie alle Gerichte und Behörden".

Example 1: The Crucifix decision

In August 1995, the GFCC ruled on the constitutionality of a Bavarian school ordinance on displaying a crucifix in public elementary school classrooms.[4] The German judges decided that displaying a cross or crucifix in class rooms is unconstitutional. This provoked harsh critique and public protests, and Church leaders, but also politicians (mostly of the Christian Democrats) took public position against the ruling. Edmund Stoiber, the former Bavarian prime minister, officially vowed that crucifixes would remain in the classrooms. With respect to the implementation of the court's decision, he deemed, subtly but unequivocally, that "the crucifix decision will be *respected*, but never *accepted*".[5] As a reaction to the decision, the Bavarian parliament passed a revision of the school ordinance where **not** displaying the crucifix is only possible in rare, atypical cases after all other means are exhausted and no compromise between the different interests is possible. Constitutional scholars, therefore, evaluate that the crucifix decision of the GFCC had no practical consequences overall (Schaal, 2007). In fact, a judge of the GFCC later said during a lecture that "there are more crucifixes hanging in Bavarian school rooms now than before the decision." (Vanberg, 2005, 4).

Example 2: The inheritance tax decision

In a landmark decision on inheritance tax in Germany in 2014, the GFCC declared the existing inheritance tax law with respect to family-owned companies as unconstitutional.[6] In their ruling, the judges instructed the federal legislature to revise the law and set a deadline. The court's decision received harsh critique by governmental officials. In an anonymous interview, government officials commented on the decision, stating that "what Karlsruhe brought to us is indescribable. The decision is written in a way that shows that the judges just write what comes to mind in their professorial delusion. And we are facing the trouble afterwards because it is ill-conceived".[7]

The Bundestag instituted a committee to study revisions of the inheritance tax code, although no legislation was initiated. After two years, the deadline expired without a legislative response. This lead the President of the First Senate of the German court to write an open letter to the German legislators, highlighting that the court felt obliged to deal with this matter again.[8] It was only after this threat that the government (hastily) adopted a law revision, which is currently discussed to be appealed to the GFCC again by the opposition.

[4] Reference number: BVerfGE 93, 1.

[5] `http://www.bpb.de/apuz/33164/regiert-karlsruhe-mit-das-bundesverfassungs-gericht-zwischen-recht-und-politik?`, accessed 26.04.2019.

[6] Reference number: 1 BvL 21/12.

[7] `https://www.zeit.de/zeit-magazin/2016/12/andreas-vosskuhle-bundesverfassungsgericht-verfassung-praesident/seite-2` (Wefing, 2016), accessed 24.04.2019.

[8] `https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2016/bvg16-041.html`, accessed 12.04.2019.

Example 3: The NPD campaign event provisional order

In March 2018, the third chamber[9] of the first Senate of the GFCC ruled in a provisional order that the National Democratic Party of Germany (German: Nationaldemokratische Partei Deutschlands, NPD) is allowed to host an election campaign event in the city hall of Wetzlar, a city in Hesse.[10] Prior this decision, the city of Wetzlar had attempted to prohibit the event for formal reasons. The NPD appealed against the prohibition to an administrative court, and won. The responsible administrative court ruled in a provisional order that the city of Wetzlar must confer the right to use the city hall to the NPD. Because the city of Wetzlar ignored the provisional order, the NPD appealed to the GFCC. In a provisional order, the GFCC decided that the city of Wetzlar must respect the administrative court ruling and that the NPD is explicitly allowed to host an election campaign in the city hall.

In the following, the city of Wetzlar also ignored the decision of the GFCC, whereby ultimately the election campaign event did not take place in the city hall. The court condemned the actions of the city of Wetzlar with unusual sharp words in a press release, stating that "it will inform the Prime Minister, the Minister of Justice and the mayor of Wetzlar to elucidate the incidents which lead to ignoring the court order".[11] In a follow-up statement, the court reports that the city of Wetzlar has clarified the circumstances that led to ignoring the decision, whereby there were "misconceptions about the binding power of judicial decisions and the remaining scope for own actions". However, there was nothing the court could do but "encourage the municipal supervision to ensure that judicial decisions will be respected in the future".[12]

In the light of these events, Andreas Vosskuhle, the sitting President of the German Federal Constitutional Court, stated in a recent interview that:

> "Judicial decisions, be they of first-instance courts or of the Federal Constitutional Court, must be respected and implemented by other public authorities. Otherwise, it is a violation of the lawful promise that we have given each other in the Federal Republic. A violation that cannot be tolerated."[13]

[9] Chambers are panels of three judges that assist the GFCC to make timely decisions for a large number of cases that are deemed to be not sufficiently important or controversial to be deliberated among all judges on the bench.

[10] Reference number: 1 BvQ 18/18.

[11] `https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2018/bvg18-016.html`, accessed 26.04.2019.

[12] `https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2018/bvg18-026.html`, accessed 26.04.2019.

[13] German original: "Gerichtliche Entscheidungen, seien sie von erstinstanzlichen Gerichten oder vom Bundesverfassungsgericht, sind von anderen Hoheitsträgern zu respektieren und umzusetzen. Andernfalls ist es ein Verstoß gegen das rechtsstaatliche Versprechen, das wir uns gegenseitig in der Bundesrepublik gegeben haben. Ein Verstoß, der nicht zu tolerieren ist". `https://www.das-parlament.de/2018/40_41/im_blickpunkt/571052-571052` (Dolderer, 2018), accessed 20.04.2019.

Although these examples are all from Germany, case studies of legislative noncompliance in other political systems are also available, e.g. in Italy (Volcansek, 1991), Russia (Trochev, 2002, 2008), or the United States (Vanberg, 2005). The occasional use of noncompliance is also documented in other contexts (Staton, 2006, 2010; Carrubba, Gabel and Hankla, 2008; Carrubba and Zorn, 2010). All of these episodes underscore that the mere presence of a constitutional court does not ensure that it can effectively exercise constitutional review and control the other branches of government. Ignoring judicial decisions is a viable option for politicians if they are not satisfied with judicial outputs.

1.2.3 The Efficacy of Constitutional Review and Public Support

Overcoming the threat of legislative noncompliance is a fundamental challenge of judicial decision-making. Previous judicial scholarship has identified *public support* for courts as one solution to this challenge. Public support is also called *diffuse support* and is generally defined as a "reservoir of favorable attitudes or good will that helps members to accept or tolerate outputs to which they are opposed or the effects of which they see as damaging to their wants" (Easton, 1965, 273). With respect to courts, diffuse support is understood here as the "support for maintenance of the institution" (Caldeira and Gibson, 1992, 638), namely the public support for the court as a legitimate institution and central feature of a constitutional state and the rule of law.[14]

With respect to the importance of public support for courts, the linchpin of the argument is that with sufficient public support, legislative attempts of noncompliance will result in a public backlash which is electorally costly, whereby the legislative thinks twice before it evades a decision. Therefore, public support is the most important weapon that judges have in the judicial-legislative relations when they expect governmental resistance, and thus the public plays a major role for the efficacy of constitutional review. In this regard, public support is like a "reservoir of good-will" that courts can rely on. Consequently, courts have a strong interest in maintaining this support and are "with limited institutional resources, [...] uncommonly dependent upon the goodwill of their constituents for both support and compliance" (Gibson, Caldeira and Baird, 1998, 343).

If public support alone were sufficient to achieve legislative compliance, then decisions of courts with such diffuse support would never be evaded. However, as previously outlined, legislative noncompliance is an empirical reality. To explain this, Vanberg (2001, 2005) argues that public support can only effectively help judges to constrain governments if the *public awareness* for a decision is sufficiently high. Only if the citizens are aware of a judicial decision and the legislative response, they can successfully

[14] Diffuse support is different to specific support because it describes the support for a court although the court makes unpopular decisions from time to time.

monitor compliance, and eventually punish the government for noncompliance. I will outline in the next section that several possibilities exist for courts to increase the probability that the public will catch governmental attempts of evasion.

1.3 Approach of this Dissertation

In this dissertation, I draw on the comparative judicial politics literature on the separation of powers, public support and legislative noncompliance and extend existing theory in two regards. First, in line with an emerging strand of literature, I understand judges as "active" actors who can strategically resort to institutional tools at the court's disposal to increase the likelihood of legislative compliance. Second, I argue that not all courts possess the sufficient level of public support necessary to effectively constrain the government. Varying degrees of public support strongly affect the leverage that courts posses in the judicial-legislative and judicial-public relationship.

Courts as Active Actors

Existing studies often assume that constitutional courts are "passive" actors. With respect to the interplay between courts, governments and the public, this means that the public awareness for a given decision is assumed to be exogenous. For instance, a key argument in Vanberg's (2005) formal model is that courts rule differently in cases where the necessary public attention for a decision is given compared with cases where it is not.

However, an emerging strand of literature demonstrates that courts are more "active" than previously thought and not entirely helpless when governmental resistance is likely. In context of the Mexican Supreme Court, Staton (2006, 2010) shows that courts strategically issue press releases announcing the result of a decision. Krehbiel (2016) finds evidence that the GFCC uses oral hearings to raise public awareness of a case, and that the German judges strategically time decisions based on the temporal proximity of the next election (Krehbiel, 2019). In the same line, Engst (2018) shows that judges strategically use directives as a way of requesting political actions depending on the costs for the government to comply with decisions. All of these findings suggest that courts actively use the institutional means at their disposal in an attempt to strategically raise public awareness when they expect legislative noncompliance or to make compliance more likely in other regards.

This has severe consequences for the amount of self-restraint that judges face. When courts are assumed to be "passive" actors, every threat of legislative noncompliance without sufficient public awareness for a case must lead to judicial self-constraint. This means that if it is not sufficiently likely that citizens become aware of an evasion attempt, the court has no other choice than to restrain itself. However, if courts can

manipulate the public awareness for a decision, they are in a much more powerful position because self-restraint is no longer necessary.

In this dissertation, I follow the emerging strand of literature and understand *judges as "active" actors who can strategically resort to institutional tools* at the court's disposal to increase the likelihood of legislative compliance. Instead of press releases, oral hearings or the timing of a decision, I look at another means that judges have at their disposal: *decision language*. Put simply, decision language offers courts a mechanism to either increase the pressure on the legislature for compliance or hide legislative noncompliance from public view (Staton and Vanberg, 2008). It can be therefore used as a tool to control the public awareness surrounding a decision, and thus it has stark implications for the effectiveness of constitutional review in judicial-legislative relations. I will further elaborate on this argument in Chapter 4.

Courts With Varying Degrees of Public Support

High diffuse support for constitutional court is the center piece of many studies on the interaction between courts, governments and the public. While it is acknowledged that the diffuse support for court experiences some ebbs and flows in response to popular and unpopular rulings (Durr, Martin and Wolbrecht, 2000; Casillas, Enns and Wohlfarth, 2011), a central assumption in many studies is that diffuse support is permanently high. Many formal models of judicial decision-making in the context of noncompliance rely on the assumption that all courts enjoy a sufficient and high level of public support. These model's central implications only hold as long as the court's diffuse public support is sufficiently large (e.g. Vanberg, 2001, 2005; Staton, 2006; Krehbiel, 2016, 2019). For instance, Vanberg (2001, 2005) argues that public support for the German constitutional court itself is permanently high, whereas the public awareness of a case is not always guaranteed (Vanberg, 2005, 21). Thus, observing legislative noncompliance in some cases but not in others is dependent on the likelihood that the public will take notice of the evasion. In the same line, Krehbiel (2016) assumes that "the court enjoys a high level of public support such that the government is always punished when the public observes noncompliance" (Krehbiel, 2016, 992).

This assumption does not account for the empirical reality. The notion that courts substantially differ in their diffuse support is already documented early in survey research on mass attitudes towards constitutional courts and supreme courts in advanced democracies (Gibson, Caldeira and Baird, 1998). While it is true that courts in general are highly-respected institutions in many countries, there is considerable cross-national variation in the level of public support. As Gibson, Caldeira and Baird (1998) note, "national high courts vary enormously in the degree to which they have achieved institutional legitimacy" (Gibson, Caldeira and Baird, 1998, 356). In Chapter 2, I support this claim with evidence from a cross-national survey conducted in Germany and France.

The reasons for low institutional legitimacy are manifold. Some courts, such as those newly installed during the third wave of democratization, have had little opportunity to build institutional trust compared with long-established courts such as the US Supreme Court or courts installed after the Second World War (Gibson, Caldeira and Baird, 1998, 350). The legitimacy of other courts such as the French Conseil Constitutionnel suffers from a highly politicized appointment process (Hönnige, 2009) and the notion that it has only received the power of protecting citizens' rights rather recently (Brouard, 2009).

Given that public support is the most powerful weapon that judges possess when dealing with governmental resistance, and thus it directly affects their leverage in the judicial-legislative interaction, varying levels of public support also imply varying levels of leverage. *Courts without robust and high levels of public support must behave differently than courts with strong foundations of public support*. Those who enjoy reasonably high public support can use this resource against the government in the way outlined above. However, courts that enjoy much lower levels of public support do not have this leverage.

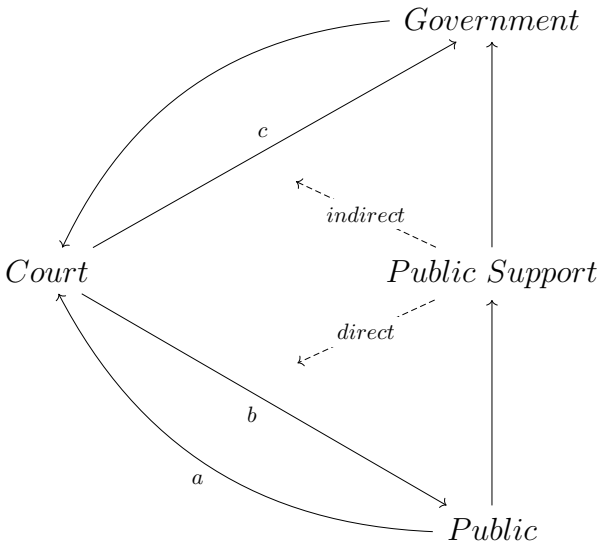
Unfortunately, there is little comparative work that systematically accounts for varying degrees of public support across countries empirically and theoretically. In this dissertation, I therefore incorporate the variation in diffuse support in my theoretical treatments and empirical analyses. In particular, *I explore how varying degrees of public support affect the possibility of courts to force governmental compliance and the legitimacy-conferring capacity of these courts*.

1.4 Research questions

In this dissertation, I study the relationship between constitutional courts, legislative majorities and the public, devoting particular attention to the intersection between *court-government* and *court-public*. I extend previous work in two regards. First, I follow the emerging strand of literature and understand judges as "active" actors who can strategically resort to institutional tools at the court's disposal to increase the likelihood of legislative compliance, namely decision language. Second, I explore how varying degrees of public support affect the possibility of courts to force governmental compliance and the legitimacy-conferring capacity of these courts. Figure 1.1 shows the basic structure of the relationships analyzed in this study. The figure shows the importance of public support for the relationship between the court and government (indirect effect) and between the court and public (direct effect).

To start with, public support has a direct effect on the court-public relation because courts are "with limited institutional resources, [...] uncommonly dependent upon the goodwill of their constituents" (Gibson, Caldeira and Baird, 1998, 343). Because diffuse support and specific support are often viewed as being linked (Caldeira and Gibson,

Figure 1.1 – Relationships Analyzed in this Study



1992; Gibson, Caldeira and Baird, 1998; Gibson, Caldeira and Spence, 2003*a*), judges need to be sensitive to public opinion. If judges too often make unpopular decisions and repeatedly issue deviant rulings, they risk losing their "reservoir of good will" and therefore their diffuse support in the long run. In other words, judges have to adjust their rulings to the prevailing public opinion. This relationship is denoted with *a)* in Figure 1.1. The judge's sensitivity to public mood can also be illustrated with a statement of a personal interview conducted with a GFCC judge in Vanberg's (2005) book:

> "There cannot be a long-running divergence between the views of the public at large and the jurisprudence of the court. The court must be carried by a consensus of the citizens...it is important to take care that a decision does not hit on a weak spot in public consensus [...] The decisions have to be understandable and acceptable." (J8,6) (Vanberg, 2005, 126)

This direct effect of public opinion on judicial decision-making has already been investigated in various studies with respect to the US Supreme Court (e.g. Gibson, Caldeira and Spence, 2003*b*; Giles, Blackstone and Vining, 2008; McGuire and Stimson, 2008; Casillas, Enns and Wohlfarth, 2011; Hall, 2014; Clark and Kastellec, 2015), and sporadically with evidence from European constitutional courts (Sieberer, 2006; Sternberg et al., 2015). I build on this work but analyze the importance of public opinion on court decision-making *from a predictive rather than an inferential perspective*. I use a machine learning algorithm to evaluate the contribution of public opinion and other political aspects of judicial decisions for predicting the decision-making of the GFCC. My first research question is thus:

To what extent do political context factors (including public opinion) contribute to the prediction of court decision-making?

While the question above explores the consequences of public opinion for judicial decision-making, in the next question I reverse the causal arrow and analyze how judicial decisions affect public opinion on specific issues. This is denoted with *b)* in Figure 1.1. Whereas this effect is already studied with respect to the US Supreme Court (Hoekstra, 1995; Clawson and Kegler, 2001; Bartels and Mutz, 2009; Marshall, 1987; Stoutenborough, Haider-Markel and Allen, 2006), I extend previous work by looking at two European constitutional courts (the German Federal Constitutional Court and the French Conseil Constitutionnel) in a comparative perspective. I expect to find systematic differences between the legitimacy-conferring capacity of these courts because the two courts substantially differ in their level of public support. In particular, I am interested in whether public opinion on a governmental policy moves in response to a court's decision output. My second research question is thus:

Can constitutional court decisions shape public opinion on a governmental policy?

The next question explores the indirect effect of public support on the *court-government* relationship, denoted by *c)* in Figure 1.1. The effect is only indirect because public support does not directly affect the behavior of the court, but indirectly through its function as a pressuring tool. In the judicial-legislative relation, previous literature mainly treats public support and public awareness for a given case as an exogenous factor. Courts are understood here as "helpless" actors that have no possibility to influence how the public thinks about a case or whether citizens even take notice of a decision. I challenge this view and follow the emerging literature highlighting the ability of judges to strategically use the institutional tools at their disposal to control the public awareness of a case, and thus understanding judges as active actors. I argue that vague language is yet another tool that judges can use to promote compliance (Staton and Vanberg, 2008). My first research question concerns measuring vagueness in judicial texts across different languages. Because most of the work on the measurement of vagueness in political texts still has a lot of development potential, I raise the question:

How can we automatically measure vague language in written court decisions?

My final research questions assesses the value of vague language for judges. I conduct an empirical test of the implications of a game-theoretic model developed by Staton and Vanberg (2008). This model argues that courts use vague language strategically as a function of their (limited) policy expertise, the preference divergence with the

legislator and the public support of the court. Varying degrees of public support affect the decision of courts to either use decision language to apply pressure to the government or hide likely noncompliance from public view. Again, I employ a comparative perspective using the French and German constitutional courts. With respect to the strategic usage of language by judges, my final research question simply asks:

Why do courts craft vague decisions?

Having provided a summary of the research questions that I will answer in this dissertation, I will subsequently discuss the key innovations and contributions of my study.

1.5 Key Innovations and Contributions

In the following section, I outline the specific innovations of my work and highlight the relevant contributions for the field of judicial politics and political science in general.

Text As Data: a Novel Measure for Judicial Vagueness

Automated methods for content analysis are a common approach to extract information from political texts. These automated methods such as dictionary methods, supervised methods for classification or unsupervised methods for clustering promise to overcome researchers' need to manually read through massive collections of documents or hire expensive human coders to undertake this task. Such methods are also increasingly used in the context of judicial decisions (e.g. Owens and Wedeking, 2011, 2012; Cross and Pennebaker, 2014; Black et al., 2016*b*; Wedeking and Zilis, 2018). Current researchers exhaustively rely on dictionary methods. Dictionaries use the frequency of the occurrence of key words in a text to classify documents into categories. Such dictionaries are used, for instance, to measure the "cognitive complexity" of Supreme Court justice's opinions (Owens and Wedeking, 2012), the "legal clarity" of Supreme Court decisions (Owens and Wedeking, 2011) or the "textual readability" of Supreme Court rulings (Black et al., 2016*b*, Chapter 3).

These dictionaries should be used with caution or at least with detailed validation, because serious error can occur when dictionaries from one domain are applied to another (Grimmer and Stewart, 2013, 275). Unfortunately, this is not always the case when e.g. dictionaries originating from the field of psychology are applied to judicial decision texts. The consequence is that most of these analyses are built on "shaky foundations" (Grimmer and Stewart, 2013, 275).

In this dissertation, I overcome this problematic practice by providing two different measurement strategies to automatically detect vague language in judicial decisions. I rely on recent advances in computational linguistic to *a)* demonstrate how a general

dictionary can be expanded to a specific domain using word embeddings and *b)* how to develop and benchmark state-of-the-art supervised machine learning classifiers for this task. I also provide an exhaustive validation of the proposed measures.

Three contributions are made. First, the construction of these measurement approaches is not unique to judicial texts but can also be applied to other problems. Accordingly, my approaches can work as a blueprint for any other application where scholars are interested in automatically identifying a latent concept in large collections of text. Second, these measurements are not developed for the sake of creating a new textual measure; rather, the clear focus is on application, whereby I use these measures to test a game-theoretic model in a subsequent chapter. Finally, recent trends in judicial politics stress the importance of more closely accounting for content-specific characteristics of judicial decisions rather than simply coding them in binary outcomes (Staton and Vanberg, 2008; Clark and Lauderdale, 2010; Engst, 2018). This dissertation is therefore a step forward to close the gap between legal scholarship and quantitative judicial politics by considering the actual content of court decisions.

The Value of Predictive Modeling for the Field of Social Science

Social scientists traditionally prioritize inferential methods over other methodologies to draw causal inferences. This often includes strong assumptions about the data generating process, the development and testing of carefully designed theories and the formulation of testable observable implications. By contrast, predictive modeling, namely "the use of the available data to produce the best possible predictions of the outcome variable" (Cranmer and Desmarais, 2017, 146), is still relatively rarely used. This also applies to research on judicial politics.

This is unfortunate, given that predictive modeling and artificial intelligence (AI) play an increasingly important role in law. Estonia plans to replace human judges with an "AI judge" in the summer of 2019 to settle conflicts in private law based on an algorithm.[15] Daniel Martin Katz, a legal scholar working on the intersection between law and computer science (computational legal studies), states that quantitative legal prediction will replace human assessment in many instances of typical legal work, a "law's information revolution" that he calls the "data driven future of the legal service industry" (Katz, 2013) (see also Surden (2014) for a discussion of the general relationship between machine learning and law).

In this dissertation, I argue that as social scientists we can also learn a lot from predictive models. Predictive modeling offers an excellent possibility to compare competing theories that examine the same outcome (Cranmer and Desmarais, 2017, 149) (see also Shmueli (2010) for an in-depth discussion of causal versus predictive

[15] `https://www.deutschlandfunknova.de/beitrag/digitalisierung-ki-richter-in-estland-faellt-urteile-per-algorithmus`, accessed 20.04.2019.

approaches). In particular, focusing on the prediction of a phenomenon is a simple mean to verify the extent to which "theoretically informed models anticipate reality, and which among those models does a better job of it" (Cranmer and Desmarais, 2017, 149).

In Chapter 5 of this dissertation, I follow such a predictive approach and test the predictive contribution of legal and political context for the forecast of GFCC decision-making in a machine learning application. I offer three distinct innovations. First, predictive modeling allows me to test whether variables associated with the legal context of a decision are sufficient to predict court decision-making, and whether political context adds to the prediction. This in turn has implications for our understanding of the predictability of judicial decision-making. Second, I show that machine learning methods are helpful to detect non-linearities in the data that conventional regression models most likely would not have detected. Third, the combined results of the first two points prompt new research questions that require novel or refined theories, and allow me to create the first benchmark of predictive accuracy for decisions of the GFCC. While I do not argue that machine learning will replace conventional statistical social science methods, algorithmic procedures will become increasingly common as a supplementary tool in the tool box of quantitative social scientists to tackle research questions from different perspectives.

Beyond the US Supreme Court: A Comparative Perspective on Germany and France

In 2011, Hönnige (2011) stated that we need more comparative research regarding European constitutional courts, especially compared with the US Supreme Court, which has been at the center of scholarly attention for decades. The extent of research on the US Supreme Court strongly differs from that on European constitutional courts, and is fairly well developed. To illustrate, whereas US scholars have recently used the level of emotional arousal in Supreme Court justices' voices during oral arguments to predict the voting behavior of these justices (Dietrich, Enos and Sen, 2018), European scholars still face a severe lack of available data and common measures that are critical to analyze judicial behavior (e.g. measures of ideology or public mood). These obstacles are even more prevalent in comparative research.

These problems are also reflected in the amount of scholarly work on these courts. In an analysis of scholarly literature from 1995 to 2008, Hönnige (2011) shows that more work has been published on the US Supreme Court in the American Journal of Political Science alone than on other courts in six comparative and two national journals together. In Table 1.1, I provide a similar count of the number of articles in two US journals (*American Journal of Political Science, American Political Science Review*), two European comparative journals (*European Journal of Political Research*, *West European Politics*) and two national outlets (the German *Politische Vierteljahresschrift* and the French *Revue Française de Science Politique*) spanning from 2011 to 2019. I searched for all articles

Table 1.1 – Number of articles concerned with courts in selected journals

Journal/Year	2011	2012	2013	2014	2015	2016	2017	2018	Sum
American Journal of Political Science	1	3	2	6	5	3	1		19
American Political Science Review		2	1	1		1	1		6
European Journal of Political Research		2					1		3
West European Politics	3		2	2			3		10
Politische Vierteljahresschrift					1		1		2
Revue Française de Science Politique									0

that either deal with the US Supreme Court or another US Court (such as state courts) and all articles concerning European national high courts and the European Court of Justice. My brief analysis reaches a similar conclusion to Hönnige (2011): even eight years after his call for more (comparative) research on European constitutional courts, the US Supreme Court still dominates the literature on judicial politics. Again, the American Journal of Political Science published more (19) articles on courts than all European outlets together (15).

In order to move beyond the focus on the US Supreme Court and to follow the call of Hönnige (2011), this dissertation offers a comparative perspective on the relationship between the court, government and the public by employing a comparative research design throughout the chapters. I systematically compare the decision-making of two European constitutional courts: the German Federal Constitutional Court and the French Conseil Constitutionnel (hereafter *CC*). The countries are selected because they share many institutional features but substantially differ in one aspect central to this dissertation: their diffuse support. I elaborate more on this point and provide a more substantial discussion of my case selection in Chapter 2.3. Put briefly, both courts are established constitutional courts, both have the right of constitutional review and often they have to deal with political questions of major societal significance in their rulings. However, as previously outlined, in this dissertation I want to carve out the implications that varying levels of diffuse support hold for judicial decision-making. The approximation of these variation via the case selection is an elegant way to overcome the lack of cross-national survey data on the diffuse support of courts over time.

Using a comparative approach is associated with in increased effort compared with analyzing a single court, given that the data must be collected or be available for the two countries, comparable measurements must be created, and the validation of these measures must be undertaken in a comparative manner. In Chapter 3, for instance, I must show that my vague language measure is not an artifact of the language in which it is written in, but rather that it is comparable across countries.

Nonetheless, the comparative approach also has several advantages. First, it allows me to approximate the varying degrees of institutional public support via the case selection instead of relying on only rarely and not comparably appearing cross-national

surveys. Second, the external validity of comparative studies is higher than that of single-country studies (see e.g. Newton and Van Deth, 2012, 2-5). Third, my work is a step forward towards promoting more comparative research on European constitutional courts, which remain heavily under-researched. This is particularly important with respect to the generalization of the findings of this study. The new courts in Eastern Europe, for instance, are mainly modeled along the lines of the German Federal Constitutional Court, and thus my results also have implications for these types of courts.

1.6 Plan of the Dissertation

In **Chapter 2**, I directly investigate the relationship between courts and the public by asking whether constitutional courts can shape public opinion on a governmental policy? In a nutshell, I argue that the institutional legitimacy of courts allows them to move public opinion on public policies into the direction of their rulings. I test this using a comparative survey experiment conducted in Germany and France by confronting citizens with different court endorsements on a public policy. I find evidence that public opinion, even amongst those with strong prior attitudes, is shifted towards the court's ruling. However, and this is the novelty presented in this chapter and the connection to my theoretical arguments, I demonstrate that this only holds for courts who possess about sufficient institutional legitimacy.

Chapter 3 is devoted to the measurement of vague language in court decisions. I argue that currently used measurement approaches are insufficient because they ignore the peculiarities of judicial texts. To overcome the limitation of existing work, I develop the concept of judicial policy implementation vagueness as a particular form of vagueness unique to judicial decisions. Based on recent advances in computational linguistics, I show that this concept can be measured with a dictionary-based approach using word embeddings and a machine learning classifier trained and benchmarked on a novel self-collected data set. Because both measurements are tailored to the judicial domain, they outperform existing measures of vague language.

In **Chapter 4**, I analyze why courts craft vague decisions and explore the decision language of the German and French constitutional court in a comparative study. I use the judicial policy implementation vagueness scores from Chapter 3 to test the implications of a game-theoretic model of Staton and Vanberg (2008). This model argues that courts use vague language strategically as a function of their (limited) policy expertise, the preference divergence with the legislator and the public support of the court. I find support for most of the central implications of the formal model. Most importantly, my results show that the level of public support of a court is crucial for explaining whether courts use decision language to strategically pressure the government of to hide likely noncompliance from the public.

Chapter 5 leaves behind the field of traditional inferential statistics and enters the world of predictive modeling. I evaluate whether it is feasible to correctly forecast the decision-making of the GFCC using a machine learning algorithm. I find that it is possible to correctly predict three out of four outcomes of over 2,900 proceedings of the GFCC decided between 1972 and 2010, just using information that are available in advance of a proceeding. What is more, I explicitly tease out the predictive contribution of legal and political context in the forecasting framework. While legal context itself already provides a reasonable baseline for a forecast, I demonstrate that the predictive performance is considerably improved when the political context of a proceeding is considered, too. This chapter not only supports the view of a multifaceted decision-making of constitutional courts which is best characterized by the ensemble of legal and political context. It also demonstrates the value of predictive modeling for the field social science as a fruitful complement to traditional causal inference approaches.

CHAPTER 2

Constitutional Courts as Opinion Leaders: Evidence From a Comparative Survey Experiment

2.1 Introduction

Can constitutional court decisions shape public opinion on a governmental policy? In this chapter, I explore the relationship between the court and the public by examining the extent to which courts can influence public opinion. Put simply, I argue that the public support for courts allows them to move public opinion on public policies into the direction of their rulings. However, and this is the novelty presented in this chapter, I show that this only holds for courts that possess sufficient institutional legitimacy. Therefore, this chapter tests the theoretical argument raised in the introduction whether varying degrees of public support lead to systematic differences in the judicial-public relationship.

Legitimacy is perceived to be the major source of power of constitutional courts as they have no formal means to ensure compliance with their decisions. As such, legitimacy has been the subject of decades of scholarly attention. Most of the work focuses on the question of whether the court's legitimacy suffers when it releases unpopular decisions (e.g. Caldeira and Gibson, 1992; Gibson and Caldeira, 2009; Bartels and Johnston, 2013; Gibson and Nelson, 2015). Another strand asks whether courts can draw on their institutional legitimacy to move public opinion on public policies in the direction of their ruling. This is called the "*legitimacy-conferring capacity*" of courts. The evidence, mostly from the US Supreme Court, of such a legitimacy-conferring capacity of courts is mixed. Most observational studies (e.g. Marshall, 1987; Stoutenborough, Haider-Markel and Allen, 2006) find no sign for such a legitimating power of courts.

By contrast, experimental studies (Hoekstra, 1995; Bartels and Mutz, 2009) find that the Supreme Court is able to move public opinion in the direction of the public policy that it endorses.

In order to advance our understanding of European constitutional courts and their importance as a "legitimizer" of public policy, I improve existing work in two directions. First, previous studies exclusively focus on the US Supreme Court, whereas not a single study analyzes the legitimacy-conferring capacity of courts such as European constitutional courts. Therefore, it is unclear whether the (mixed) findings from the US Supreme Court can be generalized to other court types such as European constitutional courts. Second, prior work constantly assumes that courts belong among the most trusted branches of the government. However, this does not hold empirically given the varying degrees of public support for national high courts worldwide (Gibson, Caldeira and Baird, 1998). Consequently, the legitimacy-conferring capacity of courts also varies: very popular courts are expected to have a higher legitimacy-conferring capacity than unpopular courts. Despite the simplicity of this argument, it has never previously been tested in a comparative scenario.

I put this theory to the test by comparing the legitimacy-conferring capacity of two European constitutional courts, namely the French Conseil Constitutionnel (a rather unpopular court) and the German Federal Constitutional Court (a prime example of a very popular court). Using several survey experiments in a unique, cross-institutional comparative design embedded in large, representative surveys in both countries, I find that the German court can move public opinion into the direction of its decision by placing its stamp of approval or disapproval on public policies. This effect is sufficiently strong to even shape the opinions of those who have strong pre-existing attitudes towards the issue. I attribute this to the broad institutional support for the German court. By contrast, I find no such legitimacy-conferring capacity for the French Conseil. The findings of this chapter thus have implications for both our understanding of the role of constitutional courts in democratic politics and for public opinion formation in general.

2.2 Court Decisions, Governmental Policy and Legitimacy

Legitimacy is perceived to be the major source of power of courts. While much of the existing research on court legitimacy focuses on whether citizens agree or disagree with court decisions and what this implies for the court's popularity and reputation, it is also important to consider the reversed causal direction: what effects do judicial opinions have on public opinion?

Courts are often among the most popular (political) institutions in Western democracies, and they are generally perceived as highly legitimate (Caldeira and Gibson, 1992; Gibson, Caldeira and Baird, 1998; Gibson and Nelson, 2014, 2015). One of the

consequences of this property is the ability of courts to pass their legitimacy to public policies. This argument dates back to Dahl (1957), who argues that the US Supreme Court is a "legitimizer" of majority coalition's policies. Dahl (1957) argues that this power stems from the Supreme Courts function as the sole legitimate interpreter and protector of the constitution (Dahl, 1957, 293). Supreme Court decisions are, therefore, viewed as credible, legitimate and rightful. This phenomenon is called the "*legitimacy-conferring capacity*" of courts. In other words, courts are able to use their diffuse support, or their "reservoir of goodwill"(Easton, 1965), to induce public (non)-acceptance of governmental policies via their rulings. This mechanism of changing public opinion in the direction of an institution's endorsement is generally known as an "*endorsement effect*". Following Zaller (1992), an endorsement effect is defined as an increase or decrease in support for a policy that occurs when people learn that a trusted source supports or does not support the policy (Zaller, 1992, 33). Such an endorsement effect may manifest either in a "whole" or "soft" opinion change. A "whole" opinion change occurs when people either change from opposing to favoring a policy, or vice versa, whereas a "soft" opinion change is only a shift in the degrees of favoring or opposing.

Extensive empirical literature exists on the legitimacy-conferring capacity of the US Supreme Court and accompanying endorsement effects using different measures and methods, overall with mixed evidence. Unfortunately, the findings of these studies depend, at least to some extent, on the nature of the research design. Experimental studies tend to find relative consistent endorsement effects of Supreme Court decisions (Hoekstra, 1995; Clawson and Kegler, 2001; Bartels and Mutz, 2009). For instance, Bartels and Mutz (2009) use a survey experiment to compare the Supreme Court and the US Congress's ability to move opinion and find that the court is more influential than the congress in using its institutional credibility to shape mass opinion. By contrast, observational studies mostly find no evidence of a legitimacy-conferring capacity of the Supreme Court (Marshall, 1987; Stoutenborough, Haider-Markel and Allen, 2006; Hanley, 2008), although sometimes decisions can polarize public opinion (Hoekstra and Segal, 1996; Johnson and Martin, 1998; Brickman and Peterson, 2006). To further complicate matters, other observational studies find that Supreme Court decisions only induce changes in public opinion under certain conditions, such as salient decisions or salient issues (Christenson and Glick, 2015; Tankard and Paluck, 2017; Christenson and Glick, 2018; Zilis, 2014), media coverage (Linos and Twist, 2016) or the legitimacy of lower courts (Gibson and Nelson, 2018).

When looking at other court types such as European "Kelsenian" constitutional courts, we must recognize that little to nothing is known about the interplay between court legitimacy and public opinion with respect to these courts. If at all, scholars have examined the role of public support for constitutional court decision-making in a separation-of-powers framework (Vanberg, 2001, 2005; Sieberer, 2006; Staton and Vanberg, 2008; Brouard, 2009; Sternberg et al., 2015). In these studies, the authors examine

whether constitutional courts have to adjust their decision-making in accordance with public opinion. They do not investigate the effect of court decisions on public opinion. Only a few studies explicitly investigate court legitimacy and diffuse support in the European court context. However, most of these studies are rather descriptive. For instance, with respect to the GFCC, Vorländer and Schaal (2002) find that the diffuse support of the GFCC remains constantly high, independent of individual decisions that might have been against the public will. The outstanding diffuse support of the GFCC, its strong legitimacy and the resulting role of the GFCC as an "interpretative authority" of the constitution is also documented in other studies (Vorländer and Brodocz, 2006), albeit mostly using only anecdotal evidence (Vorländer, 2006). No existing work has directly tested the legitimacy-conferring capacity of any European constitutional court.

In order to advance our understanding of European constitutional courts and their importance as a "legitimizer" of public policy, I improve existing work in two directions. First, to the best of my knowledge, no study exists that analyzes the effect of court decisions on public opinion outside the US. Therefore, it is unclear whether the (mixed) findings from the US Supreme Court can be generalized to other court types such as European constitutional courts. Second, previous studies work under the assumption that courts enjoy consistently high public support. While this might be true, at least to some extent, for the Supreme Court (Gibson and Nelson, 2014, 2015), it does not hold empirically given the varying degrees of public support for national high courts worldwide (Gibson, Caldeira and Baird, 1998). In this chapter, I therefore approach the question of whether courts can move public opinion from a different angle than previous studies. The novelty lies in the fact that I do not argue for different conditions under which changes in public opinion could be induced by court decisions, but instead I look at a "counter-factual" scenario: what happens to public opinion if a court with high diffuse support decides on public policy compared with a court with low diffuse support deciding on the same issue?

My central theoretical claim is the same as in previous studies: constitutional courts receive their legitimacy-conferring capacity from their perception as the only legitimate and credible interpreter of the constitution. This legitimacy is therefore grounded in their diffuse support, and it allows them to move public opinion into their direction by institutional endorsement. However, I argue that this does not hold ultimately for all constitutional courts. As Gibson, Caldeira and Baird (1998) note, "national high courts vary enormously in the degree to which they have achieved institutional legitimacy" (Gibson, Caldeira and Baird, 1998, 356). These varying degrees of public support (for an empirical overview, see Gibson, Caldeira and Baird, 1998) should also be considered theoretically.

For instance, the German court enjoys consistently high public confidence and its public support often exceeds that of other political institutions (Vanberg, 2005; Vorländer and Brodocz, 2006). By contrast, there are constitutional courts such as the one of

Russia or Bulgaria that possesses much lower levels of public support (see Gibson, Caldeira and Baird (1998), Trochev (2008), and Staton and Vanberg (2008)). The reasons for such a low institutional legitimacy are manifold. Some courts, for instance those newly installed during the third wave of democratization, have had little opportunity to build institutional trust compared to long-established courts such as the US Supreme Court or courts installed after the Second World War (Gibson, Caldeira and Baird, 1998, 345). The legitimacy of other courts, such as the French Constitutional Court, suffers from a purely politicized appointment process (Hönnige, 2009) and the fact that it received the power of protecting citizens' rights rather lately (Brouard, 2009).

If the central argument that the legitimacy-conferring capacity of courts roots in their legitimacy is correct, then unpopular courts are expected to have a much lower legitimacy-conferring capacity than popular courts. Therefore, we should be able to observe endorsement effects, namely people moving in the direction of the institutionalist endorsement in the context of constitutional courts with high diffuse support, but not in the context of constitutional courts with low public support. The observable implication for a popular court is then as follows:

Observable Implication 1: *If a constitutional court is popular, then we should observe an endorsement effect of court decisions, whereby public opinion concerning a governmental policy should move in the direction of a popular court's ruling.*

Given that it is the institutional legitimacy of courts that gives them their legitimacy-conferring capacity, there should be no endorsement effect observable in systems with a court with low diffuse support. The observable implication for such an unpopular court is then as follows:

Observable Implication 2: *If a constitutional court is unpopular, then we should not observe an endorsement effect of court decisions, whereby public opinion concerning a governmental policy should not move in the direction of a popular court's ruling.*

Im summary, I improve previous research by considering that not all constitutional courts possess the same high level of diffuse support. Instead, I argue that the varying degrees of diffuse support for constitutional courts affect their legitimacy-conferring capacity, such that courts with high public support are able to change public opinion while unpopular courts are not.

2.3 Research Design

In this section, I introduce an experimental research design that allows me to test the two competing observable implications in a comparative setting.

2.3.1 Case Selection

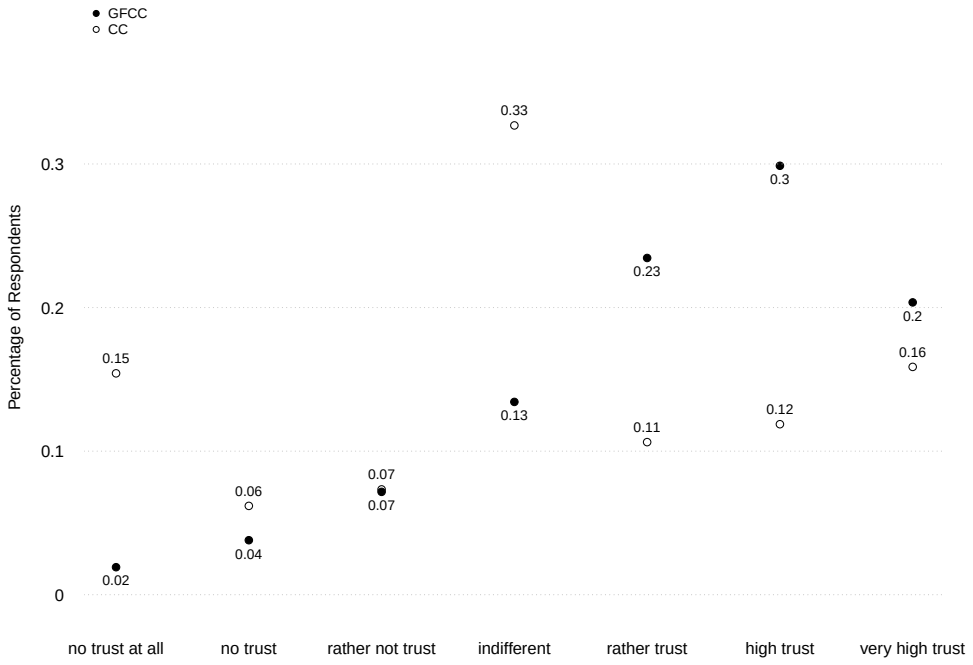
The case selection of the two constitutional courts in this comparative study is motivated by the most different system design (Przeworski and Teune, 1970, 34). The previous section outlined that the legitimacy-conferring capacity of constitutional courts depends on their diffuse support. The implication is that a court with high diffuse support should have a larger legitimating capacity than a court that is viewed rather negatively by the public. Therefore, in order to test these two observable implications, the two courts have to meet the following requirements. First, both constitutional courts must possess the right of judicial review, because otherwise courts cannot decide about governmental policies, and the logic of the legitimacy-conferring capacity could not be applied. Second, and more important, the theoretical argument requires two courts that considerably vary in their degrees of public support.

For my comparative design I chose the German Federal Constitutional Court (*GFCC*) and the French Conseil Constitutionnel (*CC*). Both courts have the right to exercise judicial review (Hönnige, 2009; Brouard and Hönnige, 2017), and thus meet the first condition. Moreover, the GFCC is a prime example of a constitutional court which enjoys high public confidence. The GFCC plays an important role in the German political system, and repeatedly finds itself in the middle of controversial political conflicts. Its prominence and reputation as the "guardian of the constitution" (Rudzio, 2006, 282) has made it one of the most recognized and respected institutions in German politics. Indeed, the German court enjoys extremely high support vis-à-vis other institutions, and its public support consistently exceeds that of the other major German political institutions (Vanberg, 2005; Vorländer and Brodocz, 2006).

By contrast, the French CC is one of the few constitutional courts in Western Europe which cannot rely on broad public support as other courts can do. This is mainly due to three historic reasons. First, the preeminence of parliamentary sovereignty in France forbade judicial review in the country for a long period (Stone, 1992). It was only 1959 when the CC was effectively set up as the last institution to be created, which symbolizes its second order status for the designers of the Fifth Republic (Brouard, 2009). Second, the CC was initially thought to limit parliamentary power, not to protect individual rights.[1] This only changed rather lately. As until 2010, any type of constitutional complaint was not allowed, and therefore citizens were kept at bay from

[1]Charles Eisenmann, a contemporary law scholar who was very critical of the 1958 constitution, said in a famous quote that the CC is thought as a "canon pointed on the parliament". This substantially changed after the enlargement of the initiation process of constitutional review to Members of the Parliament, so that the CC had a significant impact on the regulation of the French political system (Brouard and Hönnige, 2017).

Figure 2.1 – Comparison of the Institutional Trust in the Constitutional Courts of Germany and France



Note: Comparison of Institutional Trust in the Constitutional Courts in Germany and France. Data from GIP Wave 26 and ENEF Wave 15.

judicial review. Last but not least, the Conseil's alleged politicization induced by the appointment process has always hampered its legitimacy(Brouard, 2008).[2]

Recent numbers confirm this difference in public support. Figure 2.1 shows the percentage of the mean trust rating of respondents of representative surveys in both Germany and France on the Y-axis over the range of institutional trust[3] on the X-axis. It is evident that the GFCC enjoys a much higher public support than the CC: the percentage of respondents at the higher trust levels in Germany is always above the percentage of respondents in France, and vice versa. In Germany, for instance, every second respondent (50%) has high or very high trust in the GFCC, while only 28% have the same trust in the CC. Moreover, while in France more than every fifth (22%) respondent has no or no trust at all in the CC, only 6% have the same low levels of

[2]The political appointment process of French constitutional court judges is unique in Europe: "The only major country without any restrictions to a purely politicized appointment process is France" (Tsebelis, 2002).

[3]Respondents were asked about their perceived trust in the institutions. See Appendix A.1 for the exact wording of the trust questions in both surveys and supporting information.

trust in Germany. These numbers also correspond to the findings of previous studies (Gibson, Caldeira and Baird, 1998; Vanberg, 2005; Vorländer and Brodocz, 2006).

Moreover, the GFCC is also perceived as significantly more trustworthy compared with other institutions such as the government (GFCC mean 5.24 compared with government mean 3.47, $p < .01$). The level of trust in the CC is also significantly higher compared to other institutions such as the parliament, but this difference is comparably smaller (CC mean 4.16 compared with parliament mean 3.68, $p < 0.01$).

2.3.2 Experimental Design

In order to investigate whether constitutional courts can move mass public opinion, I use an experimental design embedded in two national, representative surveys in Germany and France. The exact same experiment was implemented in both countries. In Germany, the experiment was implemented as part of Wave 26 (November 2016) and Wave 27 (January 2017) of the *German Internet Panel* (hereafter GIP). The GIP collects information on political attitudes and preferences of respondents through bimonthly longitudinal online panel surveys. Although administered online, all surveys are based on a random probability sample of face-to-face recruited households from the German population, which were provided with access to Internet and special computers if necessary (Blom, Gathmann and Krieger, 2015). Wave 26 and Wave 27 include N = 2,867 registered participants[4] and are representative of both the online and offline population aged 16-75 in Germany.

In France, the experiment was embedded in Wave 16 (July 2017) and Wave 17 (November 2017) of the *French National Election Study* 2017 (l'enquête électorale française, hereafter ENEF). ENEF 2017 was a panel survey conducted online by IPSOS. As with almost all surveys in France, sampling is conducted with a quota method based on age, gender, occupation, region and type of residential area.[5] The experiments were allocated to a random sub-sample of $N = 2,661$ respondents in Wave 16.

In the experimental design, I compare the legitimacy-conferring capacity of the GFCC and the CC with other political actors. I employ a survey priming experiment where respondents are provided with a (hypothetical) public policy issue. The proposed policy in the experiment is a "school security law". According to this (hypothetical) school security law, private security companies can equip armed security forces that are allowed to search students and their lockers. The measure is said to be justified to prevent the increasing school violence. This issue was chosen for two reasons. First, it is an issue that could credibly be addressed by the constitutional courts and other political

[4]For both surveys, respondents who refused to answer or had no opinions on the relevant issues were eliminated from the sample. These were 9.1% in the GIP and 3.6% in the ENEF.

[5]See `https://www.enef.fr/données-et-résultats/` for more information (accessed 02.05.2019).

institutions. Second, it is an issue where enough polarization across the respondents can be expected. This ensures sufficient variation in the outcome variable.

Along with the issue, respondents were randomly assigned to an institutional endorsement manipulation. This manipulation occurred in the form of different political institutions/actors either stating that they *approve* or *disapprove* the school security law. Overall, I used three different institutions/actors in Germany and two in France. Accordingly, I am able to compare the legitimacy-conferring capacity of these institutions vis-à-vis the constitutional courts. In France, I used the CC and a high organizational school committee called the "Haut Conseil de l'Éducation". In Germany, I employed, in line with the French survey, the GFCC and the German equivalent to the school committee called the "Conference of the Ministers of Education". Both committees have consultative character regarding essential knowledge and educational questions, and could thus credibly express their opinion on a school security law as presented in the experimental issue. Additionally, the German survey experiment additionally used the Federal Commissioner for Data Protection and Freedom of Information (in short, the Data Security Official) in the second wave. A control group received no endorsement manipulation at all. Overall, the survey experiments thus contains seven experimental conditions in Germany (3 sources [GFCC, Conference of the Ministers of Education, and Data Security Official] $\times$ 2 arguments [approves or disapproves] + 1 control group $= 7$) and five experimental conditions in the French survey experiment (2 sources [CC,Haut Conseil de l'Éducation] $\times$ 2 arguments [approves or disapproves] + 1 control group $= 5$). Appendix A.2 includes further details about the wording of the endorsement manipulations and measurement of variables. After the experimental manipulation, respondents were asked to give their opinion on such a school security law on a five-point scale, ranging from "fully disagree" to "fully agree".

The chosen experimental design provides several methodological strengths for assessing the legitimacy-conferring capacity of the different institutions. First, because it contains a true control group which did not receive any institutional endorsement manipulation, I have a reasonable baseline for comparison. Any systematic shift in opinion away from the control group can be attributed to the legitimacy-conferring capacity of either institutional source. This is an advantage compared with some of the existing experimental studies (e.g. Hoekstra, 1995; Clawson and Kegler, 2001), which did not include a control group. Second, the data quality and size of the survey experiments are superior to other experimental studies, which mainly rely on laboratory studies involving student samples (Baas and Thomas, 1984; Hoekstra, 1995). Having the same experimental design administered in national representative mass surveys in two countries increases the external validity of my study, while internal validity remains high. Moreover, the panel design of both surveys in Germany and France enables me to add additional variables ex post if required; for instance about certain attitudes or issue preferences asked in earlier or later waves.

2.4 Results of the Survey Experiments

Due to the categorical, ordered nature of the dependent variable, I use an ordered probit model to analyze the experimental data.[6] In order to ease the result presentation and further analyses, the originally five-point scale dependent variable was recoded into a three-point scale variable with three ordered categories, namely *disagree, indifferent and agree*.[7] For the German analysis, I aggregate both GIP waves into one data set to increase computational efficiency. Because the respondent's answers are then no longer independent, the standard errors are clustered by respondents in the German analysis. In the robustness section later, a variety of alternative models is estimated to demonstrate that the results also hold when the original five-point scale dependent variable is used or when the German data is not aggregated. A detailed outline of the estimation strategy is available in Appendix A.4. In Appendix A.3, I also give a descriptive overview about the distribution of attitudes towards the school security law across German and French respondents. The main differences between the two countries is that the majority of the German respondents opposes the proposed school security law (57%), whereas the majority of French respondents supports it (48%).

For both countries, I estimate a simple ordered probit model with the three-point scale ordered respondent's opinion on the school security law as the dependent variable and each experimental treatment group as a dummy independent variable. The control group is used as the reference category. The simplicity of the model derives from the experimental design with the random assignment of the respondents to the treatment groups.

Figure 2.2 reports the ordered probit results from both countries. The respective regression tables are in Appendix A.5. For Germany, there is a statistically significant effect[8] of both GFCC endorsements. This means that compared with the control group, the GFCC approving or disapproving the school security law leads to a positive or negative, statistically significant change in public opinion. This is exactly what *Observable Implication 1* predicts: due to its reputation as a credible and legitimate interpreter of the constitution, the GFCC is able to confer legitimacy by placing its stamp of approval or disapproval on the governmental policy.

The German national expert bodies do not have the same legitimacy-conferring capacity as the GFCC. If the Data Security Official approves or disapproves the governmental policy, we observe a shift of public opinion in the corresponding direction, but these

[6] Ordered probit models require the proportional odds assumption. A likelihood-ratio test of whether the coefficients are equal across categories shows that this assumption is not violated.

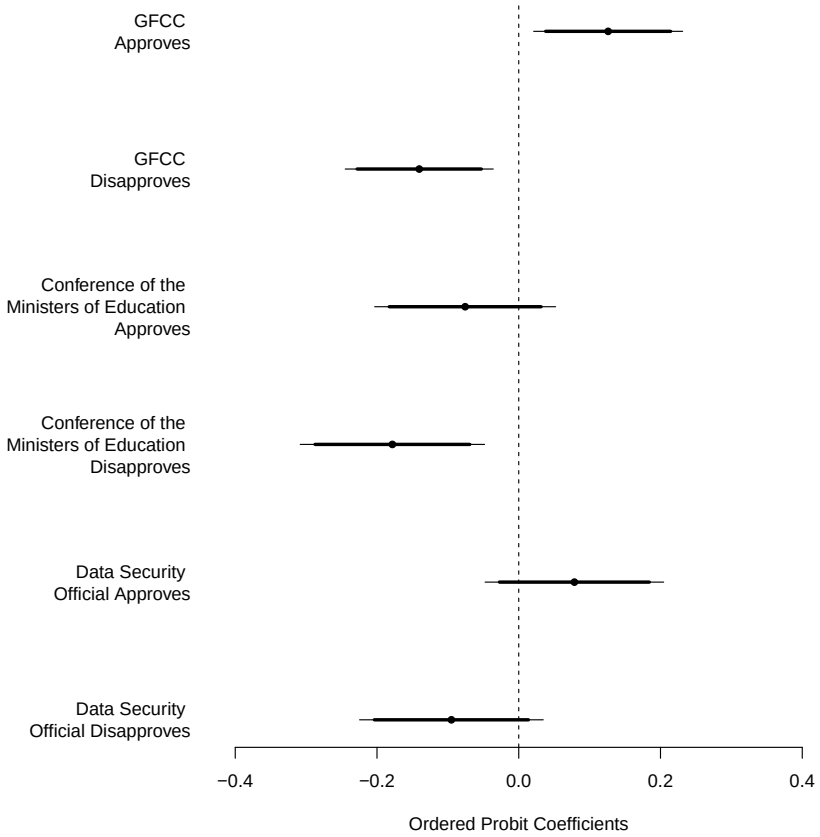
[7] The two highest (agree and fully agree) and the two lowest (fully disagree and agree) categories are summarized into agree and disagree, respectively.

[8] Statistically significant indicates a statistically significant effect on the conventional levels of statistical significance ($p < 0.05$, two-sided test) through the following section, if not stated otherwise.

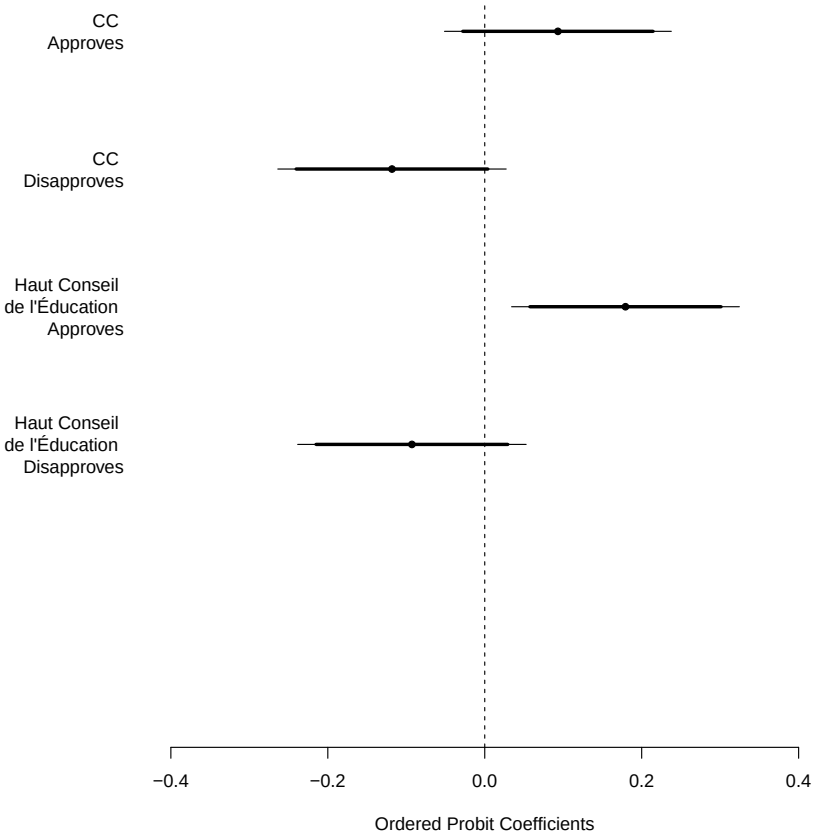
effects are not statistically significant. With respect to the Conference of the Ministers of Education, we see that both coefficients are negative, indicating that independent of whether the Minister of Education approves or disapproves the school security law, respondents dislike this policy. However, only the coefficient for the Minister of Education disapproving the law is statistically significant. In summary, the survey experiment in Germany shows that an endorsement by the GFCC indeed leads public opinion to move in the corresponding direction.

In the analysis of the French experiment, both CC treatment coefficients show the expected direction (a positive effect for the CC approving and a negative effect for the CC disapproving the governmental policy), but both coefficients are not statistically significant ($p > 0.10$). This means that in contrast to its German counter-part, the French constitutional court does not possess about the same legitimacy-conferring capacity. Therefore, an endorsement of the CC does **not** lead to a change in public opinion in either direction. Again, this is exactly what *Observable Implication 2* predicts: if a constitutional court is unpopular (such as the CC), then we should not observe public opinion moving into the direction of this unpopular court's ruling, and thus there are no endorsement effects. With respect to the Haut Conseil de l'Éducation, there is a statistically significant positive effect for the Haut Conseil approving the school security law, but no significant effect for the Haut Conseil disapproving a law. This is an interesting finding, as it suggests that the Haut Conseil at least partially occupies a larger legitimacy-conferring capacity than the CC. In summary, the survey experiment in France shows that the French constitutional court does not possess the same legitimacy-conferring capacity as the GFCC.

Figure 2.2 – Ordered Probit Regression Results of Survey Experiments in Germany and France



(a) Results for Germany



(b) Results for France

Note: This figure shows the estimates of the ordered probit regression for the survey experiment in both Germany and France. The points represent the ordered probit point estimates and the thin and thick bars represent 95% and 90% confidence intervals. The intercepts (cut points) are omitted. Standard errors for coefficients are clustered by respondents in the German analysis. See Appendix A.5 for the corresponding regression tables.

In order to evaluate the substantive relevance of the results, I calculate quantities of interests using simulations[9] (King, Tomz and Wittenberg, 2000). This allows me to provide substantial interpretations of the effect magnitudes.[10] I only present results for the German analysis. The simulations using the French data confirmed that there is no statistically significant legitimacy-conferring effect of the CC. Because the presentation of simulation outcomes of ordered probit models is not as straight forward as it is for standard probit models, I use so-called "parallel coordinate plots" to visualize the simulation results. Parallel coordinate plots allow me to visualize the probabilities of all three response categories simultaneously. This is important, because in ordered probit models the probabilities of each outcome are not independent. An alternative visualization using more commonly employed "ternary plots" (King, Tomz and Wittenberg, 2000, 358) is available in Appendix A.6.

The parallel coordinate plot on the left side of Figure 2.3 shows 1,000 simulated expected values for a respondent in the control group (black lines) and a respondent who received the endorsement that the GFCC approves the school security law (grey lines). The expected values are predicted probabilities across the three outcome categories (which sum to one). Each line in the plot corresponds to one draw of the simulation, whereby the spread of these lines represents the uncertainty of the estimates. For the simulations, only the respective experimental dummy (control group, GFCC approve treatment) is set to one and all other dummies of the model are set to zero.

The simulation results show that even if respondents receive the treatment that the GFCC approves the school security law, the probability of disagreeing remains relatively high (because the black and gray lines are still closely located on the upper end of the scale). However, we can also observe an decrease in the probability of disagreeing, as the gray lines are located below the black control group lines for the disagree category and the other way around for the agree category. To illustrate, respondents in the control group have, on average, a 57% predicted probability of disagreeing with the school security law, a 9% probability of being indifferent and a 34% probability of agreeing. By contrast, respondents who received the GFCC approves endorsement have on average a probability of 52% of disagreeing, 9% of being indifferent and 39% of agreeing.

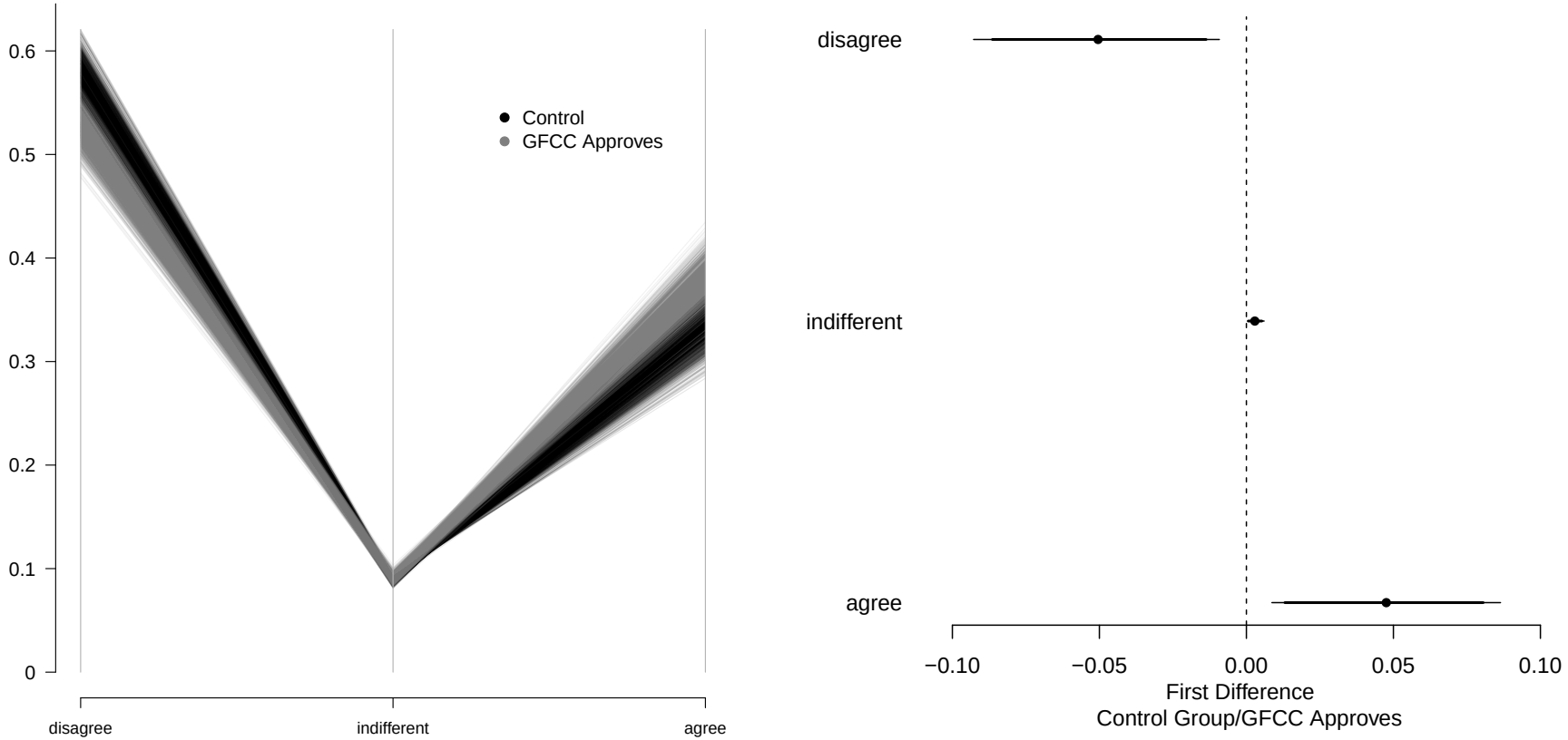
In order to better understand this relationship, I plot the corresponding first difference between the predicted probabilities of the control group and the GFCC approves endorsement group on the right side of Figure 2.3. The first difference shows that

[9]Simulations for France were carried out using the *Zelig* ordered probit implementation (Venables and Ripley, 2011). For Germany, an own implementation was used because to date, *Zelig* does not include an option for clustered standard errors.

[10]Note that using the experimental data, you obtain the same results independent of using the "observed value" or "average case" approach. This is because for the simulations, just the experimental dummy variables are varied (set to either zero or one). There is no need to fix other covariates at their mean or another arbitrary value.

when switching from the control group to the group where respondents received the treatment that the GFCC approves the law, the probability of agreeing with the school security law increases by about 5 ($\pm$.2) percentage points on average, while the probability of disagreeing decreases by about 5 ($\pm$.2) percentage points. It also becomes more likely to be indifferent. All first differences of these effects are significantly different from zero at the 95% level.

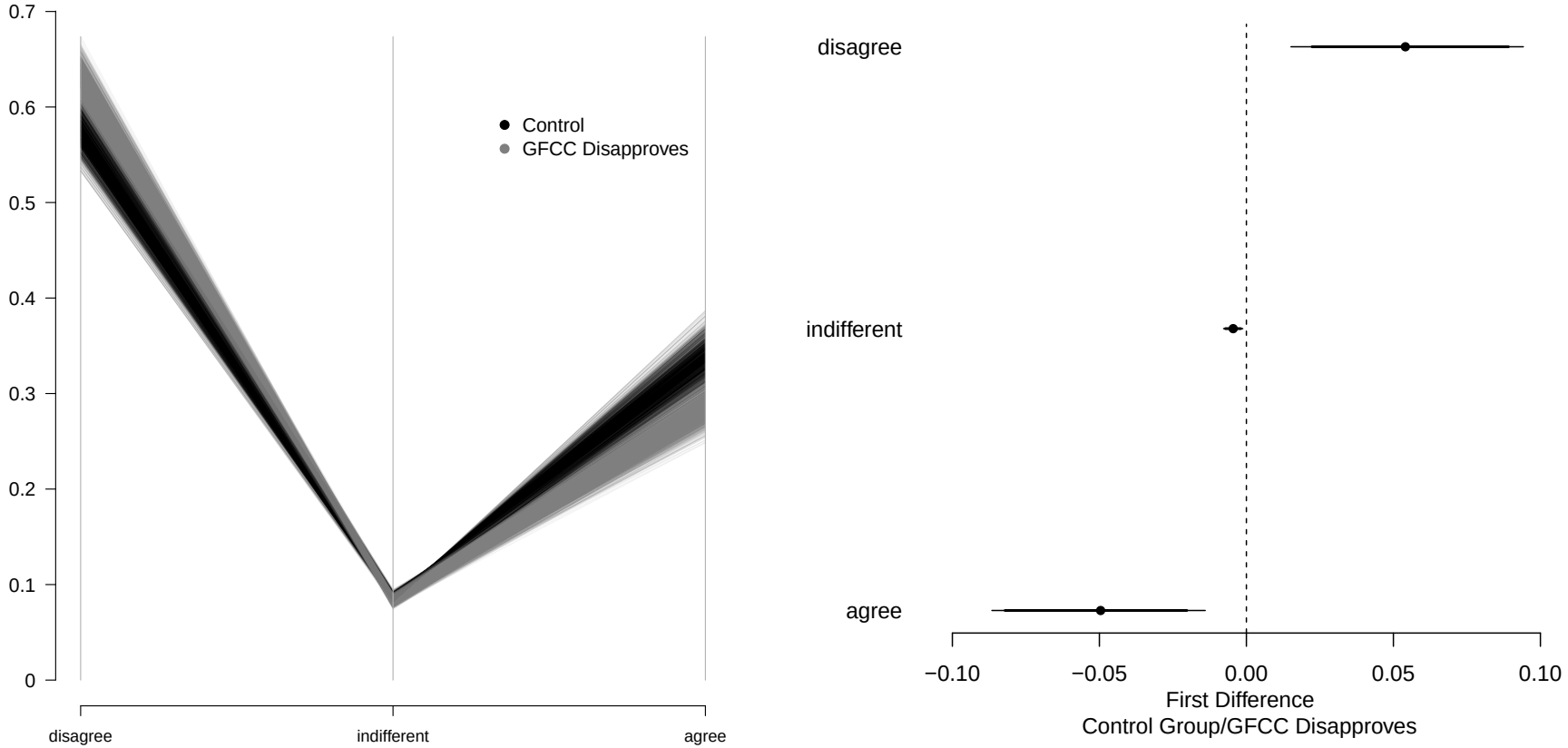
Figure 2.3 – Predicted Probabilities and First Difference of Control Group and GFCC Approves Endorsement



Note: Parallel coordinate plot of the simulated predicted probabilities for the school security law. N of simulation $= 1,000$. The probabilities are calculated by setting the respective dummy (control group and GFCC approves) of the model to one and all other dummies to zero. Each line represents one draw of the simulations. Simulations are based on the ordered probit model of Table A.1 in Appendix A.5.
Right side: First differences between the predicted probabilities of the control group and the GFCC approves treatment group from the same simulation. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals.

Figure 2.4 plots the simulation results for the difference between the control group (black lines) and the group which received the GFCC disapproves endorsement (gray lines). Looking at the parallel coordinate plot on the left side of the figure, we see that the predicted probabilities of the simulated outcomes for respondents who received the GFCC disapproves endorsement are located above the outcomes of the control group for the disagree category and below for the agree category. This means that respondents who received the GFCC disapproves endorsement have a higher predicted probability of disagreeing with the school security law than respondents of the control group. In fact, the probability of disagreeing of respondents in the treatment group is 63%, compared with 57% in the control group. The right side of Figure 2.4 shows that this difference is also statistically significant at the 95% level. Switching from the control group to the group where respondents received the treatment that the GFCC disapproves the law increases the probability of disagreeing by a about 5 ($\pm$ .1) percentage points on average, while the probability of agreeing decreases by about 5 $\pm$.2) percentage points.

Figure 2.4 – Predicted Probabilities and First Differences of Control and GFCC Disapproves Treatment



Note: Left Side: Parallel coordinate plot of the simulated predicted probabilities for the school security law. N of simulation $= 1,000$. Expected values are calculated by setting the respective dummy (control group and GFCC disapproves) of the model to one and all other dummies to zero. Each line represents one draw of the simulations. Simulations based on the ordered probit model of Table A.1 in Appendix A.5.
Right side: First differences between the predicted probabilities of the control group and the GFCC approves treatment group from the same simulation. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals.

Are these substantial changes, given that the GFCC endorsements "only" lead to a five percentage points change in the predicted probability of either agreeing or disagreeing with the school security law? I argue that there are nonetheless at least three reasons to consider this a substantial effect. First, although the absolute change seems small, it should be considered that this is the aggregate change in opinion. Previous studies demonstrate that it is difficult to detect an aggregate change in opinion, because issue polarization can lead different sub-groups of the sample to move in different directions. This, in turn, can cancel out observable opinion change effects on the aggregate-level (e.g. see Christenson and Glick, 2015). Second, most of the German respondents rather disagree with the school security law. This makes it a hard-case scenario to test the legitimacy-conferring capacity of the GFCC, and the found endorsement effects are rather conservative estimates. Finally, my experimental manipulations only comprise one added sentence to the described case context. Other studies might find larger effects, but also employ more profound endorsement manipulations. Some authors provide, for instance, detailed court reasonings and arguments against or in favor of the governmental policies (Hoekstra, 1995; Bartels and Mutz, 2009), while others offer news storylines that include counter-narratives to the court's ruling (Zilis, 2014). Moreover, the observed "soft" change in public opinion corresponds with the findings of previous studies. Using a similar survey experiment, Bartels and Mutz (2009) e.g. find that "the institutional endorsements generally soften respondents' existing positions in the direction of the institution's decision, but do not often wholly change them from opposition to advocacy" (Bartels and Mutz, 2009, 259).

2.4.1 Pre-existing attitudes and the Court's Legitimacy-conferring Capacity

The previous analyses show that the GFCC as a popular constitutional court is capable of moving public opinion in the direction of its decision at the aggregate level due to its legitimacy-conferring capacity, while the CC is unable to do so. However, do these effects also hold at the individual level, and how strong are they? In this section, I therefore seek to investigate whether constitutional courts have the power to overcome preexisting attitudes and induce opinion change even among those who are initially either strongly in favor or against the governmental policy.

Individuals do not form their opinion in a vacuum. Instead, they have pre-existing attitudes which might lead them to support – or not support – a policy's goal. It is widely accepted that people develop their policy preferences using their party identification as a heuristic (Downs, 1957; Campbell et al., 1960; Zaller, 1992; Bartels, 2002). In the early work of Campbell et al. (1960), for instance, they note that "identification with a party raises a perceptual screen through which the individual tends to see what is favorable to his partisan orientation" (Campbell et al., 1960, 133). Later work in this area confirms that partisan attitudes are powerful cues for citizens to evaluate which

candidate or party they should support (Lodge and Hamill, 1986; Rahn, 1995; Bartels, 2002).

I use the existence of these pre-existing attitudes to investigate whether the GFCC's legitimacy-conferring capacity is sufficiently strong to even change the opinion of those who hold strong prior attitudes with respect to the school security law. I expect that the very popular German Federal Constitutional Court is able to confer legitimacy to the school security law, namely to induce at least a "soft" opinion change among those who either strongly support or oppose the governmental policy. In the same line, I expect that the French CC, a court that is viewed rather negatively by the public, should not be able to induce such an opinion change.

I approximate the pre-existing attitudes of the respondents towards the school security law via their party affiliation.[11] Figure 2.5 plots the distribution of the respondents opinion on the school security law over different party affiliations in Germany and France. Only the respondents in the control group are analyzed, so that their opinion is "honest" and not manipulated by the endorsements. When looking at the German respondents, we observe that partisans of the right-wing party *Alternative for Germany* (AfD) are supportive towards the proposed school security law, while, for instance, members of the *Greens* strongly oppose such a law. Therefore, I use members of the AfD and members of the Greens to test the legitimacy-conferring capacity of the GFCC at the individual level. If the GFCC is truly perceived as a highly legitimate institution, then the endorsement of the GFCC disapproving the law should shift AfD-partisans (who initially support, namely agree with such a law) towards more disagreement, and vice versa for the Greens. This would be additional evidence in favor of the outstanding legitimating power of the GFCC.

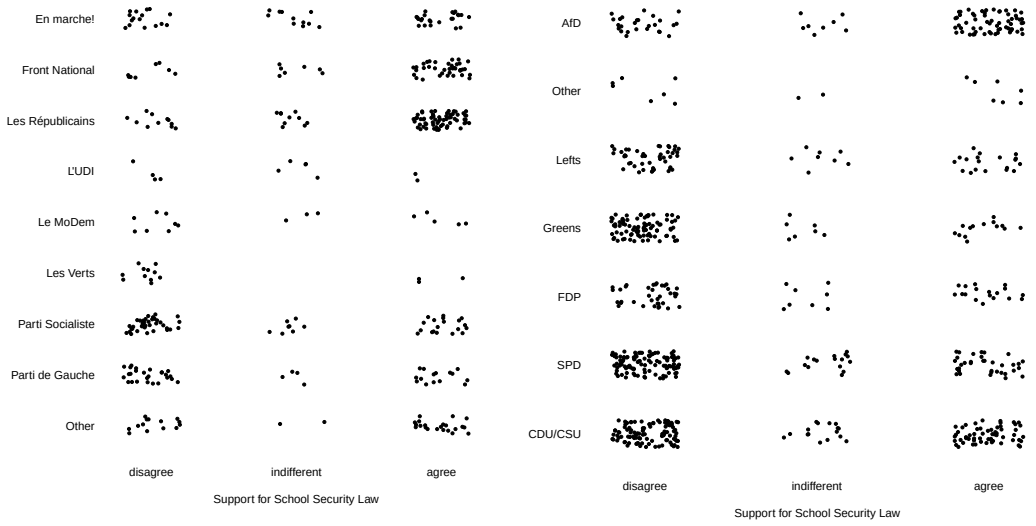
When looking at the French respondents, we observe that partisans of the right-wing *Front National*[12] (National Front) and the *Republicans* (Les Républicains) exhibit strong support for the school security law, while, for instance, members of the *Socialist Party* (Parti Socialiste) oppose such a policy. However, in contrast to the GFCC, I do not expect that an endorsement by the CC leads these party members to shift their opinion in direction of the CC's endorsement. This is due to the low public support and institutional legitimacy of the CC.

In order to test these implications, I run additional ordered probit models where I include an interaction between the experimental groups and the party affiliation. Four separate models are run: one for AfD-partisans and one for Green-partisans using

[11] The party affiliation of respondents is measured via an opinion poll on their voting preferences. For German respondents, this information is included in the core study of the GIP in Wave 25 (September 2016). The information about the party affiliation of the French respondents is taken from Wave 15 of the ENEF (June 2017).

[12] The party changed its name to National Rally (Rassemblement National) in June 2018. The time the French survey was conducted it was still named Front National, which is why I stick to this name in the text.

Figure 2.5 – Support for the School Security Law over Party Affiliations in Germany and France



(a) France *(b) Germany*

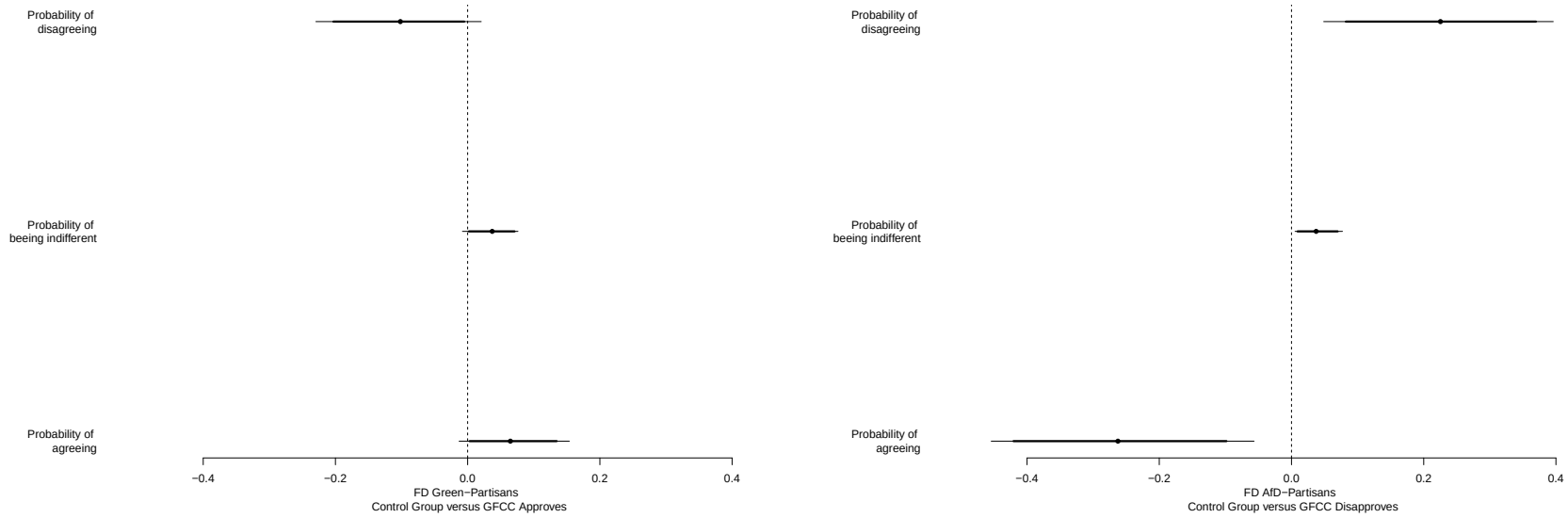
Note: This figure shows the distribution of the support for the school security law over party identification in Germany and France. Each point represents a single respondent (in each cell, little random noise is added to improve the visualization). $N = 1,039$ in Germany, $N = 521$ in France. The evaluation is based on the control groups to avoid that institutional endorsement effects bias the respondent's opinion. Jitter inside each cell is randomly added for a better visualization.

the German sample, and one for Socialist Party partisans and one for National Rally partisans using the French data. Party affiliation is incorporated via a dummy which indicates whether someone is affiliated to the corresponding party (= 1) or not (= 0). Appendix A.7 provides the corresponding regression tables.

Simulations are used again to provide a substantial interpretation of the results. The left side of Figure 2.6 plots the first differences of the predicted probabilities for Green partisans in the control group and Green partisans who received the GFCC approves endorsement.[13] I find that Greens have a 20 percentage points higher probability to disagree with the school security law than non-Green partisans, just looking at the results of the baseline model (not included in the graph). Nonetheless, Green partisans' opinions are affected by the court's ruling. A Green partisan in the control group has an 80% ($\pm$.4) probability of disagreeing with the school security law. However, this probability decreases by about 8 ($\pm$.8) percentage points on average when a Green partisan receives the endorsement that the GFCC approves the school security law. This first difference is statistically significant at the 90% level.

[13]For the simulation scenarios, this means that the control group dummy and the Green partisan dummy are set to one and all other dummies to zero. For the second scenario, all dummies but the GFCC disapproves dummy and the Green partisan dummy are set to 0.

Figure 2.6 – Effect of GFCC Endorsement on Partisans of the AfD and Greens

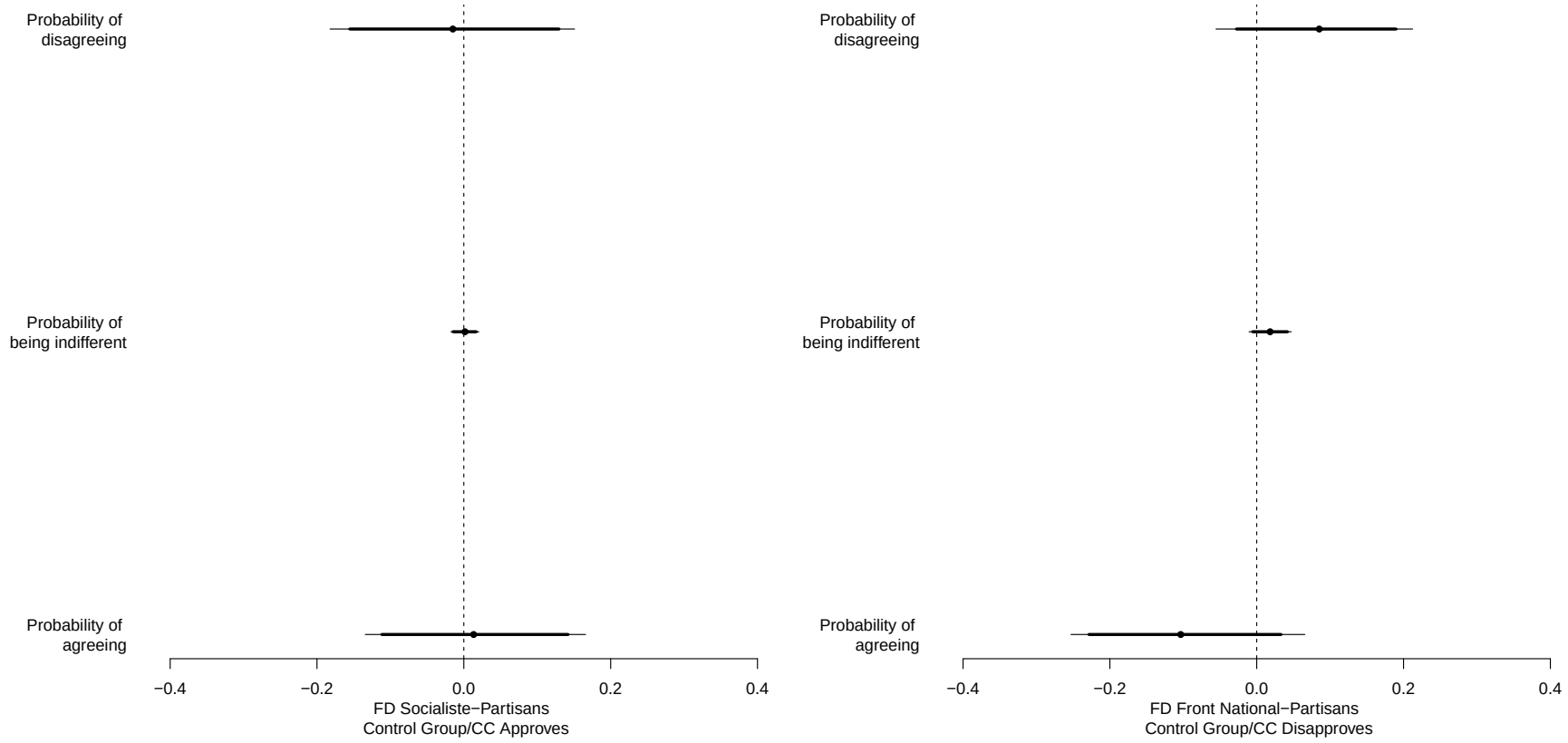


(a) FD Green-partisans

(b) FD AfD-partisans

Note: This figure shows the effect of GFCC endorsement on partisans of the Greens and the AfD. The first differences are calculated based on a simulation with $N = 1,000$ draws. The first difference on the left is the difference in the predicted probabilities of a Green partisan in the control group and the GFCC approves treatment group. The first difference on the right is the difference between an AfD partisan in the control group and the GFCC disapproves treatment group. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals. The corresponding regression tables are in Appendix A.7.

Figure 2.7 – Effect of CC Endorsement on Partisans of the Socialist Party and the Front National



(a) FD Socialist Party-partisans

(b) FD Front National-partisans

Note: This figure shows the effect of GFCC endorsement on partisans of the Socialist Party and the Front National. The first differences are calculated based on a simulation with $N = 1,000$ draws. The first difference on the left is the difference in the predicted probabilities of a Socialist Party-supporter in the control group and the CC approves treatment group. The first difference on the right is the difference between a Front National-supporter in the control group and the CC disapproves treatment group. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals. The corresponding regression tables are in Appendix A.7.

The same effect is also observable for AfD partisans, albeit in the opposite direction (right side of Figure 2.6). An AfD partisan in the control group has a 65% ($\pm$.5) probability of agreeing with the school security law. However, this changes when AfD partisans are exposed to the treatment where the GFCC disapproves the law. When receiving the treatment that the GFCC disapproves the law, the probability of disagreeing with the school security law increases by about 15 ($\pm$.7) percentage points on average. This first difference is statistically significant at the 95% level.

I conduct a similar analysis for the CC looking at respondents who are affiliated to the National Rally and the Socialists. The results are displayed in Figure 2.7. As expected, neither partisans of the Socialist Party nor of the National Rally are affected by the institutional endorsement of the CC. This means that in contrast to the GFCC, the French constitutional court does not possess sufficient legitimating power to move respondents with strong prior attitudes in the direction of its decisions.

My analyses show that the legitimacy-conferring capacity of the GFCC is sufficiently strong to even overcome strong pre-existing attitudes of individuals on the micro-level. The German court is capable of shifting individuals' positions in the direction of its decision, even if these initially have diverging preferences. This is particular strong evidence in favor of the perception of the GFCC and its judges as an extremely legitimate institution. In Appendix A.8.1, I further elaborate on both the GFCC's and the CC's diffuse support across different party affiliations. In short, whereas the majority of the French respondents, independent of their party affiliation, hold rather indifferent or even more negative views towards the CC, the GFCC enjoys substantial support across the entire political spectrum. This is additional evidence in favor of the outstanding popularity of the GFCC and its distinguished role in the political system of Germany. Especially the findings concerning the AfD partisans are remarkable, as the party leadership of the AfD recently started to verbally attack the GFCC after it has rejected an AfD's application against the refugee policy of the government as inadmissible.[14] This offers an promising initial point for follow-up studies which should investigate why individuals who intend to vote for the AfD nevertheless trust the GFCC sufficiently to follow its opinion leadership.

2.4.2 Institutional Trust and the Court's Legitimacy-Conferring Capacity

I have argued that the observed differences in the legitimacy-conferring capacity of the German and the French constitutional courts are due to varying degrees of public support. I documented the different levels of public support by means of the institutional trust of respondents in Figure 2.1, and showed that the GFCC is in fact

[14] See `https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2018/bvg18-087.html;jsessionid=9B4059135C36B058F13BCA005CB00895.2_cid394`, accessed 02.05.2019, for more information on the decision.

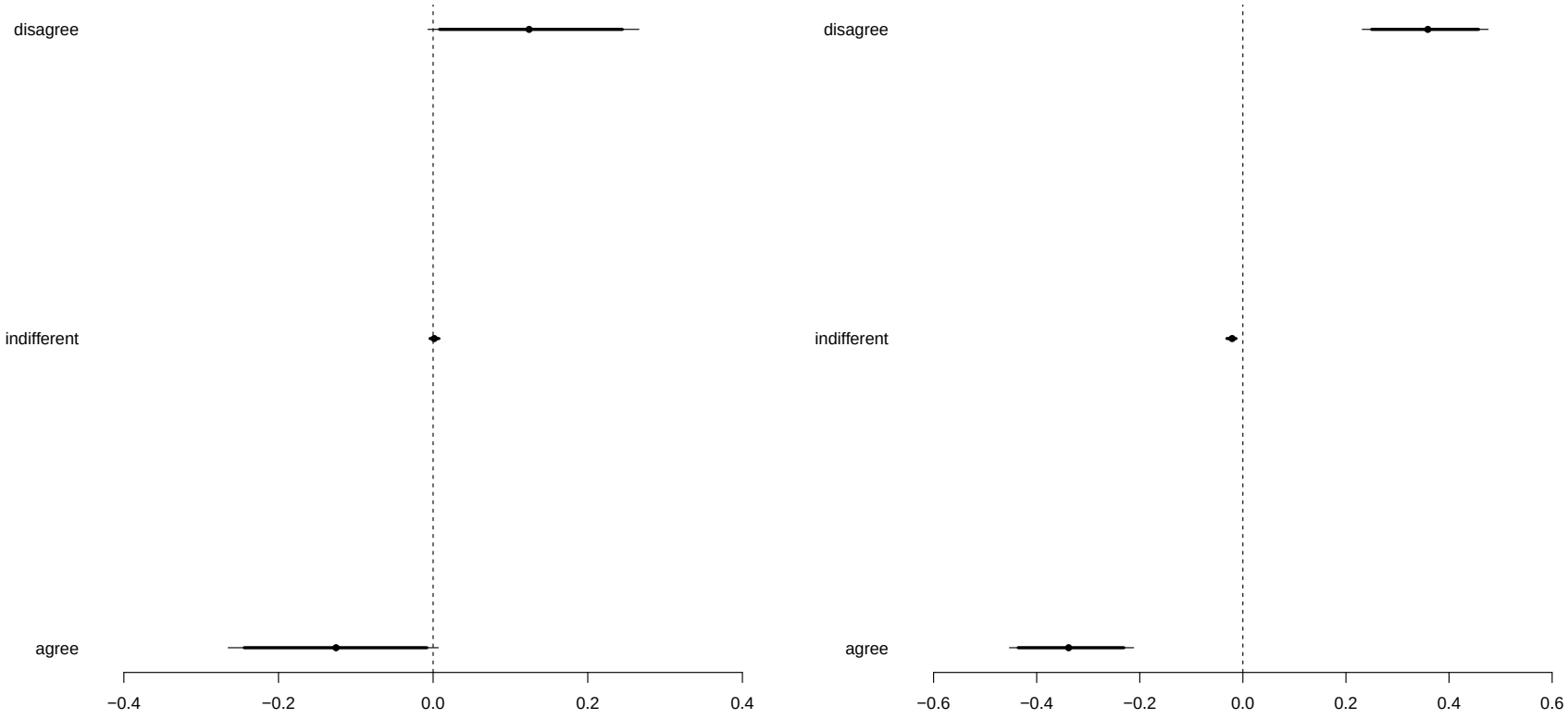
considerably more trusted than the French CC. Following this argument, I develop an additional observable implication that aims to disentangle the relationship of public support and the court's legitimacy-conferring capacity in further detail.

If the differences in the legitimacy-conferring capacity between the two courts are due to different levels of trust, respondents with low levels of institutional trusts should not perceive the respective court as a legitimate actor. Therefore, the institutional endorsement should not or only weakly affect such respondents, independent of whether the endorsing institution is the French or the German court. By contrast, someone who trusts the court also perceives it as legitimate and is therefore expected to be affected by the institutional endorsement. The observable implication is thus as follows: if it is truly institutional trust and resulting legitimacy that decides about the efficiency of court endorsement effects, then respondents with a high level of trust in the respective court should have a higher probability to follow the court endorsement than those with only low levels of trust, independent of whether they are German or French. In other words, if two respondents, one with a high level of trust and one with a low level of trust, are endorsed, then the endorsement should affect the respondent with high trust more strongly than the respondent with low trust.

This argument is tested by looking at the treatment groups which received the disagreement endorsements comparing respondents with low levels of trust to respondents with high levels of trust in these groups.[15] If trust plays an important role, we should observe significant differences between the two groups. In particular, respondents with high levels of trust should have a higher probability of disagreeing with the school security law (because this is the direction of the endorsement) than those who do not trust the court at all. This should hold true in both countries.

[15]This was tested using simulations again ($N = 1,000$). In order to calculate the quantity of interests, I include an interaction between the experimental groups and the trust variable. In order to compute the first differences, I set the dummy for the disagreement endorsement to one and all other experimental group dummies to zero, and varied the trust level between very low levels of trust (= 1) and very high levels of trust (= 7). See Appendix A.8 for more details.

Figure 2.8 – First differences between Low Trust/High Trust Respondents in Germany and France, Disapproval Endorsement



(a) FD Respondents Low Trust/High Trust, France

(b) FD Respondents Low Trust/High Trust, Germany

Note: Left Side: First Difference (FD) between French respondents with low trust and high trust who received the CC disapproves endorsement. N of simulation $= 1,000$. First Differences are calculated by setting the respective endorsement dummy to one and all other dummies to zero. Trust was varied between low (= 1) and high (= 7). Simulations based on an ordered probit model including the interaction of the experimental treatment groups and respondent's trust (Appendix A.8). The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals.
Right side: First Difference (FD) between German respondents with low trust and high trust who received the GFCC disapproves endorsement. The simulation is carried out following the same logic than in the French case.

Figure 2.8 shows the first differences between respondents with high and low trust levels who received the CC or GFCC disapproval endorsement. With respect to the French court (left side), we observe that respondents who have high trust in the court, and thus potentially perceive it as a legitimate actor, have an about 12 percentage points higher probability of disagreeing with the school security law than respondents who have only a very low trust in the court. Thus, the endorsement effect is considerably stronger among those who have confidence in the court than among those who have not. This is a remarkable change: respondents with low levels of trust have a 50% probability to agree with the school security law and only a 35% probability of disagreeing, although they received the endorsement that the CC disapproves the law. Thus, for these respondents the most likely outcome is to agree, and the treatment does not seem to work. A change of 12 percentage points means that with high levels of trust, the most likely choice of the respondents is now to disagree (47% probability to disagree, 38% probability to agree), which is in line the CC endorsement. This difference is statistically significant at the 90% level.

We observe the same pattern, but even stronger when looking at the German court (on the right side of the figure). Respondents who have high trust in the GFCC have a 36 percentage points higher probability of disagreeing with the school security law than those with only low levels of trust. This is a strong effect: respondents who do not trust the court have a probability of 53% of agreeing (and only 38% of disagreeing) with the school security. Thus, they behave in the opposite direction as the court endorsement suggests. Perceiving the court as legitimate leads respondents to entirely change their mind about the school security law, as the probability of disagreeing is now 74% (and only 20% for agreeing).

The analysis of the interaction between court endorsement and varying levels of trust confirms that institutional trust plays an important role in the perception of courts as legitimate actors. Even in the French data, where there is no evidence of a legitimacy-conferring effect of the CC on the aggregate level, I find that those who perceive the CC as legitimate react to the CC's endorsement.

2.4.3 Test of Model Assumptions and Robustness Checks

In this section I report five different robustness and diagnostic tests. In the first robustness test, I check whether the joint estimation of both GIP waves (in November 2016 and January 2017) affects the results (see Table A.10 and Table A.11 in the Appendix). Using only the data of Wave 26, the coefficient for the GFCC disapproves endorsement is barely not statistically significant ($p = .14$, two-sided test), and in Wave 27 the FCC approves endorsement is just not significant ($p = .11$). All other effects are similar to the aggregated analysis.

For the second diagnostic test, I explore the effect of potential individual heterogeneity. Previous research shows that with respect to an individual's legitimacy perception,

knowledge about the constitutional court can introduce heterogeneity (e.g. Hoekstra, 2000; Sen, 2017). If the legitimating power of the constitutional courts systematically differs depending on how knowledgeable respondents are, then an individual's knowledge should be taken into account. In order to test this, I use two questions in the surveys which measure the respondent's knowledge about the courts. I find the same patterns than in the main analysis, independent of how knowledgeable respondents are (see Appendix A.9.4 for more details).

In the third robustness check, I assess whether the school security law as governmental policy presented in the experiment might alter the results. Respondents have (in both Germany and France) a rather negative opinion towards such a policy. In order to test whether the findings are dependent on the choice of governmental policy, a similar experiment was implemented in Wave 16 of the ENEF again, but this time another governmental policy issue was chosen (a potential increase of the retirement age). Using this data, I am able to replicate my findings from the initial experiments: even when considering a different policy, the French CC has no legitimacy-conferring capacity.

In the fourth robustness test, I replicate the main analyses again but this time I use the original five-point scale of the respondent's opinion on the school security law as dependent variable instead of the recoded three-point scale (Table A.13 and A.12 in the Appendix). Using the original five-point scale does not alter the results. In fact, in the German analysis the effects become more profound.

In the fifth and last robustness check, I evaluate whether the people's trust in court rating might be affected by the treatment group they received. This could be possible because the trust rating in the GIP was asked after the experimental manipulation in the form of the institutional endorsement. If the treatment and the trust-rating are not independent, then the previous analyses would suffer from the problem of endogeneity. However, *t*-tests of the trust rating individuals in each experimental group provide insignificant results. This shows that there are no statistically significant differences in the trust ratings of respondents receiving different institutional endorsements.

2.5 Conclusion

In this chapter, I have tackled the question of whether constitutional courts can change public opinion by endorsing a certain policy position. Scholars have long debated whether courts have a legitimacy-conferring capacity and can move public opinion by placing their stamp of approval or disapproval on a certain public policy. The answer to this question is yes, but my results show that specifying the conditions under which one should expect opinion changes is more complex than previously thought. My comparative analysis of respondents in Germany and France demonstrates that the legitimacy-conferring capacity of constitutional courts is highly depended on their institutional reputation and thus, their legitimacy. The GFCC, a court that enjoys

considerably high public support, is capable of moving public opinion in the direction of its decisions. This effect is so powerful that even respondents with strong pre-existing attitudes are affected. By contrast, the French CC, a rather unpopular court, does not possess the same legitimacy-conferring capacities. Regardless of whether it approves or disapproves a public policy, I find no evidence of a shift in respondent's opinion.

These findings have important implications at large for both our understanding of the role of constitutional courts in democratic politics and for public opinion formation in general. What is known from the US Supreme Court does not hold unconditionally for all European constitutional courts. The legitimacy-conferring capacity of courts is highly dependent on their institutional legitimacy and thus, their diffuse support. This aspect should be considered within the concept of comparative politics.

My work also opens up new avenues for further research. Because survey experiments are often criticized with respect to their external validity, it is necessary that additional studies confirm the experimental evidence with observational data, for instance from a panel where respondents are asked before and after a (landmark) decision takes place. Furthermore, future studies should take into account different salient and non-salient public policies, the role of the media as a mediator of how the public learns about a decision or how individuals form their attitudes when they have access to competing arguments. Finally, further research is required to disentangle the causal mechanisms of how public support and institutional legitimacy translate into the legitimacy-conferring capacity of constitutional courts.

CHAPTER 3

The Automatic Detection of Vague Language in Constitutional Court Decisions

3.1 Introduction

Vague language is a common phenomenon in political communication. It is used as a communicative strategy in the relationship between the mass public and political actors. For instance, vague language can be strategically used to increase the flexibility of future policy actions (see Alesina and Cukierman, 1990; Marschall and McKee, 2002; Meirowitz, 2005; Eichorst and Lin, 2019). Moreover, vague language may be used in contexts where information verification is costly, and being incorrect might be connected to paying reputation costs (e.g. Chortareas, Stasavage and Sterne, 2002). Third, and most important for this chapter and this dissertation, vague language can be used as a strategic tool to influence the behavior of other (political) actors.

The argument of vague language as a strategic tool also appears in work on judicial politics (Staton and Vanberg, 2008; Owens and Wedeking, 2011; Owens, Wedeking and Wohlfarth, 2013; Corley and Wedeking, 2014; Cross and Pennebaker, 2014; Black et al., 2016*a*,*b*). Staton and Vanberg (2008) for instance argue that judges use vague language strategically to leverage the policy expertise of other actors or hide likely legislative non-compliance from the public. Others argue that judges strategically vary the clarity of their decisions when they rule contrary to public opinion in an effort to maintain public support as best as they can (Black et al., 2016*a*). Related to this, other authors show that judges strategically obfuscate their opinion language to circumvent unfavorable responses from political branches (Owens, Wedeking and Wohlfarth, 2013).

Unfortunately, although vague language plays an important role in (theoretical) political science literature in general and research on judicial politics in particular, the discipline lacks the appropriate tools to *automatically*[1] detect vague language in applied work. The reason for this situation is mainly the lack of appropriate data and that computer linguists only recently became interested in interdisciplinary work on this topic.[2]

Motivated by this gap, the goal of this chapter is *to develop a measure for vague language in written decisions of the German and French constitutional court*. I propose two different methodological approaches. The first approach is a dictionary approach and relies on word embeddings to expand a widely used general vagueness dictionary (the Linguistic Inquiry and Word Count dictionary, hereafter *LIWC*) to the legal domain. In the second approach, I benchmark a set of machine learning algorithms to develop a binary sentence classifier that is able to classify sentences of GFCC decisions into vague and not vague. For the training, I use a novel self-collected data set consisting of over 3,500 sentences. I find that both approaches are able to automatically detect vague language in written court decisions and provide valid and robust measures of a decision's linguistic vagueness. Moreover, my benchmark tests show that both methods outperform the generic LIWC dictionary approach. My results thus have implications for both scholars of social science and computational linguistics.

This chapter proceeds as follows. I first discuss the concept of linguistic vagueness in the legal context and provide a brief overview of related work on linguistic vagueness detection in both the social sciences and the computational linguistics literature. I then propose two different methodological approaches for measuring vague language and apply them to decisions of the German and French constitutional court. I finally demonstrate the validity of the new measures and discuss several avenues for further research.

3.2 Related Work and Challenges

This section gives a brief overview of the classification task by defining the concept of linguistic vagueness in the context of constitutional court decisions. In particular, I introduce the concept of *judicial policy implementation vagueness* and provide an overview of the detection of linguistic vagueness in the field of social science and computational linguistics.

[1] Automatically here refers to methodological approaches which require little or no direct human intervention for the vagueness classification.

[2] Notable exceptions include Štajner et al. (2016), Štajner et al. (2017) and Theil, Štajner and Stuckenschmidt (2018*b*).

3.2.1 Linguistic Vagueness in Judicial Decisions

The computational linguistics literature distinguishes between *ambiguity* and *vagueness*. These two concepts are not equivalent. An expression is ambiguous if it has two or more distinct denotations – that is, if it is associated with more than one region of a meaning space. A standard example is the word "bank", which can denote the rim of a river or a financial institution (Poscher, 2012, 2). By contrast, an expression is vague if the region it denotes does not have perfectly well-defined boundaries. A simple example of a vague word is "tall". It is unclear what exactly constitutes a "tall" person. Someone who might be considered "tall" on the street might not be considered tall in a basketball team. Even if you state that a person of 185 centimeters is considered tall, then would a person of 184 centimeters not be considered tall?

As noted in Štajner et al. (2017), when social scientists speak about vagueness, they actually mean *linguistic hedging*. Linguistic hedging is an umbrella term for the use of speculative, uncertain or vague language[3], and is defined as "any linguistic means used to indicate either a) a lack of complete commitment to the truth value of an accompanying proposition, or b) a desire not to express that commitment categorically" (Hyland, 1998, 1). These so-called "hedges"[4] are words "whose job is to make things fuzzier or less fuzzy" (Lakoff, 1973, 471). Put differently, hedges related to the intentional use of certain words or terms to modify the information content or range of possible interpretation in a sentence. Eichorst and Lin (2019) refer to this as the "intentional feature of word choice that modifies and scales this range" (Eichorst and Lin, 2019, 17).

In consideration of the definition of hedging and the particular context of constitutional court decisions, I introduce a new concept called "*judicial policy implementation vagueness*". I define judicial policy implementation vagueness as a particular form of hedging in the context of court decisions and as the *intentional choice of words or terms that give the legislator a wide decision leeway and room to maneuver in how it can implement a court decision*. The following two statements in court decisions illustrate how judicial policy implementation vagueness appears in the context of court rulings (German original below):

*The **decision leeway of the legislator** allows for **different regulatory mechanisms**.*[5]
(1 BVL 1/98, GFCC in the context of unemployment benefits 2000; own translation)

[3]The terminology of linguistic hedging is not used consistently even in the computational linguistics literature. For instance, in the context of academic writing and the biomedical domain, sentences with hedges are referred to as speculative sentences.

[4]Hedges are also called "weasel words", e.g. in the context of Wikipedia articles. Contributors and editors of Wikipedia are encouraged to tag these weasel words in articles for further improvement.

[5]Original quote in German "Der **gesetzgeberische Gestaltungsraum** lässt hier **verschiedene Regelungsmöglichkeiten** zu."

> *It is the **responsibility of the legislator** to differentiate the cases of medically indicated and not medically indicated abortion.*[6]
> (BVerfGE 39, 1, GFCC in the context of the abortion decision 1974; own translation)

Both statements contain several hedges (shown in bold), for instance: it is the "decision leeway of the legislator" that allows for "different regulatory mechanisms", and it is the "responsibility of the legislator" to differentiate cases of medical indicated and not indicated abortion. In both cases, the GFCC attributes the legislator a wide room to maneuver and gives it large freedom in how the actual implementation of the decision could look like.

Court decisions are different to other (political) text data such as Tweets, party manifestos or speech texts. Three properties are important regarding the measurement of vagueness. First, court decisions are, in general, rather long texts with a complex structure. Decisions of the GFCC, for instance, have an average length of 5,232 words (embedded in 316 sentences) and decisions of the French CC have an average length of 2,234 words (in 36 sentences). This makes it challenging to detect hedges or sentences with hedges given these long texts. Second, any measurement strategy must consider that the probability of observing judicial policy implementation vagueness is not the same across different text sections in a decision. Often, a considerable part of a decision body contains a detailed summary of a decision's context (e.g. the different stages of appeal), the legal framework of a law, and a description of the plaint. In these text sections, it is rather unlikely to observe judicial policy implementation vagueness. This is why they must be excluded from the measurement. Third, hedges that are common in other domains such as Wikipedia articles or scientific abstracts might not be a useful indicator of hedging in court rulings. This is because different domains often use domain-specific language, and any measurement strategy must account for that.

3.2.2 Linguistic Vagueness in Social Sciences and Computational Linguistics

In this section, I will discuss related work in the social sciences, legal scholarship and the computational linguistics literature and demonstrate that the appropriate measurement of judicial policy implementation vagueness requires new and innovative methodological approaches.

Related Work in Social Science and Legal Scholarship

The systematic and automatic detection of vague language in (written) texts is still in its infancy in the social sciences. The majority of studies, for instance in political

[6] Original quote in German: "Es ist **Sache des Gesetzgebers**, die Fälle des indizierten und des nicht indizierten Schwangerschaftsabbruchs näher voneinander abzugrenzen."

discourse analysis, rely on qualitative approaches, extensively studying the use of vague language in specific events or certain situations (Gruber, 1993; Fraser, 2010; Giuseppina Scotto di Carlo, 2013). Most of the legal scholarship is mainly engaged in a normative or philosophical discussion of the meaning and value of vague language in certain domains of law (Post, 1994; Waldron, 1994; Jónsson, 2009; Poscher, 2012; O'Rourke, 2017). None of these approaches are beneficial for the automatic detection of vagueness in judicial decisions.

The bulk of quantitative work of social scientists on vague language uses dictionary-based approaches, with varying degrees of complexity. Dictionaries use the frequency of the occurrence of key words in a text to classify documents into categories. The simplest of these dictionary methods count how often modal auxiliaries such as "may, might, can, should" appear in certain texts (e.g. Rabab and Rumman, 2015). In the political science literature, vague language is primarily identified using such dictionaries. In this context, recent scholarship mostly relies on the dictionary of vague words from the *Linguistic Inquiry and Word Count* (LIWC) platform (Tausczik and Pennebaker, 2010). LIWC simply counts the frequency of vague words which are defined in a pre-specified dictionary. Vague words of this dictionary are, for instance: *possible*, *or*, *some*, *unclear*, and *perhaps*. Due to its simplicity and availability in multiple languages, LIWC and its collection of dictionaries are currently the state-of-the-art measure for vagueness (or similar concepts) in political science (e.g. Owens and Wedeking, 2011, 2012; Cross and Pennebaker, 2014; Black et al., 2016*b*; Wedeking and Zilis, 2018; Eichorst and Lin, 2019). I will elaborate on the usage of dictionary-based approaches in the context of judicial politics in the next section.

These dictionary based approaches suffer from at least two limitations with respect to the measurement of judicial policy implementation vagueness, which I will highlight in the following. I focus my discussion on LIWC, although my critique in principle also applies to other dictionaries.

1. Many dictionaries originate from the field of psychology. LIWC, for instance, was developed to identify cognitive properties of human individuals from experiments about expressive writing in English language (Tausczik and Pennebaker, 2010). The validity and reliability of the vagueness dictionary of LIWC and its translations to other languages have been extensively tested (Wolf et al., 2008; Piolat et al., 2011; Pennebaker et al., 2015). However, it remains unclear how LIWC performs when being applied to different domains, and judicial decisions in particular. The problem of using dictionary methods for different domains is also discussed in Grimmer and Stewart (2013) in the context of political texts.

2. Dictionaries only employ simple counts of words which define a particular dimension. LIWC and related programs cannot detect semantic relationships in texts, for instance whether the meaning of a word changes in different contexts.

They can also not measure negations or latent traits of human language such as sarcasm or irony.

Due to these shortcomings, I argue that LIWC (or any other *general* vagueness dictionary) is not a reasonable choice to measure judicial policy implementation vagueness. However, due to its popularity in the discipline I will use it as a baseline to compare my own approaches with. In summary, the (quantitative) detection of linguistic vagueness in the social sciences and legal scholarship mainly relies on the use of general dictionaries, which raises concerns about the quality of these measurement.

Related Work in Computational Linguistics

In the natural language processing (NLP) and computational linguistics literature, the detection of linguistic vagueness has attracted considerable scholarly interest. Extensive work exists on how to properly identify and classify linguistic vagueness. This work often relies on large data sets and uses different unsupervised and supervised machine learning approaches to detect linguistic vagueness.[7] Moreover, computer linguists can draw on established, annotated data sets which often serve as benchmarks for new methods (e.g. Farkas et al., 2010; Vincze et al., 2008).

Most of the work on linguistic vagueness in the NLP literature deals with two types of texts: scientific texts from the (bio)medical domain (Light, Qiu and Srinivasan, 2004; Medlock and Briscoe, 2007; Vincze et al., 2008; Szarvas, 2008), and Wikipedia articles (Ganter and Strube, 2009; Farkas et al., 2010). These classification approaches reach accuracies between 80% and 95%, depending on the classification task. Other NLP applications include the detection of linguistic vagueness in website privacy policies (e.g. Reidenberg, Breaux and Norton, 2016; Liu, Fella and Liao, 2016). Only recently, the attention of computer linguists has turned to the domain of social science. This, often interdisciplinary, work includes studies on speculative sentences in transcripts of monetary policy meetings on the U.S. central bank (Štajner et al., 2016, 2017) or the detection of uncertain statements of stakeholders in the financial market (Theil et al., 2018; Theil, Štajner and Stuckenschmidt, 2018*b*).

Although computer linguists increasingly work on social science problems, there are at least three major obstacles that circumvent directly adapting these approaches to measure judicial policy implementation vagueness in court rulings:

1. Most of the existing algorithms and methodological approaches are only developed for texts in English language.

2. The existing algorithms and approaches are not specifically developed (and therefore, not directly applicable) to the judicial domain. The work of Theil et al.

[7] However, also dictionary-based approaches are used (for a dictionary-based approach in the financial domain see e.g. Loughran and Mcdonald, 2011).

(2018) and Theil, Štajner and Stuckenschmidt (2018*b*) shows that the performance of NLP classifiers suffers if models designed for one domain are applied to others without adaption.

3. No large data set of court decision texts exists that are annotated with respect to linguistic vagueness that could be used to train NLP classifiers. The lack of training data is not only a problem of German or French texts, but also generalizes to English.

In summary, although computer linguists traditionally have a long history and extensive experience with the detection of linguistic vagueness, there is not much prior work that I could rely on when trying to detect judicial policy implementation vagueness in constitutional court decisions. In the next sections, I will therefore propose two approaches that break new ground from both a social scientist and computational linguistics perspective.

3.3 Method 1: Exploiting Word Embeddings for Domain-Specific Dictionaries

The first method I use to automatically detect judicial policy implementation vagueness in court rulings is a dictionary-based approach. In particular, I show how word embeddings can be used to find meaningful word analogies to expand a general dictionary to a specific domain. In what follows, I briefly discuss the role of dictionary-based approaches in NLP and introduce word embeddings, a popular method in NLP to map words from a given vocabulary to vectors of real numbers. I then demonstrate that it is straight forward to use these embeddings to tailor a general dictionary such as LIWC[8] to the specific domain of court rulings.

3.3.1 Dictionary-Based Approaches in Judicial Politics

Dictionary based approaches are a popular approach in text sentiment analysis and opinion mining. A dictionary (also called lexicon) contains a collection of words which are mapped to certain categories or dimensions. The Dictionary of Affect in Language (DAL) for instance is a dictionary which is designed to measure the emotional meaning of words and texts (Whissell et al., 1986). The LIWC dictionary (Pennebaker and King, 1999) is another popular dictionary which contains a collection of words belonging to over 70 predefined dimensions. Most often, these dictionary simply count the occurrences of words belonging to a certain dimension relative to the overall amount of

[8]I want to thank Dr. Markus Wolf for providing me with the German LIWC dictionary (Wolf et al., 2008).

words in a text. Such dictionary-based methods are also increasingly used in the context of judicial decisions (e.g. Owens and Wedeking, 2011, 2012; Cross and Pennebaker, 2014; Black et al., 2016*b*; Wedeking and Zilis, 2018). Dictionaries are used, for instance, to measure the "cognitive complexity" of Supreme Court justice's opinions (Owens and Wedeking, 2012), the "legal clarity" of Supreme Court decisions (Owens and Wedeking, 2011) or the "textual readability" of Supreme Court rulings (Black et al., 2016*b*, Chapter 3).

I argue that these dictionaries should be used with caution or at least with detailed validation, because "applying dictionaries outside the domain for which they were developed can lead to serious errors" (Grimmer and Stewart, 2013, 268). A dictionary-based approach only works when the meaning (e.g. vagueness) associated with a word is closely aligned with how the word is used in a certain context. Unfortunately, this is not always the case when e.g. dictionaries originating from the field of psychology are applied to judicial decision texts. A dictionary developed for a certain domain that is then applied to another often lacks discriminative capacity in most contexts. Therefore, they need to be adapted to the specific domains to which they are desired to be applied to provide valid results (see Loughran and Mcdonald, 2011; Young and Soroka, 2012). Loughran and Mcdonald (2011) provides a clear example for this in an analysis of the tone of corporate earning reports in the accounting literature. They show that many words that have a negative connotation according to widely used sentiment dictionaries are words typically not considered negative in the financial context. For instance, the word "cancer" has not necessarily a negative connotation if it is mentioned by a health-care company. Although the "off-the-shelf" usage of general dictionaries is at least questionable, there are several examples in recent judicial political research where dictionaries are used without any domain adaption. The consequence is that most of these analyses are built on "shaky foundations" (Grimmer and Stewart, 2013, 275).

Wedeking and Zilis (2018) for instance use the aforementioned DAL and LIWC dictionaries to measure disagreeable rhetoric in Supreme Court majority opinion. In almost the same manner, Owens and Wedeking (2012) collapse ten LIWC indicators into one measure which they say to capture a judge's cognitive complexity. Both dictionaries have originally emerged from the field of psychology, and it is questionable whether such dictionaries can be carelessly applied on other specific contexts such as written opinions of judges. Wedeking and Zilis (2018) for instance discuss a couple of examples to demonstrate the face validity of their approach, whereas Wedeking and Zilis (2018) just refer to the overall internal and external validity of the LIWC approach demonstrated elsewhere (Owens and Wedeking, 2012, 493). I argue that more elaborated methods are necessary to demonstrate the validity of their measurement, for instance the agreement of their measures with human-based annotations or the correlation with other vagueness scores.

Researchers interested in domain-specific dictionaries have to devote considerable time and resources into a manual expansion of dictionary terms using the domain-specific texts. This means that researchers and domain experts have to manually work through the texts and identify the domain-specific terms which are semantically similar to the entries in the general dictionary. This procedure is also known as finding *word analogies* or *expansion candidates*.

With recent advances in NLP and neural networks and the increase in computational capacities, an alternative to the manual expansion became available: *dictionary expansion using word embeddings*. The idea behind this approach is simple: automatically identifying domain-specific words which are semantically related to the words in the general dictionary. Stated differently, an algorithm takes over the task of finding word analogies for which usually human researchers and experts are required. Such an approach is currently the state-of-the-art in terms of automatic dictionary expansion.[9] Tsai and Wang (2014) automatically expand a set of sentiment dictionaries containing vocabulary specific to the financial domain by training word embeddings on a corpus of financial texts and adding the 20 most similar terms to each original dictionary entry. Theil, Štajner and Stuckenschmidt (2018*b*) follow a similar strategy and use word embedding models for automatically expanding a financial dictionary of uncertainty triggers. Both studies show that the expanded dictionaries improve the detection of uncertainty triggers in the financial domain. In a similar fashion, Setiawan, Widyantoro and Surendro (2017) employ word embeddings to address the problem of vocabulary mismatch in Tweets. Nonetheless, while the idea of using word embeddings to expand a general dictionary is a well-established methodology in the NLP community, to the best of my knowledge it is not commonly used in political science research to date.[10]

3.3.2 Introducing Word Embeddings

Word embeddings are continuous high-dimensional vector space models based on shallow *neural networks* to learn vector representations of each word in a background corpus, such that similar words are close to each other in the word embedding space. In this geometric space, the geometric relationships between word vectors reflect the semantic relationships between these words. In a reasonable embedding space, for instance, you would expect synonyms to be embedded into similar word vectors. Moreover, the geometric distance between any two word vectors should relate to the semantic distance between the these words. This allows for simple algebraic operations on these vectors: "King - Man + Woman" results in a vector very close to "Queen"

[9] Word embeddings are also increasingly used in NLP because they provide more powerful word representations than other approaches such as Bag-of-Words.

[10] Word embeddings in general are new to political science. Only few research studies have begun to take advantage of them. For instance Rheault and Cochrane (2018) use word embeddings for the estimation of ideological placements in a parliamentary corpora.

(Mikolov, Yih and Zweig, 2013, 747). Word embeddings have recently become popular as a text representation, since the vectors produced can be compared to find semantically (rather than textually) similar words using similarity metrics (e.g. cosine similarity). This makes them useful for finding semantically similar words analogies given a list of input words. This list of input words is usually called "*queries*" or "*seeds*".

Although there exist already pre-trained word embeddings,[11] there are two reasons not to use them for developing a domain-specific dictionary. First, most of the popular and validated embeddings are in English language, and thus not useful for expanding the dictionary to German and French. Second, the pre-trained embeddings are trained on general large corpora such as the Google News dataset (Mikolov, Yih and Zweig, 2013) or different Wikipedia corpora (Pennington, Socher and Manning, 2014). However, what makes a good word-embedding space strongly depends on the task: the perfect word embedding space for a costumer rating analysis model may look different from the perfect embedding space for a financial-document-classification model, because semantic relationships might vary from task to task (Allaire and Chollet, 2017, 171). In the next section, I explain the estimation of word embeddings in further detail.

The word2vec CBOW Architecture

In general terms, a word embedding can be described as a mapping $V \rightarrow \mathbb{R}^D : w \rightarrow \overrightarrow{w}$ that maps a word w from a vocabulary V to a real-valued vector $\overrightarrow{w}$ in an embedding space of dimensionality D. There exist multiple embedding methods, but here I focus on one the most popular techniques, the *continuous bag of words* (CBOW) approach proposed by Mikolov, Yih and Zweig (2013). This method, commonly referred to as *word2vec*, is a probabilistic prediction approach. Given a number of context words around a target word w, these models try to predict w using the context words w around it.

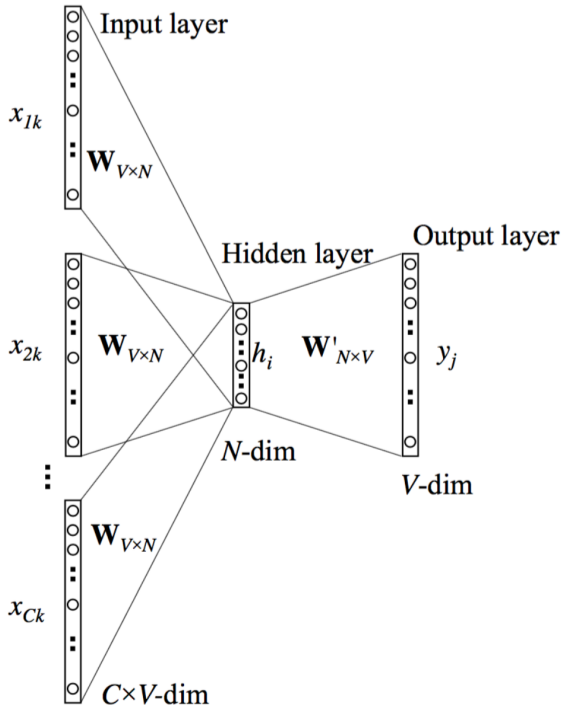
Figure 3.1 shows the general network topology of the CBOW model.[12] There are only three layers in this network: an input layer, a hidden layer, and an output layer. The input layer is represented by a one-hot[13] encoded vector x of context words $x_1, x_2, ..., x_C$ for a word window of size C and a vocabulary size V. C is a hyper-parameter and often a window size between five and eight is used. This window is usually symmetrical to the left and the right. An alternative way of imagining this is to think about a sliding window over the text, that includes the central word currently in focus, together with

[11] The pre-trained *word2vec* model of Google developed in the original *word2vec* paper of Mikolov, Yih and Zweig (2013) is trained on roughly 100 billion words from a Google News dataset. Other popular pre-trained word embeddings are GloVe embeddings (Global Vectors for Word Representation) developed by Stanford researchers (Pennington, Socher and Manning, 2014).

[12] In the original paper of Mikolov, Yih and Zweig (2013), another word embedding architecture is introduced. This architecture is called *skip-gram*, and actually reverses the logic of CBOW: the surrounding words of w are predicted given w.

[13] One-hot encoding means that for the input vector $x = \{x_1, x_2, ..., x_V\}$, $x_k = 1$ and all other $x_{k'} = 0$ for $k \neq k'$.

Figure 3.1 – General Network Topology of the CBOW Model



Note: This figure shows the general topology of the CBOW model. Figure taken from Rong (2014, 6).

the C preceding and C trailing words. The central word in the middle of the sliding window is the target word we like to predict in the output layer. The output layer y is a one-hot encoded vector of length V.

The hidden layer h is an N dimensional vector, where N is the number of dimensions chosen to represent the words. It is arbitrary, which means the researcher can set it prior to training. The one-hot encoded input vectors are connected to the hidden layer via a $V \times N$ weight matrix W and the hidden layer is connected to the output layer via a $N \times V$ weight matrix W'.[14] Interestingly, in *word2vec* we are not interested in actually using the neural network for the task we trained it on (predicting a word based on its context). Instead, the goal is to learn the weights: the weight matrix W' is the final embedding matrix that encodes the meanings of words as context that is used to find the semantically similar candidates later. The forward propagation in this simple neural network works as follows. The first step is to evaluate the output of the hidden layer h:

[14]Note that W' is not the transpose of W, but stands for a different matrix.

$$\begin{aligned} h &= \frac{1}{C} W^T \cdot (x_1 + x_2 + ... + x_C) \\ &= \frac{1}{C} (v_{w_1} + v_{w_2} + ... + v_{w_C})^T \end{aligned} \tag{3.1}$$

where C is the context window, $w_1, ..., w_C$ are the words in the context, and v_w is the input vector of a word w. In other words, given C input word vectors, the activation function for the hidden layer h sums up the corresponding rows in W, and divides by C to take their average. The hidden layer is connected to the output layer via another weight matrix W'. Using these weights, it is possible to compute a score u_j for each word in the vocabulary:

$$u_j = {v'_{w_j}}^T \cdot h \tag{3.2}$$

where v'_{w_j} is the j-th column of the output matrix W'. These are the inputs for the output layer. The output of the output layer is obtained by passing u_j into the $softmax$ function to obtain the conditional probability of words using $softmax$[15]:

$$y_j = p(w_j | w_1, ..., w_C) = \frac{exp(u_j)}{\sum_{j'=1}^{V} exp(u'_j)} \tag{3.3}$$

where y_j is the output of the j-th unit in the output layer. In the next step, the weight matrices W and W' are learned. In the beginning, W and and W' are randomly initialized. Training examples are sequentially fed into the model and the error is observed. The error is defined by a loss function of the difference between the expected output and the actual output. The training objective (for one training sample) is to maximize 3.3, namely the conditional probability of observing the actual output word w_O given the input context words $w_{I,1}, \ldots, w_{I,C}$ regarding weights in weight matrix W and W'. The loss function $\mathcal{L}$ is defined as:

[15] $Softmax$ is known as multinomial logit in the political science context.

$$\begin{aligned}\mathcal{L} &= -\log \mathbb{P}(w_o|w_{I,1},\ldots,w_{I,C}) \\ &= -u_{j^*} + \log \sum_{j'=1}^{V} \exp\left(u_{j'}\right) \\ &= -{v'_{w_O}}^T \cdot h + \log \sum_{j'=1}^{V} \exp\left({v'_{w_j}}^T \cdot h\right)\end{aligned} \tag{3.4}$$

here j^* is the index of the actual output word w_O in the output layer. The loss function is minimized finding the values for W and W' that minimize the loss via back propagation using stochastic gradient descent methods.[16] Readers interested in more details about updating the weights are directed towards the excellent introduction paper of Rong (2014).

3.3.3 Application to German and French Court Decisions

In what follows, I apply a three-step procedure to obtain a measure of judicial policy implementation vagueness for all French and German court rulings. In the first step, I employ a corpus of decision texts of the German and French constitutional courts to train judicial decision-specific word embeddings. In the second step, I use the words from the LIWC dictionary as seeds to obtain the top-n most similar terms to the seed words. These are called the top-n candidate terms. Finally, I utilize these top-n candidate terms to expand the list of vague words from the LIWC dictionary and to obtain a domain-specific vagueness dictionary. This allows me to measure the judicial policy implementation vagueness in German and French decision texts with dictionaries tailored to the judicial domain.

Step 1: Training Word Embeddings

I train judicial decision-specific word embeddings on two corpora of decisions of the French and German constitutional courts. As with most deep learning algorithms, learning improves with the amount of data provided. For the German case, I use all Senate decisions available in the Constitutional Court Data Base[17] (2,006 documents) plus all so-called chamber decisions[18] (6,415 documents) available on the website of

[16] There exist several efficiency optimization tricks to train the model faster, such as hierarchical softmax and negative sampling. These are beyond the scope of this short introduction.

[17] The Constitutional Court Database (CCDB) (Hönnige et al., 2015) contains data on all decisions of the German Federal Constitutional Court between 1972 and 2010. This database is part of the research project entitled "The German Federal Constitutional Court as a Veto Player" funded by the German Research Foundation, and is located at the University of Hannover and the University of Mannheim.

[18] Chambers are panels of three judges, which help the GFCC to make timely decisions for a large number of cases that are deemed to be not sufficiently important or controversial to be deliberated among all judges on the bench.

the GFCC as of August 2018. The final training data includes a vocabulary size of 148,465 (unique) words and terms (uni-grams, bi-grams, tri-grams), overall more than 5.1 million words.

For the French court, the training data contains 1,343 decisions obtained via the Conseil Constitutionnel's website.[19] The final training data includes a vocabulary size of 56,301 (unique) words and terms, overall accounting for more than 3.7 million words.

Although these corpora might look small compared with the corpora used in the original *word2vec* implementation (where billions of words where used), several studies suggests that the specificity of a corpus has much stronger influence on the quality of the embeddings than its size (Lai et al., 2016; Dusserre and Padró, 2017). Moreover, when the goal is to automatically find word analogies, it is more important to have a corpus that correctly represents the usage of these specific words than to have huge amounts of text data unrelated to the targeted context (Dusserre and Padró, 2017, 2).

Text preprocessing was kept at a minimum because its has proven to have a significant influence on the results obtained (Denny and Spirling, 2018). For this reason, the texts are mainly left untouched: I did not remove punctuation, numbers or stopwords.[20] Also, I applied no lowercasing of words. Especially the removal of stopwords would not be meaningful in the context of my application, because modal words such as "can" or "should" are considered to be stopwords (for instance in the most popular list of stopwords of Porter (1980)), but are in fact of high relevance for the informative structure of a sentence (Theil, Štajner and Stuckenschmidt, 2018*b*, 6), and therefore, for my analysis.

As unique terms I considered uni-grams, bi-grams, tri-gram. This is because some words have a highly ambiguous meaning when taken out of context. For instance, the word "sun" has substantially different interpretations when used in the expressions: "sun spider" (an order of animals in the class Arachnida) and "sun bath". *N*-grams help to capture these context dependent differences, where formally a *n*-gram is defined as a contiguous sequence of tokens of length *n* (Manning and Schütze, 1999). For example, the multiword expression "a contiguous sequence" from the previous sentence would be referred to as a trigram. In order to train the *word2vec* model, I use the *R word2vec*[21]

[19] I only use QPC (question prioritaire de constitutionnalité) and DC (Contrôle de constitutionnalité des lois ordinaires, lois organiques, des traités, des règlements des Assemblées) decisions. Other types of decisions, such as LP (Loi du pays de Nouvelle-Calédonie), LOM (Compétences outre-mer) or FNR (Fins de non-recevoir) do not meet the requirements of my definition of judicial policy implementation vagueness in court rulings: these decisions are not directed at governmental laws, and are often very context specific.

[20] Stopwords are frequently appearing words such as "the", "of", "in".

[21] The package is called WordVectors created by Benjamin Schmitt. The package can be found at Github: `https://github.com/bmschmidt/wordVectors`.

implementation with standard parameters.[22] The result is a high-dimensional (100 dimensional) vector representation, where each unique word is embedded in a word vector.

Before moving to Step 2, the identification of the top-n expansion candidates, it is important to check whether the trained embeddings capture useful information about the word's representations and their contexts. There is no clear agreement in the computer linguistic literature about how to evaluate the results of word embeddings (Schnabel et al., 2015; Jastrzebski, Leśniak and Czarnecki, 2017), nor is there a simple goodness-of-fit measure. I use a dimensionality reduction technique called *t-Distributed Stochastic Neighbor Embedding* (*t*-SNE) (van der Maaten and Hinton, 2008) to reduce the dimensionality of the embeddings. In short, t-SNE is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data.[23] The output is a two-dimensional "map" where semantically similar words are close to each other. Figure 3.2 shows a two-dimensional reduction of the *word2vec* model for Germany using *t*-SNE based on the 300 most frequently occurring words or terms in the rulings.[24]

Similar to a Principle Component Analysis, the dimensions of the map are not identified. Nonetheless, we see that the *word2vec* model successfully learned to group similar words closely together. In the lower right, for instance, all of the names of months like August, January etc. are grouped together. Furthermore, on the upper left, there is a cluster of numbers. In the top left of the map, often appearing numbers of paragraphs are clustered, with the corresponding laws and norms next to it (such as "SGB" (Sozialgesetzbuch), "StGB" (Strafgesetzbuch) or the "BVerfGG" (Bundesverfassungsgerichtsgesetz)). These laws appear often in the context of these numbers that are close. § 90 SGB for instance is a paragraph of the SGB that appears very often in the context of social legislation, which is, in turn, often part of the GFCC jurisdiction. Furthermore, semantically very similar words such as "Auffassung" and "Rechtssprechung" are grouped together (top middle). Given this visual inspection, it appears that the trained judicial decision-specific word embeddings are meaningful and can be used to find the top-n candidate terms in the next step.

Step 2: Find the Top-n Candidates for Expansion

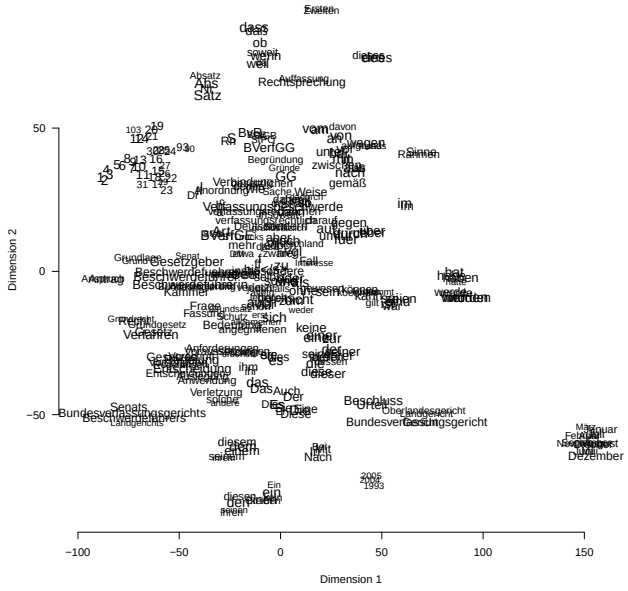
In the second step, for each word in the list of vague terms from LIWC, I consider the top-20 most similar words according to cosine similarity as new candidates for the expanded dictionary. In the context of *word2vec*, word A is said to be similar to Word B

[22] N = 100, C = 5, min count = 5, negative sampling = 0. Different values for N were tried (100, 150, 200, 500). $N > 100$ did not lead to better representations, but increased the computational burden. This is why $N = 100$ was used.

[23] The output of t-SNE is comparable with the output of a simple Principal Component Analysis (PCA), but shows better separation between different groups in the data in some cases.

[24] A similar graph for the French embeddings is in Appendix B.1

Figure 3.2 – Two Dimensional Reduction of the German word2vec Model Using t-SNE



Note: This figure shows a two dimensional reduction of the German word2vec model using t-SNE for the 300 most frequently occurring words in the German court rulings. Although the dimensions are not identified, it is evident that the word2vec model successfully groups similar words closely together.

if "1) A could be used interchangeably for B, or 2) A appears in a similar context as B" (Theil, Štajner and Stuckenschmidt, 2018*a*, 6). Cosine similarity is a common measure to find similar words (Mikolov, Yih and Zweig, 2013, 749), and is defined as:

$$similarity(w_1, w_2) = \frac{\overrightarrow{w_1} \cdot \overrightarrow{w_2}}{||\overrightarrow{w_1}||||\overrightarrow{w_2}||} \tag{3.5}$$

for all similarity calculations in the embedding space, where w_1 and w_2 can be any word from the vocabulary. The following list gives some exemplary similarity candidates in the trained GFCC[25] embeddings for different seed words according to close cosine similarity:

- **Regierung**: tragende Mehrheit, Exekutive, Koalition, Antragsgegnerin, politische Parteien
- **Richter**: Anwalt, Rechtsanwalt, Verteidiger, Notar, Spruchkoerper
- **Gestaltungsraum**: Gestaltungsspielraum, Spielraum, Ermessensspielraum, Entscheidungsspielraum, Beurteilungsspielraum, Entscheidungsraum, Einschätzungsraum,

[25] A French example would be the seed term "juge", with close candidates such as "procureur"(prosecutor) or "avocat" (lawyer).

> Ausgestaltungsspielraum, Einschätzungsspielraum, Gestaltungsfreiraum, Typisierungsspielraum

Especially the last term is a good example of how word embeddings can support human expert judgment. The GFCC typically uses a term like "Gestaltungsraum" to indicate that the legislator has a large leeway:

> "Der Gesetzgeber wird unter Nutzung seines weiten **Gestaltungsspielraums** zu entscheiden haben, auf welche Weise er den verfassungswidrigen Zustand beseitigt. (BVerfGE 110, 33)[26]"

However, the GFCC not always uses the term "Gestaltungsspielraums", but rather many very similar word analogies. An expert might have identified a couple of these similar terms by manually reading through a large number of decisions. Nonetheless, it is questionable whether all related terms, often just different by nuances, would have been found. However, a good domain-specific dictionary must capture all of these terms, and word embeddings provide all of these similarity candidates on an objective basis.

In order to expand the initial LIWC dictionary I use the twenty most similar words as expansion candidates according to cosine similarity. Twenty is an arbitrary value, although it ensures that a sufficient number of candidates is considered. Using a $n > 20$ reduces the quality of the candidate terms as they become less similar to the seed. These candidate terms are then manually filtered and added to the list of seed terms to obtain the expanded dictionary.

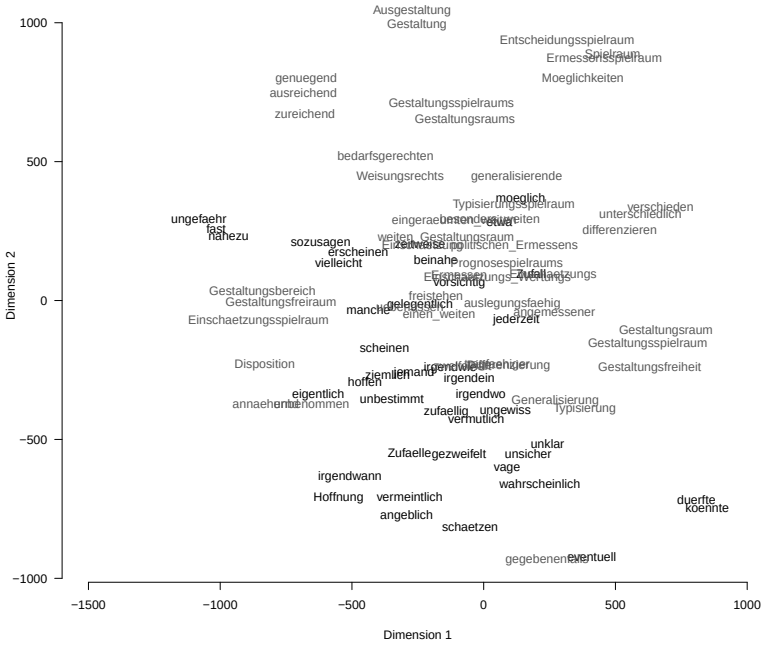
The original German LIWC dictionary contains 57 vague words in their inflicted form, and 49 candidates terms are added.[27] The original French LIWC dictionary contains 48 vague words, and I added seventeen candidate terms. The expanded German judicial policy implementation vagueness dictionary now contains legal domain-specific terms such as "*auslegungsfähig*" (open to interpretation, discretionary), or "*Ermessensspielraum*" (latitude of judgment, discretion), which have not been in original LIWC dictionary before. Word embeddings are also able to find semantically similar terms consisting of more than a single word. The French expanded dictionary now contains words such as "*d'une façon générale*" (generally speaking), or "*interprétation large*" (wide interpretation), while the original French LIWC dictionary only includes terms consisting of single words. The full list of words for both dictionaries can be found in Appendix B.1.

Figure 3.3 shows a two-dimensional mapping using t-SNE of the terms of the original LIWC dictionary (black) and the expansion candidates (gray) for the German appli-

[26] Own idiomatic translation: "The legislator will have to decide, using its wide leeway in decision-making, how to eliminate the persisting unconstitutionality".

[27] I removed fourteen terms of the original LIWC dictionary because these terms are not meaningful in the context of judicial decisions, and also had no reasonable expansion candidates. Such words are, for instance, "Vorahnung", or "Glueck".

Figure 3.3 – Two-Dimensional Mapping of Original LIWC and Expansion Terms, GFCC



Note: This figure shows a two-dimensional reduction of the German word2vec model using t-SNE for terms of the initial LIWC dictionary (black) and the expansion terms (grey).

cation. Note that the dimensions are not meaningful nor identified, but can be used to investigate how close or distant the original seed terms and expansion terms are positioned to each other. The smaller the distance between two words in the graph, the more similar they are semantically. The graph illustrates that using word embeddings, I am indeed able to identify semantically similar words. In the bottom right, for instance, the word "eventuell" (possibly) from the original LIWC dictionary is complemented by the similar word "gegebenenfalls' (if applicable), what makes intuitively sense. The very legal word "auslegungsfaehig" is close to "freistehen" and "angemessen" (center of the figure), and is a good example of how the expansion approach helps to incorporate more domain-specific terms.

Step 3: Applying the Judicial Domain-specific Dictionary to Court Rulings

Now that I have the judicial domain-specific list of vague words, I apply these to the decision texts of the German and French constitutional court. Before, I lemmatized the terms in the expanded dictionary and the decision texts.[28] Lemmatization refers to the process of grouping together the inflected forms of a word into a single item using the lemma or lexeme of a word (Manning and Schütze, 1999, 132). For example, the

[28]Lemmatization was carried out using Python and the library *spacy* (`https://spacy.io/models/de`).

English verb “to decide” appears as “decide”, “decided”, “deciding”, “decides” etc., but all inflected forms have the same lemma, which is “decide”. In my application, lemmatization is superior to simple stemming (cutting words to their root) because lemmatization also takes into account the context of a word; the word “meeting” for instance can be either the base form of a noun or it can be the verb “to meet”. Unlike stemming, lemmatization attempts to select the correct lemma depending on the context by using a dictionary-lookup and rules obtained from pre-trained models. Moreover, in my application to the French decisions, stemming would lead to fundamentally incorrect matches and biased measures. The reason for this is that the French LIWC dictionary contains the word “questionner” (to call something into question), which becomes “quest" when being stemmed. In the CC decisions, they often use the word “la question” (the question), which has the same stem: “quest". Hence, using the stemmed version of dictionaries and decision texts would lead to a bias measurement of the text’s vagueness, because every “question" would be matched as a term of the vagueness dictionary.

I restrict the application of the dictionaries to all decisions dealing with the revision of governmental laws or statutes. This case selection is guided by theory, as there is no reason to believe that constitutional courts use judicial policy-implementation vagueness in decisions which, for instance, deal with formal mistakes of a lower court or related issues. For the German court, I therefore use all 875 decisions dealing with a federal- or state law[29] available between 1972 and 2010. I did not apply the expanded dictionary to the complete decision body, but only to the part of the decision where judges actually justify their decision and therefore, where judicial policy implementation vagueness might appear. This section is the so-called “*B.Part*” in a decision body. All other text passages before this section are dropped (including the header of the decision and the parts where the context of a decision is described). Separate opinions are also removed, as they do not reflect the language of the court but of individual judges. The GFCC often groups multiple proceedings with a similar issue in one main decision, with only one decision text. This means that it is, therefore, not possible to obtain vagueness scores on the proceeding level, but only for each of the decisions.

For the French court, I only score decisions of the type *DC* (Contrôle de constitutionnalité des lois ordinaires, lois organiques, des traités, des règlements des Assemblées). These are the only decisions that directly deal with laws (Hönnige, 2007, 2009). The data was automatically collected from the website of the Conseil[30] and contains 1,286 decisions decided between 1974 and 2018. The year 1974 is chosen because the Conseil only became a fully developed constitutional court after this point in time. In order to

[29] This data was obtained from the CCDB. I only use decisions which deal directly (unmittelbar) with a law; I do not use decisions which only indirectly concern a law (mittelbar).

[30] `https://www.conseil-constitutionnel.fr/`, accessed 15.04.2019.

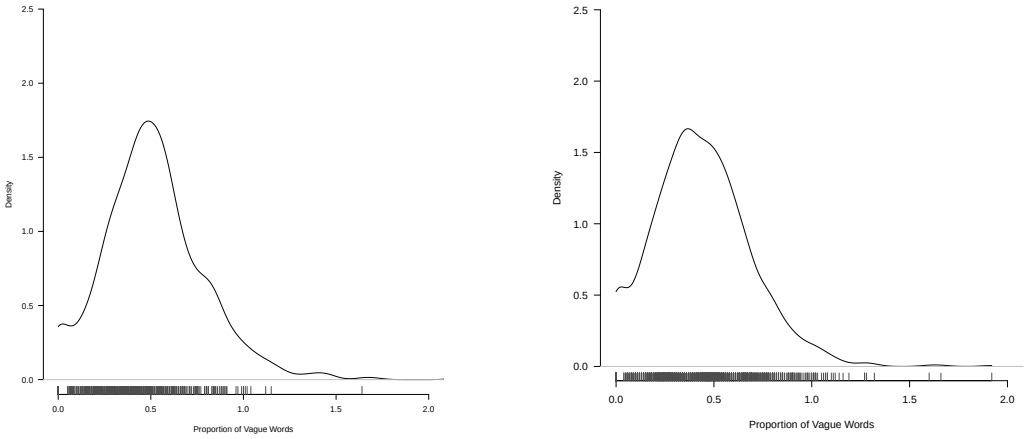
clean the French decision texts, I only dropped the header where the decision contexts is summarized, and kept the remaining text body. This is because the French judges can give policy implementation recommendations in every section of their decision. The processed texts of the German decisions finally contain 2,167 words on average, and 2,422 words for the French ones. In both text data, I did not lowercase the words since capital letters carry semantic meaning in both German and French.

In order to obtain the vagueness scores for each decision text, I split the texts into a set of sentences, and each sentence into a set of word tokens. Sentences with five tokens or less were dropped, because they are often noisy. A word token is deemed vague if it is included in the list of vague terms of the expanded dictionaries. The final vagueness measure is then on the word-level: it is the proportion of words matching one of the terms in the expanded dictionary proportional to the overall number of words in a decision,[31] where a higher percentage of vague words indicates a more vague decision.

Figure 3.4 shows the distribution of the proportion of vague words for decisions of the GFCC (left side) and the CC (right side). The average vagueness score for the GFCC texts is 0.36%, with a standard deviation of 0.21. The average vagueness score of the CC's decision is 0.43% with a standard deviation of 0.25. For both courts, there is also a number of decisions where not a single word of the vagueness dictionary appears. Although this numbers appear relatively small, they are yet meaningful: as Eichorst and Lin (2019) highlight, measuring vague language requires "counting specific, individual words within large documents of language", and that "although a single modifier can shift interpretation [...], its empirical approximation will be relatively small and difficult to discern absolutely" (Eichorst and Lin, 2019, 24). The overall small proportion of vagueness is also comparable with the findings from other studies (see Liu, Fella and Liao, 2016; Theil et al., 2018; Eichorst and Lin, 2019).

[31] One could also use the proportion of vague sentences of a decision, namely the proportion of sentences containing at least one vague token proportional to the overall number of sentences of a decision. However, this method is sensitive to outliers (very short or very long sentences), and also harder to compare with the French findings due to difference in writing style. This is because the French court, uses only a few but very long sentences, hence producing relatively high percentages of vague sentences. The proportion of vague words and the proportion of vague sentences of a text are significantly correlated with .84 using Pearson's r, $p < 0.01$.

Figure 3.4 – Distribution of Proportion of Vague Words, GFCC and CC Decisions



(a) Proportion of vague words per decision, GFCC

(b) Proportion of vague words per decision, CC

Note: Kernel density estimation of the proportion of vague words/vague sentences in GFCC decisions.

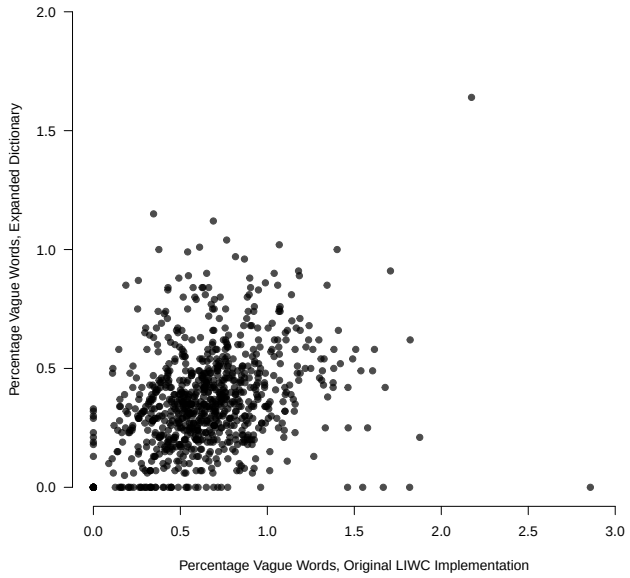
In order to illustrate how the word choice modifies the informative structure of a sentence, Table 3.1 provides some examples of how vague words are used by the GFCC. Vague words in each sentence are boldfaced. It is evident how vague words shape the decision leeway of the legislator, and thus reflect judicial policy implementation vagueness.

Table 3.1 – Vague Text Examples from German Decision Texts

Vague Text Examples
Auf dieser Grundlage darf er [der Gesetzgeber] **generalisierende**, **typisierende** und pauschalierende Regelungen treffen, ohne wegen der damit unvermeidlich verbundenen Härten gegen den allgemeinen Gleichheitssatz zu verstossen. (BVerfGE 112, 268)
Es liegt insoweit im **Gestaltungsspielraum** des Gesetzgebers, ein für gesetzliche wie private Versicherte im Grundsatz einheitliches, alle wesentlichen Krankheitsfälle befriedigend abdeckendes Versorgungsniveau festzulegen. (BVerfGE 123, 186)
Vielmehr hat der Gesetzgeber einen **weiten Gestaltungsspielraum**, welche Pflichten zur Sicherstellung von Gemeinwohlbelangen er Privaten im Rahmen ihrer Berufstätigkeit auferlegt. (BVerfGE 125, 260)

I want to emphasize again that the original LIWC dictionary would not have identified the vague terms in these sentences. Terms such as "Gestaltungsspielraum" are judicial domain-specific words and do thus not appear in the original LIWC dictionary. In

Figure 3.5 – Scatterplot of the Proportion of Vague Words, LIWC versus Expanded Dictionary, GFCC



order to illustrate how the measurement of the expanded dictionary diverges from the original LIWC approach, I scored the same decision texts but this time using the original LIWC-implementation. The measurement level of LIWC is also the same as for the expanded dictionaries, namely the proportion of vague terms in a text. Figure 3.5 shows a scatterplot of the expanded dictionary scores on the y-axis and the original LIWC scores for the same German decision texts on the x-axis. Using the expanded dictionary yields, for the same text, to quite different vagueness scores than the original LIWC dictionary (indicated by a rather low correlation coefficient of 0.32 using Pearson's r). The expanded dictionary finds less vague words than the LIWC implementation.[32] Consequently, research solely relying on LIWC would miss-classify a large proportion of the court decisions, and thus potentially draw incorrect inferential conclusions. Later in this chapter, I will further elaborate on the validity of the expanded dictionaries in comparison with LIWC.

3.3.4 Method 1 Summary

In this sub-chapter, I have demonstrated how word embeddings can be used to create a judicial domain-specific dictionary of vague terms. For this purpose, I trained

[32]There is also a number of decisions where the expanded dictionary measures zero vague terms, while the LIWC dictionary measures up to 2.8%. This is because I removed some of initial words in the LIWC dictionary because they are neither meaningful in the context of judicial decisions nor did they have reasonable expansion candidates.

word embeddings on all decision texts of the GFCC and the CC, used the words of the original LIWC dictionaries as seed terms to find word analogies, and finally applied them to a corpus of German and French decision texts. Because the expanded dictionaries are tailored to the legal context, they are better suited to identify judicial policy implementation vagueness in texts where general dictionaries fail.

3.4 Method 2: Training a NLP Vagueness Classifier

In this section, I develop a binary sentence classifier designed to automatically detect vague sentences in GFCC decisions.[33] In particular, I draw on recent advances in machine learning to train and evaluate the predictive performance of different popular NLP classifier on a corpora of over 3,500 manually annotated sentences randomly drawn from published decisions of the GFCC. I make three contributions: first, I collect a novel data set and evaluate whether and how good the classification of vague terms works via an algorithmic procedure. Second, I benchmark the predictive performance of different standard machine learning and deep learning classifiers on this data set and compare their predictive performance. Third, I compare the performance of Method 1, the expanded dictionary, with the best machine learning classifier and discuss the strengths and pitfalls of both approaches.

To foreshadow, I find that overall machine learning classifiers perform considerably well in classifying the latent concept of judicial policy implementation vagueness. Deep learning classifiers do not perform distinguishably superior in terms of the raw performance metrics, but are better able to classify unseen text instances than standard machine learning classifiers.

3.4.1 Corpus Construction and Annotation Procedure

No annotated data set exists that I could use to train the different classifiers. For this reason, I collect a new vagueness data set consisting of 3,581 sentences which were randomly sampled from all published decisions of the GFCC from 1972 to 2010, and then manually annotated. Only the text sections where judicial policy implementation vagueness can appear (the "B. Part") are used to sample from. Introductory sentences, habitual utterances and sentences shorter than four words are excluded from the sampling process, as they would dilute the results of the classification task.

The annotation was performed by three German native annotators. All of them possess domain-specific knowledge of law and social science.[34] A more detailed

[33] Due to the large effort it requires to gather the necessary training data, it was not possible to develop a similar classifier for the French application.

[34] One of them has a minor in public law. The other two annotators have had at least two public law classes in their undergraduates.

description of the annotation task with annotations instructions and more examples can be found in Appendix B.2. Each of the annotator had to manually classify all of the 3,581 sentences as either a vague sentence or a not-vague one. The following two sentences again highlight the difference between the two types:

- ***Vague:*** *'Denn der Gesetzgeber hat, wie aufgezeigt, mehrere Möglichkeiten, um den verfassungswidrigen Zustand zu beseitigen.'*[35] (BVerfGE 121, 175)
- ***Not Vague:*** *'Der Gesetzgeber hat sowohl den datenerhebenden als auch den datenempfangenden Behörden eine Kennzeichnungspflicht aufzuerlegen.'*[36] (BVerfGE 100, 313)

Additionally, the annotators also had to highlight the words or expressions that triggered their decision-making. I use this information to check whether there are frequently-occurring terms that the expanded dictionary does not yet contain.

3.4.2 Inter Annotator Agreement

The pairwise inter-annotator agreement (IAA) for labeling sentences as vague or not vague is shown in Table 3.2.

Table 3.2 – Pairwise Inter-Annotator Agreement

Annotators	Agreement	κ
1 & 2	96.95	0.63
2 & 3	98.5	0.79
1 & 3	98.45	0.84
Average	97.97	0.75

Note: Agreement = percentage of cases in which both annotators assigned the same class; κ = Cohen's Kappa.

Cohen's *Kappa* (κ) (Cohen, 1960), is between $0.63 \leq \kappa \leq 0.84$, depending on the annotator pair. The average κ of 0.75 can be, depending on the source, considered as "substantial" (Landis and Koch, 1977, 165) or "excellent" (Fleiss, Levin and Paik, 1981, 609). This is similar or better to the IAA achieved in related annotation tasks.[37] The achieved IAA suggests that the task of marking vague or not vague sentences in court rulings is equally difficult for humans as the same task on Wikipedia sentences, but seems to be more difficult than marking sentences in biomedical texts. Here, human

[35] Own translation: "The legislator has, how illustrated, multiple options to remediate the unconstitutional state."

[36] Own translation: "The legislator has – for the data receiving and the data collecting agencies – the obligation to label them."

[37] Štajner et al. (2016) for instance classify sentences of central bank transcripts using three annotators, and achieve a $0.56 \leq \kappa \leq 0.61$. Theil et al. (2018) achieve a κ of 0.73. (Ganter and Strube, 2009) achieve a $0.45 \leq \kappa \leq 0.80$ on the Wikipedia data set.

annotators achieved a $\kappa = 0.98$ (Medlock and Briscoe, 2007). For the final classification of a sentences, the majority vote was taken. This means that if two annotators disagreed on the classification, the third annotator's coding was pivotal. Overall, from the 3,581 annotated sentences, 4.2% (154) were annotated as vague. This strong class imbalance is comparable with other vagueness data sets (e.g. Theil, Štajner and Stuckenschmidt, 2018*b*; Liu, Fella and Liao, 2016).

3.4.3 Experimental Set-up

I randomly divide the annotated data set into two smaller data sets: one training set (2,864 sentences, 80% of the total data) and one test set (717 sentences, 20% of the total data). Stratified sampling is used to ensure that all data sets contain the same ration of sentences marked as vague and not vague. Ignoring the strong class imbalance would negatively affect the quality of the classifiers. The test set is never used in the training or any other validation procedure, and is strictly hold out-of-sample.

In my experimental set-up, I use two different sets of algorithms. The first set contains a wide range of different "conventional" NLP classifiers, namely support vector machines (Cortes, Vapnik and Saitta, 1995), naive bayes, random forests (Breiman, 2001), extreme gradient boosting and logistic regression. All of these algorithms have shown their value in various NLP tasks. The second set of algorithms draws on recent advances in deep learning, where particular types of neural networks for sequence classification have shown promising results for text classification.

For the first set of algorithms, the conventional NLP-classifiers, I follow standard NLP practice and use Bag-of-Words[38] vectors of sentences with the commonly used *Term Frequency-Inverse Document Frequency* (tf-idf) term-weighting as training features.[39] Tf-idf uses the raw count of a term in a document, and then weights this count by the number of occurrences of this term in other documents. More formally, the final weight w_i for a term i in document j is defined as:

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \tag{3.6}$$

where $tf_{i,j}$ is the number of occurrences of a term i in document j, df_i is the number of documents containing i and N is the total number of documents. Accordingly, less weight is given to very frequent (and thus, probably not very discriminative) terms and more weight to rare terms. Before tf-idf is applied, I use the Porter-Stemmer (Porter,

[38]Bag-of-Words is a simple way of representing text data by encoding the term frequency of a certain word of a vocabulary in a sentence into a vector.

[39]Only the 15,000 most frequently occurring terms (including uni-grams, bi-grams, and tri-grams) are used to improve computational efficiency.

1980) and removed stop-words.[40] Both increases the computational efficiency of the training because text redundancies are removed. For each of the algorithms, the hyper-parameters are optimized with respect to the area under the precision recall (AUC-PR) curve[41] via grid-search using stratified five-fold cross-validation on the training set. The model with the optimal hyper-parameter performance on the training set was then evaluated on the test set.[42]

It is important to note that hyper-parameter tuning and cross-validation in one procedure requires an out-of-sample test set to measure the so-called true error. The true error is a measure of how well a model can predict outcomes of previously unseen data (Efron and Hastie, 2016; Cranmer and Desmarais, 2017). An estimate of true error is important in practice, as it allows one to check whether a model generalizes well to unseen data or just memorizes the patterns in the training data (i.e. over-fitting). Using out-of-sample data to evaluate the predictive performance of a model allows me to obtain an estimate of true error despite cross-validation is used for hyper-parameter tuning. Hence, the procedure outlined here does not suffer from a problematic cross-validation described in Neunhoeffer and Sternberg (2019).

The second set of algorithms I explore originate from the family of deep learning algorithms. Specifically, I use two different types of deep learning classifiers, a *Gated Recurrent Unit*, a particular type of a Recurrent Neural Net, and a *Convolution Neural Network*. Both neural network types have recently shown impressive performances in sequence classification tasks and have outperform conventional NLP algorithms in several binary sentence classification.[43] (Kim, 2014; Tang, Qin and Liu, 2015; Wu et al., 2015; Lee and Dernoncourt, 2016; Yin et al., 2017). Because deep learning algorithms are still relatively unknown in the field of social sciences, I will give a very brief summary of both network's architecture in the following. A more detailed description of the architecture of both network types can be found in Yin et al. (2017) and Allaire and Chollet (2017, Chapter 6), or in the two excellent overview papers of Lopez and Kalita (2017) and Young et al. (2018).

A *Gated Recurrent Unit*, introduced by Cho et al. (2014), belongs to the family of *Recurrent Neural Networks* (RNNs). These types of networks have proven to be particular efficient with sequential data, and text analysis in particular (Tang, Qin and Liu, 2015;

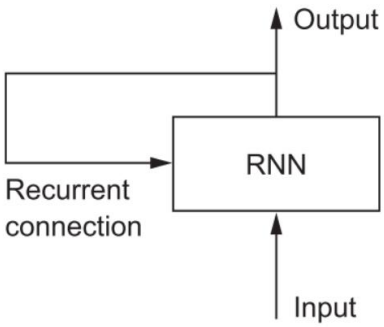
[40] For both I used the *Python* 3.6 NLTK 3.3 implementations. Classification did not improve by not removing the stopwords and stemming. Classification did also not improve using lemmatization instead of stemming.

[41] Cranmer and Desmarais (2017) recommend using the AUC-PR in imbalanced classification scenarios (Cranmer and Desmarais, 2017, 154).

[42] The *Python* Scikit-learn 0.20 (Pedregosa et al., 2011) machine learning library was used to train and evaluate the models.

[43] As of end-2018, these are the two most popular deep learning algorithms for binary sentence classification. The *R* tensorflow keras implementation was used. Computation was carried out on a Nvidia 1060 GTX GPU.

Figure 3.6 – Simplified Architecture of a Recurrent Neural Network



Note: A simple RNN as a network with an internal loop. Figure taken from (Allaire and Chollet, 2017, 180).

Dauphin et al., 2017; Rosenthal, Farra and Nakov, 2017). In order to understand the architecture neural networks for text classification, it is first necessary to explain the notion of sequential data and why it needs a particular network architecture.

Sequential data refers to data types such as time series data (e.g. stock markets, sensor data, weather forecast), but also to texts as a sequence of words or characters. Densely connected networks (standard neural networks), called feed-forward networks, have no "memory". Each input is propagated straight through from the input layer over the hidden layer to the output. In other words, the order of the data, namely the notion that related things follow each other, is completely ignored. Therefore, in standard neural networks the data is assumed to be *iid* (independent and identically distributed). By contrast, readers of this dissertation do not read every written word independently. While reading, every word is processed one after another, such that the meaning of a word is the result from the words read before.

RNNs are designed to deal with sequential data. A recurrent neural network is called recurrent because it performs the same computation for every element of a sequence, where the output is dependent on the prior computations. Put differently, a RNN iterates over the sequence elements (like an internal loop) and maintains a *state* containing information relative to what it has processed before (Allaire and Chollet, 2017, 180). This is illustrated in Figure 3.6 again. Another way of thinking about an RNN is to think about an "unfolded" network, where unfolding refers to writing out the network for the complete sequence. A sentence consisting of three words, for instance, would be unfolded into a three-layer network, with one layer for each word. Stated more generally, a RNN can map an input sequence with an arbitrary number of elements x_t at time t into an output sequence O with elements O_t, with each O_t depending on all of the previous $x_{t'}$ for $t' < t$ (Lecun, Bengio and Hinton, 2015, 442). I included a more detailed description of the network's architecture and the necessary computational steps in Appendix B.3.

The network is trained and optimized via a variant of backpropagation suited for sequence data, the so-called backpropagation through time using stochastic gradient descent. Because the weights in a RNN are shared by all time steps in the network, the calculation of the weights at time step t depend on the weights of time step $t-1$). This is why the gradient at each output depends not only on the calculations of the current time step, but also the previous time steps. This often causes vanilla RNNs to suffer from a problem called *vanishing gradient problem* (Bengio, Simard and Frasconi, 1994; Hochreiter, 1998). Put simply, the vanishing gradient problem refers to the problem that while conducting back-propagation, the gradients tend to becoming increasingly smaller the more one moves back in the network. This negatively affects the learning of the early layers in a network, which only learn very slowly. In a word sequence classification example, this would mean that it is difficult for the network to memorize words from far away in the sequence.

For this reason, I follow the current NLP state-of-the-art and use a Gated Recurrent Unit (GRU), a special type of RNN. GRUs are designed to combat the vanishing gradient problem through a gating mechanism, namely an *update gate* and a *reset gate*. This gating mechanisms allows the GRU to decide whether the current time-step information matters or not, and to control the information inflow from past steps. In short, the reset gate determines how much of the past information the network should forget. The update gate defines how much of the past information (from previous time steps) should be passed along, and how to combine this information with new, incoming information from the current time step. The special thing about GRUs is that using these gating mechanisms, the networks are able to adaptively remember and forget. A short introduction to the network's architecture is in Appendix B.3. Interested readers should also be hinted to the original paper of Cho et al. (2014). In order to illustrate why these gating mechanisms are so powerful, consider the following example (toy) sentence:

> The German Federal Constitutional Court, given the prevailing case law and weighting to the respective interests of the public and of the parties concerned, decides that the *legislator has a considerable decision leeway in the organization of the inheritance tax*.

In this case, the important information for the classifier, namely whether the legislator has discretion or not, is contained in the last part of the sentence. In that case, the network can learn to set the reset gate such that it "washes out" early and redundant information. By contrast, consider another example (toy) sentence:

> *The legislator has a considerable decision leeway in the organization of the inheritance tax*, accordingly, the German Federal Constitutional Court, given the

> prevailing case law and weighting to the respective interests of the public and of parties concerned, rejects the complaint.

Here, the important information is at the beginning of the sentence. In such a case, the GRU can learn to set the update gate such that a majority of the previous information is kept when continuing to iterate of the sequence of words of the sentence. These gating mechanisms thus help to deal with the vanishing gradient problem, because at every single time step the network can decide to keep relevant information and forward it to the next step. I decided to use a GRU rather than other popular RNN variants such as Long Short Term Memory (LSTM) networks because GRUs are computationally more efficient to train and, more important, have shown to exhibit better performance on small data sets such as I use (Chung et al., 2014; Jozefowicz, Zaremba and Sutskever, 2015). The final GRU used for the classification of my data set comprises of three stacked GRU layers, with 64, 32 and 16 units each and a drop-out rate of 0.5[44] and recurrent drop-out rate of 0.2. The input features are the training sentences encoded into a 256 dimensional word embedding space.

The second type of neural network architecture used in the experiment is a so-called *Convolutional Neural Network* (CNN). While originally used in the context of computer vision for image recognition, CNNs are increasingly used in the NLP context (see Kim, 2014). Such CNNs can be competitive with RNNs and GRUs on certain sequence classification problems, but usually come at a lower computation cost (Allaire and Chollet, 2017, 209). The main idea behind a CNN is that it extracts local patches (convolutions/sub-sequences) from a text sequence. This can be imagined as sliding a "window" of a certain size over an input sequence.[45] A CNN with a convolution windows of size five is able to learn word sequences of length five or less, and it can recognize these words in any context in an input sequence. In other words, a pattern learned at a certain position in a sentence can be recognized at a different position in another sentence. This is called translation invariance (Allaire and Chollet, 2017, 209). I provide an overview of a simple CNN architecture for sentence classification in Appendix B.3. For a more detailed description of CNNs for sentence classification see the original paper of Kim (2014). The final CNN used for the classification is similar to Kim (2014) and comprises three convolutional layers with a kernel size (window size)

[44] Over-fitting is a serious problem of deep neural networks, as usually hundreds of thousands of parameters are estimated and only a limited number of training data available. Drop-out in neural networks describes the process of randomly dropping out (setting to zero) a number of units from the neural network during training. Drop-out has shown to significantly reduce over-fitting and be superior to other regularization methods (Srivastava et al., 2014). Recurrent drop-out is a particular drop-out for recurrent layers, and specifies the drop-out rate of the recurrent units (Gal, 2016).

[45] In praxis, this sequence (e.g. a sentence) is encoded into an embedding matrix. A five word sentence using a 100-dimensional embedding matrix would result in a 5×100 input matrix (see Figure B.5 in the Appendix.).

of 3, 4 and 5 with 100 feature maps each, and a drop-out rate of 0.5. The input features consist of each sentence encoded in a 256-dimensional word embedding space.

3.4.4 Classification Results

I evaluated all algorithms using the overall accuracy (i.e. the percentage of correctly-classified instances), precision, recall, F_1-score (harmonized mean of precision and recall) and receiver operating characteristic area under curve (ROC AUC). Such an extensive set of performance measures is necessary because using accuracy alone in cases with highly imbalanced data leads to miss-leading performance evaluations (Cranmer and Desmarais, 2017, 152). A definition of these performance measures is in Appendix B.3.1. The baseline to compare each algorithm against is a naive classifier that always assigns the majority class, namely a classifier that classifies each sentence as not vague. Therefore, by definition there are no true positives (correctly classified vague sentences), which is why precision[46] and recall[47] of the majority classifier are zero. A *TruePositive* is a vague sentence that is correctly classified as vague, and a *TrueNegative* is a not vague sentence which is correctly classified as not vague. The accuracy of the naive classifier is the percentage of not vague sentences.

The results of the experiment are presented in Table 3.3, together with the majority class baseline. The reported predictive performance is strictly evaluated out-of-sample on the test set. None of the test set observations have been used at any point in the training process. The corresponding confusion matrices for each of the classifier are in Appendix B.3.2.

Table 3.3 – Results of the classification task for the test data. The best results are presented in bold. The majority class (baseline) is always classifying not vague.

	Precision	Recall	F_1	Accuracy	ROC AUC
Support Vector Machines	**0.76**	0.29	0.42	94.98	64.15
Random Forest	0.67	0.13	0.22	94.1	56.44
Naive Bayes	0.31	0.24	0.27	91.1	60.44
Logistic Regression	0.50	0.53	**0.52**	93.72	74.88
XGBoost	0.24	0.73	0.36	83.4	**78.71**
Convolutional Neural Net	0.27	**1.00**	0.42	**95.4**	63.33
Gated Recurrent Unit	0.27	0.71	0.39	94.7	62.96
Expanded Dictionary	0.60	0.20	0.29	82.01	71.74
Majority Class (not vague)	0	0	0	95.68	50.0

The highest precision is achieved by the SVM, albeit at the cost of relatively low recall. The highest recall (maximum score of 1.0) is achieved by the CNN, which means that the CNN does not classify one truly not vague sentences as vague (zero false negatives).

[46] Defined as $\frac{TruePositive}{TruePositive+FalsePositive}$.
[47] Defined as $\frac{TruePositive}{TruePositive+FalseNegative}$.

However, this comes at the cost of a relative low precision. Furthermore, it is interesting to note that the simple logistic regression achieves the best F_1 score[48], resulting from balanced precision and recall scores. In order to investigate the classification results in further detail, consider the following obviously-vague sentence:

> *Denn der Gesetzgeber hat, wie aufgezeigt, mehrere Möglichkeiten, um den verfassungswidrigen Zustand zu beseitigen.*[49] (BVerfGE 121, 175)

Both types of classifiers (the conventional classifiers like random forest and the neural networks) correctly assign high probabilities that this sentence belongs to the vague class. By contrast, all classifiers have problems of correctly classifying less obvious vague sentences with a complex structure such as:

> *Der Gesetzgeber kann die Vielfalt der Faelle, die er mit seiner Regelung erfasst, nicht im vorhinein erkennen und muss sich deswegen mit Einschaetzungen zufriedengeben.*[50] (BVerfGE 97, 186)

Both types of classifier assign rather low probabilities of this sentence being vague. This sheds light at one broader result of the classification task with respect to the positive class: even the best model only correctly identifies 33 out of 45 vague sentences (the *Xgboost* model, which also produces a large number of false positives, indicated by the low precision) in the test data. It is important to note that the sentences that are difficult to classify for the algorithms are also the same sentences with which the human coders had the largest disagreement on, meaning that different annotators labeled the same sentences differently. Given the sentence above, for instance, one of the human annotators disagreed with the other two and annotated this sentence as not vague, because it lacks a clear vagueness indicator. If even trained human annotators disagree on the classification, then an algorithm will most likely also have difficulties predicting the correct label.

Overall, when comparing the predictive performance of the traditional NLP classifiers with the deep learning algorithms, there is no large difference between the predictive power of both. Why are the neural nets not considerably better than most of the simple Bag-of-Words classifier, although they come at a much larger computational cost? The reason for this is that for classifying judicial policy implementation vagueness, there are certain trigger words (like those used in the expanded dictionary) which often indicate whether a sentence might be vague or not. Thus, the Bag-of-Word approaches (which

[48] F_1 score is the harmonic mean of precision and recall.

[49] Own translation: "The legislator has, how illustrated, multiple options to remediate the unconstitutional state."

[50] Own translation: "The legislator cannot anticipate the diversity of all possible cases with its regulation in advance, and must therefore come up with an assessment."

ignore the order of the words in the sentences) perform remarkably well because the information whether a trigger word is in a sentence is already sufficient for identifying certain vague sentences, not where exactly the trigger word is located in the sentence.

However, the deep learning classifiers are superior in capturing the semantic context of a sentence. Table 3.4 shows two toy sentences in the first column. The first sentence includes a negation and is not vague and the second does not contain a negation but is vague. The second column shows the probability of this sentences belonging to the vague category as classified by the GRU. The third column shows the same probabilities but for the logistic regression classifier.

Table 3.4 – Vague Text Examples from German Decision Texts

Sentence	GRU probability vague	Logit probability vague
Der Gesetzgeber hat **keinen** weiten Gestaltungsspielraum.	0.05	0.41
Der Gesetzgeber hat **einen** weiten Gestaltungsspielraum.	0.86	0.27

Table 3.4 illustrates that the GRU is capable of capturing the semantic meanings of sentences, such as simple negations. Although both sentences only differ in one word (in fact, only one character), the GRU correctly assigns are very low probability of being vague to the first sentence and a very high probability for the second one. By contrast, the logistic regression assigns rather low probabilities of being vague to both sentences. It even assigns a higher probability of being vague to the first, not-vague sentence.

Furthermore, it is interesting to note that the expanded dictionary correctly identifies 27 out of 45 vague sentences. However, the bad news is that the dictionary also produces a large number of false positives (111) on the test set. This means the expanded dictionary classifies sentences which are not annotated as vague incorrectly as vague. The reason for this is that the dictionary cannot capture any semantic meaning, but simply checks whether a sentence contains a certain pre-specified word of the dictionary or not. For instance, consider the following sentence (which is annotated as not vague):

> *Der Gesetzgeber hat bei der Einführung von Sonderabgaben Kompetenzschranken zu beachten, die seinen Gestaltungsspielraum im Verhältnis zur übrigen Regelungsbefugnis in der jeweiligen Sachmaterie deutlich verengen.* (BVerfGE 67, 256)

The dictionary correctly identifies that the sentence contains the trigger word "Gestaltungsspielraum" (room to maneuver), which generally implies a certain decision leeway of the legislator. However, it cannot consider that this decision leeway is "deutlich verengt" (clearly limited). The deep learning classifiers, for instance, do not wrongly

label this sentence as vague (the GRU assigns a probability of zero that this sentence is vague).

Classification Result Comparison With Similar Studies

My experimental results are comparable with those of similar classification tasks using related methodological approaches. First, I compare my results with other classification studies on domains that are also understudied by the NLP community. Theil, Štajner and Stuckenschmidt (2018*b*) classify uncertain statements in financial disclosure documents and achieve F_1 scores between 0.59 and 0.41. However, they leverage a much richer corpus and use an expanded dictionary containing over 4,100 trigger words. Štajner et al. (2017) achieve an F_1 score of 0.5 using a Bag-of-Words SVM classifier in the detection of speculative sentences in the monetary policy domain. These two examples show that my experimental results are comparable with the results of similar NLP applications in understudied areas.

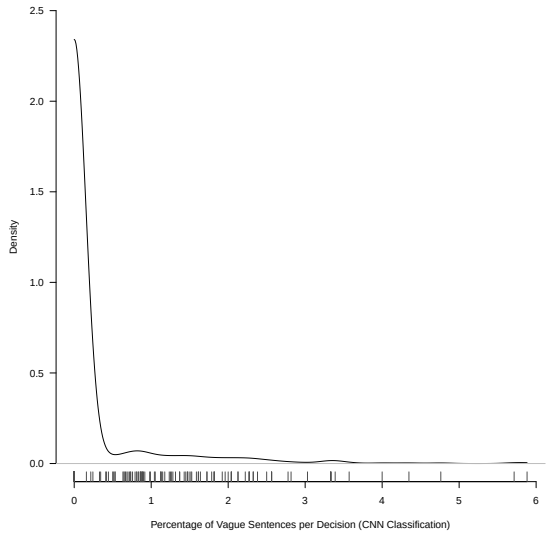
Second, I compare my results to traditionally well-studied domains such as the classification of uncertain statements in Wikipedia entries (the CoNLL-2010 shared task, (Farkas et al., 2010)) and abstracts from the (bio)-medical domain (e.g. Light, Qiu and Srinivasan, 2004). These two areas are well-studied, a large set of different classification approaches exist and there are even yearly competitions on these data sets. The best classifier in the Wikipedia domain achieves a F_1 score of 0.6, whereas the best system for the abstracts achieves a F_1 score of 0.86. This shows that my classification results are at least comparable with those of the Wikipedia data set. It is unsurprising that on average, I do not achieve the same impressive performance scores of these studies, given the sheer amount of training data and prior research available in these areas.

3.4.5 Application to GFCC Court Decisions

In the last step, I used the CNN to predict all texts from the same data set used in the application of the expanded dictionary, e.g. all decision texts of the GFCC dealing with the revision of governmental laws or statutes. I chose the CNN due to its excellent recall and its superiority in classifying unseen sentences, although for instance the logistic regression has a higher F_1 score.

Because the classifier is trained to identify vague sentences, the classification of the texts is undertaken at the sentence level. The final measure obtained by the CNN classifier is then the proportion of vague sentences in a decision, calculated as the number of vague sentences divided by the overall number of sentences. Figure 3.7 shows the distribution of the proportion of vague sentences. Again, the score is rather low with a mean of 1.8% vague sentences, a standard deviation of 0.65, a minimum of zero and a maximum of 6.

Figure 3.7 – Density Plot of the Proportion of Vague Sentences per GFCC decision, CNN Classification



Note: Kernel density estimation of the classification based on the Convolutional Neural Network, measured as the percentage of vague sentences of all sentences in decisions of the GFCC. Thicks on the x-axis show the allocation of all observed/actual percentages.

3.4.6 Method 2 Summary

In this sub-chapter, I have used used two different sets of popular machine learning algorithms for the classification of judicial policy implementation vagueness. In particular, I evaluated the predictive performance of a set of traditional machine learning classifiers and two deep learning algorithms (Gated Recurrent Units and Convolutional Neural Networks) in a binary sentence classification task on a novel, manually annotated corpus of randomly sampled sentences of written decisions of the German Federal Constitutional Court. The best classifier (the CNN) achieves a performance of 0.42 in terms of F_1 score in a strict out-of-sample prediction. This shows that although judicial policy implementation vagueness is an abstract, latent concept, staofte-of-the-art NLP classifier can capture this phenomenon. Furthermore, this study is the first application to judicial texts in German language, and the first which uses machine learning algorithms to classify and predict the occurrence of vagueness in constitutional court decisions. Moreover, the vagueness scores introduced in this chapter will be used in Chapter 2 to test a popular game theoretical model on the strategic use of vague language in court decisions. Therefore, the implications of this classification experiment go beyond the area of NLP, and have practical implications of our ability to study linguistic phenomena in political science.

3.5 Validation

Validity is an obvious concern when new measures are established. This section evaluates the validity of both the expanded dictionary and the CNN classifier results in a comparable manner. For my applications on the German court, I rely on two validity checks. The first validity check uses a data set of Engst (2018) on vague directives[51] in GFCC decisions. The data of Engst (2018) codes directives according to their level of detail, namely how specific or vague these directives are.[52] A vague directive according to Engst (2018) is a directive that, for example, only depicts some minor short-comings of the legislation under review, or can be understood as a reminder for the legislator to keep track on a certain issue (Engst, 2018, 109). In contrast, directives are not considered as vague if they contain a request for the legislator to modify a certain law or even present the legislator different alternatives or suggestions on how a revision could look like. I test the *convergent validity* of the German vagueness measures by using the expanded dictionary and the CNN classifier to obtain the judicial policy implementation vagueness scores for the texts of these directives. The observable implication is as follows:

- Directive texts that were classified as vague by Engst (2018) should receive a higher vagueness score than decisions classified as not vague.

This validity check allows me to compare the extent to which the automatic vagueness detection is comparable with human judgment. Such a validation procedure in the context of machine coding is also recommend by Grimmer and Stewart (2013) as the human gold standards of validation: if my measures are performing well, they will replicate the hand coding; if they perform bad, they will fail to replicate the human coding and introduce error (Grimmer and Stewart, 2013, 279).

My second validity check evaluates the German measure's *face validity*. The study of Staton and Vanberg (2008) explicitly labels certain GFCC decisions as prime examples of a either very vague decisions or very specific decisions. In this regard, the observable implication is as follows:

- Decisions which Staton and Vanberg (2008) describe as prime examples of vague decisions should receive a high vagueness score, whereas decisions depicted as very specific should receive a low vagueness score.

[51] A directive is defined as "a judicial directive is a statement by the judges included in a court's decision, directed at political actors to request an action by those actors to respond to a constitutional issue." (Engst, 2018, 109). I would like to thank Benjamin Engst again for providing me that data.

[52] Engst (2018) distinguishes between implicit directives, explicit directives or explicit directives with a guideline for implementation. In line with Engst (2018), I consider implicit directives as vague directives, and the other categories as not vague directives. The classification in Engst (2018) was carried out by a student assistant with expertise in political science and law.

My next check concerns the expanded dictionary and evaluates the *robustness* of my measure to idiosyncratic features of both German and French. For this, I use bridging observations (decisions available in both languages) to show that the vagueness scores of these decisions are independent of the language they are written in. The implication is as follows:

- If a decision written in French obtains a similar vagueness score using the French expanded dictionary than the same decision written in German using the German expanded dictionary, then the obtained vagueness scores are not an artifact of the language they are written in.

My final check assesses the validity of the original LIWC dictionary. At the beginning of this chapter I argue that using the original LIWC dictionaries can lead to biased measurements. To evaluate this claim I conduct the same validity analyses than for the expanded dictionaries, but this time I use the original LIWC implementation. The implication is as follows:

- If the original LIWC dictionary is not suitable for analyzing judicial policy implementation vagueness, then it should provide poor results in the validity checks.

For decisions of the French Conseil there is, unfortunately, no comparable data set available as the one of Engst (2018). For this reason, I have to limit my validation to an extensive discussion of the face validity of the decision which scored highest in vagueness according to the expanded dictionary. The results of all validity checks are again summarized in Table 3.5. Put simply, I find that my measures of judicial policy implementation are valid and robust, but that the LIWC implementation fails all of the checks.

3.5.1 Convergent Validity of the German Application

I follow the validity check procedure outlined before and leverage data set of Engst (2018). The first validity check evaluates whether directive texts of the GFCC which are coded as vague[53] according to Engst (2018) contain a higher proportion of vague words/vague sentences than directives coded as not vague. For this, all directive texts are scored using the expanded dictionary and the CNN classifier. In fact, using a *t*-test I find that for both measurement approaches, directives coded as vague by Engst (2018) contain a statistically significant ($p < 0.05$, two-sided test) higher proportion of vague

[53]Of 2,006 decisions in the data, 1,725 contain no directive, 114 contain a directive coded as vague, and 165 contain a directive coded as not vague. Only decisions which contain a directive at all are used. Otherwise, the decisions without a directive would lead to highly imbalanced groups, and biased test results.

words/vague sentences than directives which are coded as not vague. The convergent validity is therefore confirmed.

Note that in the above two validity checks, the skewness of the measures (they contain a large number of zeros, namely zero percent vague words/vague sentences) could be an issue since the *t*-test relies on the assumption of normality. In order to check the robustness of the validation checks, I re-run the same analyses but this time I use the Mann–Whitney-U test, which is a non-parametric hypothesis test that does not require the assumption of normality. The findings remain robust and yield to similar results.

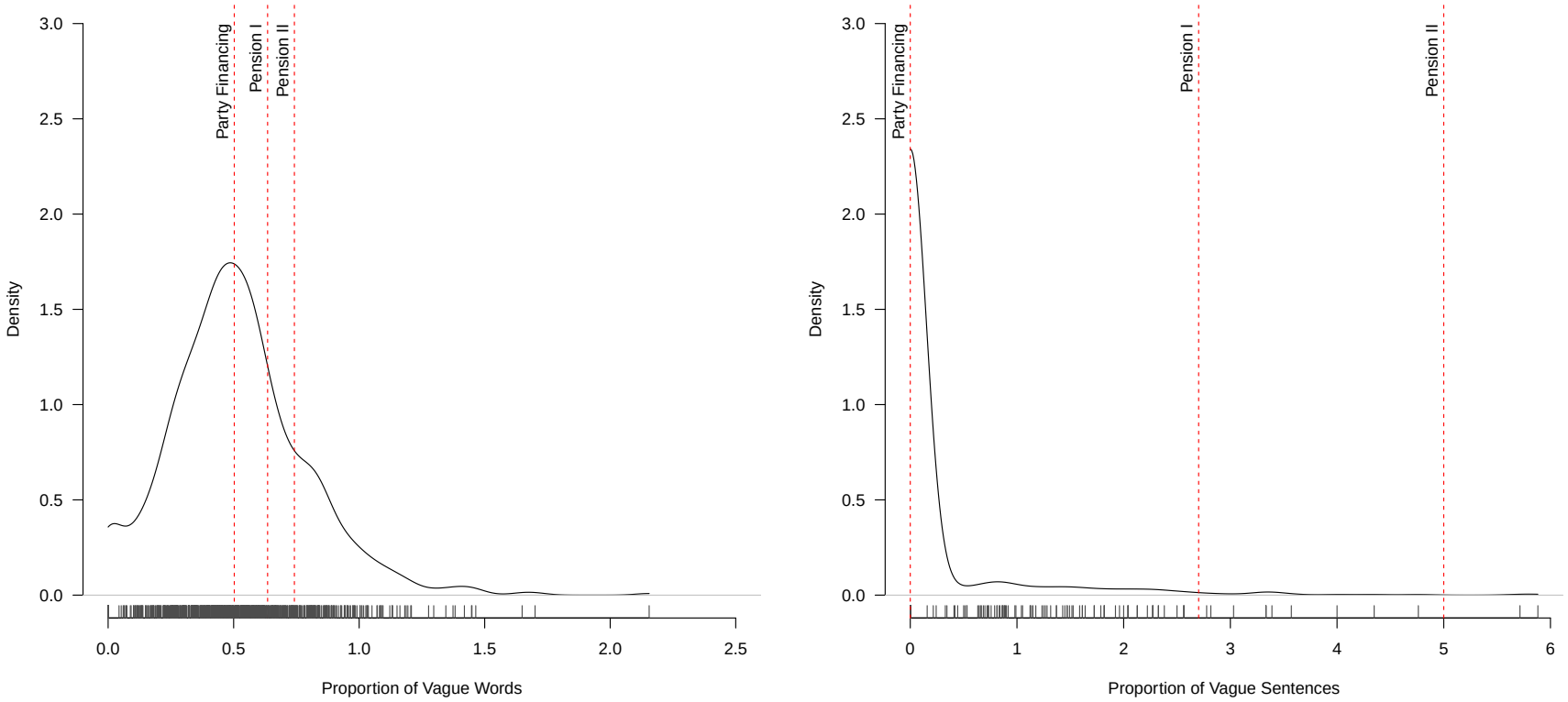
3.5.2 Face Validity of the German Application

To demonstrate the face validity of the German application, I use three decisions discussed in Staton and Vanberg (2008). In their paper, Staton and Vanberg (2008) mention three decisions of the GFCC which either stand, in the context of their paper, for very specific or very vague decisions. This allows me to formulate expectations about where these papers should be located on the vagueness continuum.

The first decision is a decision on party financing (BVerfGE 24, 300). In this landmark decision, the GFCC declared existing eligibility requirements for receiving public subsidies of small parties such as the National Democratic Party of Germany (NPD) as unconstitutional instructed the legislature to revise the party financing law such that any party receiving at least 0.5% of the votes in an election receives public subsidies. Staton and Vanberg (2008) quote this decision as a prime example of a specific, namely a non-vague decision.

The second and third decision are both concerned with pension benefits of civil servants. In 1980, the GFCC ruled about the pension benefits of civil servants who argued that the taxation of their pension benefits is a violation of the equal treatment clause of the constitution because other pensions are not taxed in the same manner (BVerfGE 54, 11). The court declared that this is indeed a violation of the equal treatment clause, but that all necessary steps towards a correction are in the hands of the legislator. Afterwards, no legislation was initiated. In a follow up decision in 1992, civil servants appealed again (BVerfGE 86, 369) and argued that the legislator has evaded the court's decision. Again, the court stated that in this case "legislative delay" was not "unreasonable" given the complexity of the issue (English translation taken from Staton and Vanberg (2008, 513)). Staton and Vanberg (2008) use both of these decisions as a prime example of a vague decision. Therefore, if the measures for the judicial policy implementation vagueness are valid, we would expect both decisions on civil servant benefits to score rather high on the vagueness measure, and the decision on party financing to score rather low.

Figure 3.8 – Distribution of Proportion of Vague Words/Sentences in GFCC Decisions



(a) Proportion of Vague Words, Expanded Dictionary

(b) Proportion of Vague Sentences, CNN Classification

Note: Kernel density estimation of the proportion of vague words/vague sentences in GFCC decisions. Dashed lines indicate the vagueness scores of the party finance decision and the civil servant decisions described in the text.

Figure 3.8 shows the density of the proportion of vague words/vague sentences in a decision, with dashed lines indicating the position of each of the three decisions on the vagueness continuum scored by the expanded dictionary (left side) and the CNN (right side). Both decisions concerning the civil servant's pension benefits are located on the right-hand side of the continuum (indicating a rather vague decision), whereas the party financing decision is on the left side (indicating a rather not vague, ergo specific decision). This is in line with the placement outlined in Staton and Vanberg (2008), and thus confirms the observable implications.

3.5.3 Face Validity of the French Application

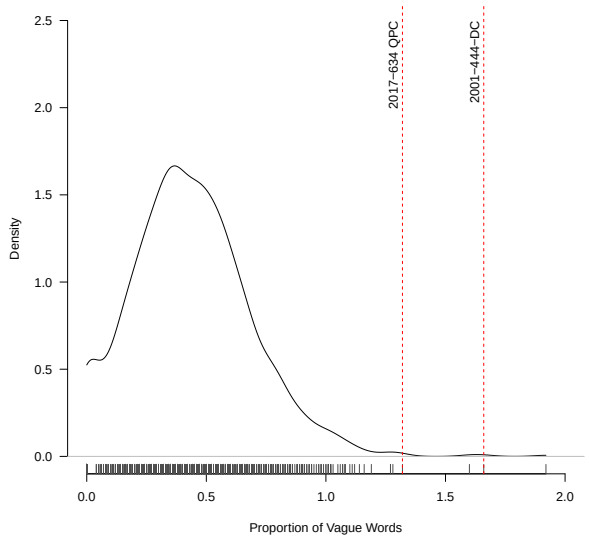
In order to assess the validity of the expanded dictionary for the application on the CC, there is, unfortunately, no comparable data set available like the one of Engst (2018). For this reason, I manually inspect the scoring of decisions which received high vagueness scores and assess whether these scores are in line with what one would expect based on the topic and the decision text. I want exemplarily highlight two decisions which have high vagueness score: the "Décision 2017-634 QPC" and the "Décision 2001-444 DC" (plotted and highlighted in Figure 3.9). Both demonstrate the usage of vague language in decisions of the Conseil.

Decision "2017-634 QPC" is among the most vague decisions in the data with a vagueness score of 1.32% vague words. In this decision, the Conseil examines governmental laws on the modernization of the economy and banking and financial regulations. In particular, it investigates whether certain sanctions mentioned in these laws are covered by the constitution and whether they are justifiable or not. The Conseil decided that the laws and their wording are constitutional. In their decision, the French judges write that the sanctions are justifiable because "the legislator pursued the objective of preserving the economic public order". In this context, the Conseil further remarks that "the Conseil does not have to the right to question the intentions of the legislator, but only gives it jurisdiction to decide on the conformity of the legislative provisions under its consideration with the rights and freedoms that the Constitution guarantees"[54] (own translation).

In decision "Décision 2001-444 DC" (revision of the electoral calendar in 2001), the self-resistance of the Conseil is even more evident. This is the decision with the second highest vagueness score (1.66%) according to the French expanded dictionary. Some quotes from the decision highlight why this is the case. The Conseil writes first that "the legislator [...] can freely modify the durations" and, as an even a more explicit sign for judicial policy implementation vagueness, that "the Constitutional Council does not

[54]French original: "la Constitution ne confère pas au Conseil constitutionnel un pouvoir général d'appréciation de même nature que celui du Parlement, mais lui donne seulement compétence pour se prononcer sur la conformité des dispositions législatives soumises à son examen aux droits et libertés que la Constitution garantit."

Figure 3.9 – Distribution of Proportion of Vague Words According to the French Expanded Dictionary



Note: Kernel density estimation of the proportion of vague words/vague sentences in Conseil Constitutionnel decisions. Dashed lines indicate the vagueness scores of the respective decisions.

have a general discretion of the same nature as that of Parliament; It is therefore not up to it to inquire whether the desired objective of the legislature could be achieved by other means [...]"[55] (own translation). Both of these decisions show that the automatic vagueness detection through the expanded dictionary identified these vague decisions.

3.5.4 Bridging Observations: Scoring the Same Decisions in Two Languages

One serious challenge of the expanded dictionary approach is that judicial policy implementation vagueness is measured across two different languages. Thus, it is important to show that the obtained results are independent of idiosyncratic features of French and German, and not just an artifact of the language they are written in. I use *bridging observations* to show that language is not an influential factor behind my vagueness scores. Bridging observations in my application are decisions which are available in both German and French language. Assuming that the meaning and the nuances of language are not lost in process of the translation, we should observe similar vagueness scores for the same decisions in two languages. In other words, a decision in French should have a similar vagueness score than the same decision in German, and vice versa. Evidence supporting this assumption has recently been published by

[55]French original: "le Conseil constitutionnel ne dispose pas d'un pouvoir général d'appréciation et de décision de même nature que celui du Parlement ; qu'il ne lui appartient donc pas de rechercher si l'objectif que s'est assigné le législateur pouvait être atteint par d'autres voies [...]".

de Vries, Schoonvelde and Schumacher (2018) who show that Google Translate works for Bag-of-Words approaches, to which dictionary classifiers also belong.

In order to obtain the bridging observations, I extracted all officially translated decisions texts of the GFCC and CC from their websites. For the French application, there are 20 decisions available which were officially translated from the court into German. It is important to note that the French Conseil states itself that only the most important decisions are translated, and thus this sample is not random.[56] These decisions are then scored using the expanded dictionaries in both German and French language. My analysis shows that both vagueness scores for German and French decisions are significantly correlated with an r of .41 ($p < 0.05$). This implies that for French decisions, the scoring of the dictionaries is not driven by idiosyncratic features of either language, as the exact same text receives similar vagueness scores independent of the language it is written in.

Unfortunately, the German court does only provide two official French translations of its rulings (while there are over 289 available in English). I therefore randomly pick 20 decisions and translate them into French using *Google Translate* machine translation. This is a common strategy in NLP if no official translation of the same text is available (see e.g. Banea et al., 2008). My analysis shows that both the scores of the original and translated decision texts significantly correlated with .39 ($p < 0.05$).[57] In summary, both analyses confirm that the obtained scores of my dictionaries are not dependent on the language where the measurement approach is applied to.

3.5.5 Validity of the Original LIWC Dictionary

Finally, I evaluate the original LIWC approach using the same validity checks for the expanded dictionary than before. Applying the general LIWC dictionaries to domain-specific texts without a critical evaluation of the result's validity is – unfortunately – still common in the field. In order to show the superiority of my proposed methods (the expanded dictionary and the CNN classifier), I replicate the validity checks, but this time I use the original LIWC implementation.

Put simply, LIWC fails all validity checks. First, the original LIWC implementation is unable to detect vague words/terms in the vague directive texts and thus, cannot statistically distinguish between directive texts coded as vague and texts coded as not vague ($p > 0.10$, two-sided test). In fact, LIWC finds that the directive texts which are vague according to Engst (2018) contain a *lower* proportion of vague words than the directive texts coded as specific. Second, when replicating the face validity analysis, LIWC fails to properly locate the selected decisions on the vagueness continuum. Third, when replicating the analysis using bridging observations, I find that the same decision

[56] Descriptive statistics of the data used can be found in Table B.1 in Appendix B.3.3.
[57] Summary statistics can be found in Table B.2 in Appendix B.3.3.

Table 3.5 – Summary of Validity Checks for Each Court and Methodological Approach

	GFCC			CC	
Validity Check	Expanded Dictionary	CNN Classifier	Original LIWC	Expanded dictionary	Original LIWC
Directives	check	check	fail	not applicable	not applicable
Face Validity	check	check	fail	check	fail
Bridging Observations	check	check	fail	check	fail

in the different languages also receives different scores, and that these scores are not significantly correlated ($p > 0.10$, two-sided test). These findings demonstrate that LIWC, while it may be useful in some domains, should always be used with caution and must be carefully evaluated.

All validity findings are again summarized in Table 3.5. While both the expanded dictionary and the CNN classifier turn out to be a valid measure of judicial policy implementation vagueness, I also demonstrate that using the general LIWC approach in comparison yields to poor results. In sum, I am confident that both measures offer a valid and reliable operationalization of judicial policy implementation vagueness in constitutional court decisions.

3.6 Summary

In this chapter, I have asked *how we can automatically measure vague language in written court decisions?* I introduced two methods to automatically detect judicial policy implementation vagueness in written decisions of the German and the French constitutional court. The first method relies on word embeddings to create legal domain-specific dictionaries in German and French. The second method benchmarks several NLP classifiers on a novel annotated data set consisting of over 3,500 randomly sampled sentences from GFCC decisions. My results show that both approaches are valid methods to measure judicial policy implementation vagueness, while dictionary approaches currently very popular in the field have difficulties measuring this concept.

I make three contributions in this chapter. First, I show that similar measurement and classification approaches already established in other fields of NLP also work properly when applied to judicial texts. Second, I make all my data and code used in this chapter publicly available. Other researchers can use the annotated data set as a benchmark to test different algorithms, and explore more complex model structures. In this regard, recent advances in semi-supervised text classification (the combination of adversarial neural nets to build more training data and supervised classification) have shown promising results, especially in situations where training data is scarce, although the classification problem is challenging(e.g. Aghakhani et al., 2018; Chen and Cardie, 2018). Third, and most importantly, my work provides a blueprint and practical guidance for other researchers in social science who face the sample hurdles, i.e. data scarcity and the lack of prior research in a certain linguistic domain. For researcher

who can rely on an existing general dictionary, I show how word embeddings can be used to expand this dictionary to a specific domain. For researchers without such an existing dictionary, I demonstrate how a state-of-the-art machine learning classifier can be developed to solve a certain classification problem.

Researchers should keep in mind that there is always a trade-off along the two dimensions of accuracy/quality of a classification approach and the required resources. An expanded dictionary does not reach the same precision as a carefully-developed and tuned algorithm, but is easy to implement and at low costs with respect to computational resources and required time. By contrast, a machine learning classifier often outperforms a dictionary in terms of accuracy but requires extensive resources to gather the necessary training data and algorithmic fine-tuning. The implications of this chapter thus go beyond the methodological application on the judicial decisions but teach us some important lessons about the general feasibility and applicability of interdisciplinary approaches.

CHAPTER 4

Why Do Courts Craft Vague Decisions? Evidence From Germany and France

4.1 Introduction

Why are judges vague about the policy implications of a decisions in one occasion and highly specific in the other? In a landmark decision on inheritance tax in Germany in 2014, the German Federal Constitutional Court declared the existing inheritance tax law with respect to family-owned companies as unconstitutional.[1] In their ruling, the judges instructed the federal legislature to revise the law, but also stated that how a revision is implemented is the legislature's task due to the issue's complexity. The Bundestag instituted a committee to study revisions of the inheritance tax code – but no legislation was initiated. After two years without a legislative response, the President of the German court wrote an open letter to the German legislators, highlighting that the court feels obliged to deal with this matter again.[2] It was only after this threat that the government (hastily) adopted a law revision, which is currently discussed to be appealed to the GFCC again by the opposition.

This decision illustrates two fundamental challenges of judicial policy-making. First, when reviewing legislation, judges often confront a wide range of issues and usually do not possess specialized knowledge about it. Second, and the inheritance tax decision illustrates this clearly, judicial decisions are not self-enforcing, and legislative compliance cannot be taken for granted. How can judges deal with these challenges?

[1] 1 BvL 21/12.

[2] `https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2016/bvg16-041.html`, accessed 12.04.2019.

In this chapter, I address this question by testing the implications of a game-theoretic model proposed by Staton and Vanberg (2008). This model is widely cited[3] but has not been empirically examined yet. In essence, the authors argue that decision language allows judges to hedge against their limited policy-expertise while simultaneously it can also be used to pressure hostile legislators or mask legislative noncompliance, depending on the public support of the court. Therefore, their formal model incorporates both of the arguments that are central to this dissertation: public support varies across countries, and courts are active actors who strategically use the institutional tools at their disposal. I put these arguments to test using again a comparative study on the German and French constitutional court and the judicial policy implementation vagueness scores established in the previous chapter.

My findings mainly confirm the implications of the formal model: vague decision language is used by both courts to give discretion to the better informed legislator in complex cases or reduce the legislative room for maneuver if preference divergence between court and legislator is high (but only in Germany). More importantly, my results show that both courts use vagueness differently when legislative noncompliance is likely: The German Federal Constitutional Court, a court with high public support, uses specific decision language to strategically pressure a hostile legislature. By contrast, the French Conseil Constitutionnel, a court with low public support, uses vagueness as a "defensive mechanism" to hide legislative noncompliance from public view.

4.2 The Challenges of Judicial Policy-Making

The relations between courts and other policy-making institutions generate serious challenges for judicial policy-making. The first challenge is the *limited policy expertise* of judges. Designing the adequate public policies for a given desired political outcome is often associated with abstract technical issues. Consider the inheritance tax decision of the GFCC again. Striking down the current inheritance tax law as unconstitutional is only part of eliminating an unconstitutional public policy. Determining which specific policies will achieve a fair taxation of family-owned companies of different sizes, and considering all associated future side-effects of these policies is a difficult problem requiring specialized knowledge.

Other policy-making institutions such as legislatures do not face the same difficulties. Legislators usually have access to a committee system and can rely on a reservoir of staff-experts and other bureaucratic support.[4] Therefore, relative to policy-makers with whom judges interact, they typically have only limited access to technical information

[3]Google scholar lists 167 studies that cite this paper as of April 2019.

[4]For instance, every member of the German Bundestag already has, on average, ten employees as personal staff (`https://www.bundestag.de/dokumente/textarchiv/2015/kw32_finanzierung_buero-384390`, accessed 06.05.2019).

which would be necessary for evaluating alternative policies.[5] In other words, judges may be in the position to evaluate whether a certain law is unconstitutional or not, for instance whether the tax inheritance law under review is in accordance with German Basic Law or not. However, they will typically be in a less advantageous position to design the specific policies that are necessary to accomplish these outcomes compared with other policy makers.

The second challenge which judges face is the *fear of legislative noncompliance*. Because judicial decisions are not self-enforcing, compliance cannot be taken for granted. Courts have no coercive means and, as one of the most cited quotes in judicial policy states, "no influence over either the sword or the purse" (Hamilton 1788, Federalist No. 78). Moreover, considering support of high courts from a cross-national perspective, Gibson, Caldeira and Baird (1998) argue that "with limited institutional resources, courts are therefore uncommonly dependent upon the goodwill of their constituents for both support and compliance" (Gibson, Caldeira and Baird, 1998, 343). When facing an undesired policy demand by the court, legislative majorities might be tempted to ignore the judicial decision. Whether this will happen largely depends on the political costs of such an evasion attempt. Because courts often enjoy broad public support, open evasion of a judicial decision can result in a public backlash that could be electorally costly for governments (Vanberg, 2005; Staton, 2006; Hall, 2014; Krehbiel, 2016).

In summary, there are two main challenges which judges are confronted with when deciding about policy. On the one hand, many decisions require specialized knowledge which judges not always do possess. On the other hand, judges fear legislative noncompliance, since judicial decisions are not self-enforcing. In the next section, I discuss how the strategic use of decision language can help judges to deal with these challenges by means of the game-theoretic model of Staton and Vanberg (2008).

4.3 The Value of Vagueness

The formal model of Staton and Vanberg (2008) offers some general lessons about how judges can manage the judicial policy-making challenges they face.[6] The author's main argument is that decision language, or more precisely, vague or specific decision language, is a tool that judges can use strategically to deal with the outlined challenges of judicial policy-making.[7]

[5] The information which plaintiffs or interest groups offer might often be biased, therefore not presenting a reliable information source for judges.

[6] For space reasons, the formal model and its central implications are only summarized here. A detailed account of the model, including the game's setup, the utility functions of both the court and government as well as the equilibrium and its interpretation is available the original paper (Staton and Vanberg, 2008).

[7] In accordance with the original paper, vague decision language can be understood as a decision that does not articulate any specific demand on the legislator about the means necessary to achieve a certain policy goal (Staton and Vanberg, 2008, 508). By contrast, specific decision language is then understood as

To foreshadow, their theory follows two different lines of arguments. First, they rely on the well-established literature on delegation between legislators and bureaucrats (Bawn, 1995; Epstein and O'Halloran, 1999) and argue that judges will *a)* write vague decisions when they want to give discretion to the better informed legislator and *b)* write less vague decisions the greater the preference divergence between court and legislator. Second, and this is foreign to standard delegation models and the major novelty of their model, they show that how judges use decision language in the face of legislative noncompliance ultimately depends on their level of public support. In what follows, I will shortly discuss the theoretical accounts of both lines of arguments and derive testable hypotheses.

Following the standard delegation story, Staton and Vanberg (2008) argue that by issuing a vague decision, judges can give other policy makers the necessary discretion to take advantage of their superior policy expertise. In other words, writing a vague decision allows judges to hedge against their limited policy-making abilities and to cope with their judicial uncertainty. With a vague decision, the better-informed legislator has any freedom to choose the appropriate steps to implement an optimal policy outcome. This argument from the delegation literature is straightforward: the more leeway the principal wishes to provide to the agent, the more vague that the instructions have to be such that the agent has sufficient maneuver room. The first implication of the formal model can thus be summarized as follows:[8]

Judicial Policy Uncertainty Hypothesis (H1): *All else being equal, the higher the judicial policy uncertainty of the court, the higher the vagueness of the decision.*

However, vagueness is not costless. Providing discretion raises the possibility that the legislator will use its expanded authority to implement an outcome which reflects its own interests. The court here faces a trade-off: on the one hand, if a decision is too vague and gives too much leeway to the legislator, the legislator might be tempted to realize its own policy preferences rather than following the court's demand. On the other hand, by crafting a highly specific decision that exactly outlines the means to achieve a desired policy outcome, judges run the risk of "locking in" an inappropriate policy due to their limited expertise.

Again, following the literature on delegation, Staton and Vanberg (2008) argue that judges are sensitive to the divergence of preferences between them and the policy makers to whom they delegate to resolve this trade-off. If both court and legislator share the same policy preferences, judges are willing to provide the policy-maker with

a decision that clearly articulates the judicial demands. I discuss the concept of decision vagueness in further detail in the operationalization part.

[8]Hypothesis 1 and Hypothesis 2 are direct implications of the first observation of the formal model (Staton and Vanberg, 2008, 511).

adequate discretion. This is because although providing discretion comes at the price that the policy outcome will reflect the legislature's preferences, this cost is more than outweighed by the informational gain the court achieves by giving the better-informed legislator the necessary room to maneuver. The more the preferences begin to diverge, however, the higher the price that the court has to pay, since the legislator will use its freedom to implement outcomes that are increasingly disliked by the court. In such a scenario, the court will write less vague decisions to ensure that the final policy outcome is close to its own ideal point. This second implication of the formal model can be summarized as follows:

Preference Divergence Hypothesis (H2): *All else equal, the higher the preference divergence between the court and the legislator, the lower the vagueness of a decision.*

To sum up the first two hypotheses, increasing judicial policy uncertainty will lead judges to write more vague decision in order to give discretion to the better-informed policy-maker, while they will write less vague decisions the higher the preference divergence between them and the legislature.

The second part of the theory provided by Staton and Vanberg (2008) takes into account that the judicial context adds an important twist to the usage of vague language that goes beyond the standard delegation story. As outlined before, judicial decisions are not self-enforcing, and judges must fear legislative non-compliance. The formal model of Staton and Vanberg (2008) makes an important contribution by arguing that the leverage of judges largely depends on the costs that other policy-makers face for resisting judicial decisions. Here, evading a decision of a very popular court is more costly for legislative majorities than ignoring a decision of a very unpopular one. Moreover, these costs depend on how easy it is for others – either other political elites or the public – to tell that a decision has not (yet) been properly implemented, and on how easy it is for courts to make a credible case to others that a decision has been ignored (see Vanberg, 2005). Previous research shows that courts have some control over the extent to which others can detect non-compliance, and take active measures to increase the chance that the public becomes aware of a decision and its legislative (non)-implementation.[9]

In the formal model of Staton and Vanberg (2008), the authors argue that decision vagueness is an additional mean which judges can use to control the extent to which observers can detect noncompliance. The more clearly the court articulates its demands, the higher the costs for legislative evasion since noncompliance is easier to detect. Vagueness, on the other hand, reduces the costs of noncompliance: if a decision is

[9]For example, Staton (2006) finds that courts strategically use press releases to increase public awareness. Krehbiel (2016) shows that judges strategically use oral hearings to draw attention to a decision.

vague, it might not be obvious that the legislative majority (or another policy-maker) is not complying with a decision, even if the final policy differs considerably from the court's demand. Specific decisions in turn increase the pressure for faithful compliance because they increase the chance that an attempt of evasion leads to considerable political costs. In other words, the more clearly judges state the policy implications of their decision, the easier it is to verify whether policy-makers have faithfully complied. This makes it more likely that the mass public becomes aware of any legislative evasion attempt. Why do judges then not always choose to write clear decisions when fearing legislative noncompliance?

As Staton and Vanberg (2008) highlight, decision language is a "double-edged sword" (Staton and Vanberg, 2008, 513). While a clearly written decision can generate high pressure for compliance, writing specific decisions when legislative noncompliance is likely is not always the best option for judges. Specific language provides judges with a tool to increase the costs of policy-makers resulting from noncompliance, but this strategy can also be unsuccessful and, like a boomerang, turn into costs for the court. If other policy-makers are willing to openly ignore even clearly articulated judicial orders, a specifically written decision of a court which is then ignored highlights the relative lack of judicial enforcement power. Such an open evasion is also costly for courts, because legislative evasion can have a corrosive effect: noncompliance by a policy-maker today may begin to undermine the general perception that court decisions must be respected, thus inducing increasingly more noncompliance tomorrow. Once the evasion of court decisions becomes "normal", courts are at risk of losing their institutional reputation, the only "weapon" against hostile legislative majorities that they have, in the long run (Carrubba, 2005). In order to prevent such an erosion of authority, judges may choose to be vague when they expect legislative noncompliance to protect the court against open legislative resistance. Accordingly, when do judges use language as a tool for pressure and when do they do so to protect themselves against open resistance when legislative noncompliance is likely? Staton and Vanberg (2008) argue that how judges use decision language when they expect legislative noncompliance is a function of their institutional reputation (and thus, depending on their public support) and the preference divergence between them and the legislator.

For a very popular court that enjoys generally robust levels of public support, judges confront a difficult calculus. These courts possess considerable leverage, which influences how they behave in the trade-off between using specific decision language to force compliance and using vagueness to hide resistance. When both the court and the legislator share similar policy-preferences, potential legislative noncompliance will lead judges to write a vague decision. This is intuitive: if the legislator shares similar preferences as the court, it is more beneficial for the court to hide potential noncompliance from the public than trying to pressure the government and move the final policy outcome closer to themselves. However, this calculus changes when such a

powerful court is confronted with potential noncompliance from a legislator who is ideologically distant. In such a scenario, a powerful court will write a specific decision to apply maximum pressure on the legislator. The intuition behind this is that if a court is confronted with an ideologically distant legislator, writing a specific decision not only draws the legislative response towards the court's ideal point, but it also increases the chance that the public will take notice of the evasion attempt of the legislature. This in turn increases the costs of legislative noncompliance. A popular court can behave in such an "offensive" way for two reasons: first, because it has sufficient leverage to credibly threat the government due to its high public support; and second, even if the legislature decides to ignore the specific decision and the public takes notice from the lack of judicial enforcement power, the court's institutional reputation is sufficiently robust to deal with such an open evasion and retain its reservoir of public support nonetheless. How a popular court uses decision language in the face of legislative non-compliance can thus be summarized as follows:[10]

Pressure Hypothesis (H3a): *All else equal, decision vagueness will* ***decrease*** *if a* ***popular*** *court faces a high risk of legislative noncompliance and the preference divergence with the legislature is large.*

The dynamics between preference divergence and risk of non-compliance play out differently for unpopular courts that only enjoy low levels of public support. These courts therefore have less leverage when facing potential legislative non-compliance. This is because the electoral costs for legislative evasion are only small, while the erosive effect of non-compliance could jeopardize the remaining institutional reputation that they possess so far. For these courts, the benefits of using specific language to pressure the legislature are outweighed by the potential costs for appearing as a powerless institution. Judges in such circumstances will use decision vagueness when they expect resistance to "mask" noncompliance. In other words, if an unpopular court's preferences considerably diverge from the legislature's and the risk of noncompliance is high, these courts will try to hide their weak position behind vague language and use it as a "defensive mechanism" to maintain institutional reputation. As Staton and Vanberg (2008) put it with respect to these unpopular courts, "because specificity is a tool with limited effectiveness, vagueness as a mask predominates" (Staton and Vanberg, 2008, 513). This can be summarized with the following hypothesis:

[10]Hypothesis 3a and Hypothesis 3b are direct implications of the second observation the formal model (Staton and Vanberg, 2008, 513).

Defensive Mechanism Hypothesis (H3b): *All else equal, decision vagueness will* ***increase*** *if an* ***unpopular*** *court faces a high risk of legislative noncompliance and the preference divergence with the legislature is large.*

In the following, I will outline a comparative research design that allows me to test these hypotheses.

4.4 A Comparative Application

In an ideal research design, one would be able to analyze the same constitutional court undergoing considerable ups and downs in its public support, in combination with a permanent measure of the court's public support over time. Like this, other institutional factors or country-specific variations would be fix, and one could isolate the effects of interest. However, such data is not available. I will, therefore, use a similar comparative research design than introduced in Chapter 2. In particular, I will again approximate varying levels of public support via the case selection.

4.4.1 Case Selection

To test the theory, I use a comparative design and selected again the French and German constitutional courts. The reasons for choosing these courts are similar to Chapter 2. First, both constitutional courts must possess the right of judicial review, because otherwise courts cannot decide about governmental policies, and the logic of the formal model could not be applied. Second, and more important, the theoretical argument requires two courts that considerably vary in their degrees of public support. In Chapter 2.3, I explained the institutional setting of both courts and discussed their policy-seeking role within the political system. Moreover, I also provided evidence from a recent, comparative survey to demonstrate that the public support for the GFCC is indeed considerably higher than the public support for the CC. Therefore, all necessary requirements to test the formal model's implications are met.

4.4.2 Data and Operationalization

For the analysis of the GFCC, I use a data set originally collected by Vanberg (2005) and extended by Krehbiel (2016). The data set contains all published decisions of the GFCC between 1983 to 2010 reviewing the constitutionality of federal and state laws.[11] Because the original data set of Krehbiel (2016) excludes decisions in which the court

[11] I not only include decisions where the status quo of a law is challenged, but also where it is upheld. This is because Staton and Vanberg (2008) explicitly state that their formal model works in both cases (see footnote 11 in the original paper for more detail).

does not have discretion over oral hearings, I also added these cases to the data set.[12] I do not include special decisions such as decisions on provisional orders (Einstweilige Anordnungen, §32 Act on the GFCC), or requests to exclude a judge from a case (Befangenheitsanträge, §19 Act on the GFCC). This is because the logic of the formal model studied here cannot be properly applied to these decision types. Furthermore, it is important to note that the analysis is conducted on the decision-level. The GFCC typically groups similar cases (proceedings) together, which could be directed against the same law(s), but also for instance against rulings of lower courts. A decision is included in the data set if a least one if its proceeding is directed against a federal or state law. This is also because the vagueness scores created in Chapter 3 are measured on the decision-level. This way of aggregating information from the proceeding-level to the decision-level is also used in similar work on the GFCC (Engst, 2018, Chapter 4). In sum, the final GFCC data set contains 372 observations.[13]

For the analysis of the French CC, I use self-collected data obtained from the web page of the CC on all decisions dealing with the abstract review of laws decided between 1974 and 2010.[14] I only use decisions from 1974 onwards because the CC only became a fully established constitutional court after a reform in 1974 (see Hönnige (2009) for more information on the choice of this time period). No decisions after 2010 are used due to the introduction of a new proceeding type in 2010 (QPC, Question prioritaire de constitutionnalité). QPCs allow the a posteriori control of laws by the CC. This new pool of litigants and extended power of the CC might have changed the court-legislator dynamics. For this reason, only decisions after 2010 are excluded. The final data set on the CC includes 558 decisions.

Measuring Decision Vagueness

Staton and Vanberg (2008) do not provide an exact definition of what makes a vague decision. All they write is that "in the context of our model, a perfectly vague opinion is an opinion that attacks the status quo policy as illegitimate, but does not impose any specific demands on the legislature for reforming that policy" (Staton and Vanberg, 2008, 508).

[12]Therefore, the final data set includes constitutional complaints, concrete reviews, public law disputes, election disputes involving the constitutionality of an electoral law, constitutional disputes between the national and state governments, constitutional disputes within a state, and abstract reviews. It does not include unpublished chamber decisions. Only decisions which deal directly (unmittelbar) with a law are used.

[13]Note that the original data set of Krehbiel (2016) contains 613 observations, but his unit of analysis is at the proceeding-level. The smaller N in my case is explained by aggregating information on the decision-level, the exclusion of decisions which only indirectly (unmittelbar) deal with a law and removing decisions dealing with provisional orders.

[14]These are decisions of the type "DC", so called "Contrôle de constitutionnalité des lois ordinaires, lois organiques, des traités, des règlements des Assemblées".

I argue that this concept can be operationalized by my newly introduced concept of *judicial policy implementation vagueness* in Chapter 3. Judicial policy implementation vagueness was defined as a particular form of linguistic hedging in the context of court decisions and as the intentional choice of words or terms that give the legislator a wide decision leeway and room to maneuver in how it can implement a court decision. Chapter 3 also proposed two different approaches to measure judicial policy implementation vagueness in decisions of GFCC and the CC, namely the usage of an expanded dictionary and a NLP classifier. Both approaches yield to vagueness scores that are tailored for the judicial domain, and are thus superior to the usage of general dictionaries such as LIWC or other approaches commonly used in judicial politics research.

For the GFCC application, I use the vagueness measure that relies on the NLP classifier (the Convolutional Neural Network classifier) because it exhibited the best out-of-sample prediction performance. The German vagueness measure thus describes the proportion of vague sentences in each decision. The mean of the decision vagueness in the German data is 0.19 (with a standard deviation of 0.5). This means that 0.19 percent of all sentences across all decisions texts in the German data are vague. For the CC application, the expanded dictionary measure is used. The French measure thus describes the proportion of vague words in each decision. The mean of the decision vagueness in the French data is 0.42 (with a standard deviation of 0.23). This means that across all words in the French data, 0.42 percent are vague.

At this point, I want to emphasize again that both vagueness scores measure the occurrence of specific words or sentences within long decision documents in a large text corpus. This is why the empirical approximation of these terms is relatively small in absolute numbers. Nevertheless, I have demonstrated in Chapter 3 that even small changes in the wording of a decision can have a considerable influence on its judicial policy implementation vagueness.

Measuring Judicial Policy Uncertainty

Judicial policy uncertainty describes a court's uncertainty about the necessary means to achieve a desired policy outcome. An ideal measure would be, for instance, the training background of each judge for each policy field, a judge's prior experience with a given topic, or, as a simpler approximation, the policy expertise of at least the rapporteur responsible for a decision. Unfortunately, data on all of this information are not available. For this reason, I argue that the judicial policy uncertainty of judges can be measured by a decision's *complexity*. To approximate this complexity, I use the dichotomous coding scheme for complexity following the measurement strategy proposed by Vanberg (2005) and recently used by Krehbiel (2016). According to this coding scheme, decisions involving taxation, budgets, economic regulation, social insurance, civil servant compensation, and party finance are coded as "complex" with

a value of 0, whereas those involving institutional disputes, family law, judicial process, individual rights, asylum rights, and military conscription are coded as "simple" with a value of 1. In the German data set, this variable is already included in the original data set of Krehbiel (2016) for most of the decisions. For the decisions I manually added, I use the data from the CCDB that contains for each decision the information on the policy area following the Comparative Agenda Project[15] (CAP) to classify these decisions accordingly. The German measure for judicial uncertainty is thus a dummy variable indicating whether a decision is complex or not.

Although a French version of the policy topic coding for each decision is available from the French CAP project, I refrain from using this data in the main analysis because it is only available until 2007. However, I will use it to test the robustness of the findings later in the robustness section. I approximate the case complexity in the French application by the number of legal doctrines the CC has to consider in a decision. In the introduction of the French rulings, the Conseil always quotes the laws, statutes and legal doctrines it has to examine in this particular ruling. I argue that the more laws, statutes or doctrines the Conseil has to consider, the more complex the issue of a ruling is.[16] This variable ranges from 1 to 28, with a mean of 5.2 and a standard deviation of 3.7.

Measuring Preference Divergence

Preference divergence describes the divergence between the ideal point of the court and the policy preferences of the government. I follow the common measurement approach first proposed by Hönnige (2009) by measuring preference divergence as the absolute ideological distance between the court and the government on a common left-right scale using the ideology scores from the Comparative Manifesto Project (CMP) (Laver and Budge, 1992). The position of the government is calculated by weighting the CMP scores by the seats of the governing parties in parliament. This allows for a more nuanced measurement of the government's policy position than using the raw CMP scores without weighting. In the German analysis, the position of each Senate of the GFCC is measured by assigning each judge the CMP score of the political party that nominated him or her on the given day this judge entered the court. Subsequently, the mean position of each Senate is calculated. In the French application, the same method was applied (but there is only one chamber). Finally, the absolute distance between the government position and the court position is calculated to obtain the ideological distance between court and government.

[15] `https://www.comparativeagendas.net/`, accessed 12.04.2019.

[16] A similar line of argument in the context of judicial decisions can be found in Wittig (2016, 103).

Measuring Risk of Non-Compliance

The measurement of the perceived *risk of non-compliance* must approximate the appraisal of the judges whether or not the government will try to evade a decision after the court has delivered a ruling. I follow the measurement strategy of Vanberg (2005) and Krehbiel (2016) and measure the risk of noncompliance by examining whether or not the government whose law is being challenged filed an amicus brief defending the constitutionality of the statute. Filing such a brief can indicate the level of importance of a law for the government, because it requires the legislator to invest resources in such a statement. Furthermore, such a public statement (which will also be published together with the judicial decision) demands the legislator to position itself publicly, and is thus risky for the government's reputation (Krehbiel, 2016, 997). For both applications, the variable is coded 1 when the challenged government files a brief defending the constitutionality of the law under review and 0 otherwise.

Confounders

For the German application, I control for two potential confounders, namely the *public awareness* for a decision and the deciding *Senate*. Vanberg (2005) and Krehbiel (2016) find that the degree of public awareness for a decision affects the behavior of the GFCC. The court's decision to write extra vague or specific decisions might be correlated with the salience of a decision. If the public awareness for a decision is high, the court could use this to additionally increase the pressure on the government and write even more specific decisions, because the government's room for maneuver is smaller than without public awareness. In the same way, case salience may correlate with the government's decision to file a brief. Therefore, failing to control for existing public awareness of a case could lead to biased results. I use the fact whether the court holds a public hearing or not as a proxy variable of a case's salience and thus, as an indicator of public awareness. This variable is already used in other studies to approximate public awareness (Vanberg, 2005; Krehbiel, 2016). Unfortunately, the Conseil does not hold such hearings, nor is another measure available that approximates the salience or public awareness for a CC decision in the French application.

I also control for the institutional structure of the GFCC. The GFCC comprises two different Senates with varying persons and jurisdiction. This could result in a systematic variation of the usage of decision language. I address this possibility with the variable Second Senate, which has the value 1 if the Second Senate is concerned with a decision and a 0 if it is the First Senate. Since the French constitutional court only comprises one chamber, there is no need to control for this factor. Summary statistics of all variables used throughout the analyses are in Appendix C.1.

4.4.3 Statistical Model

Because the dependent variable is a continuous measure bounded to the interval $[0, 1]$ (the proportion of vague sentences in a decision document)[17], I employ a fractional logit regression (Papke and Wooldridge, 1996). Fractional logit is a Quasi-MLE (QMLE) method with the conditional expectation of a fractional response variable:

$$E(y_i|x_i) = \Lambda(x_i\beta) = \frac{exp(x_i\beta)}{1 + exp(x_i\beta)} \tag{4.1}$$

where y_i with $0 \leq y_i \leq 1$ is the fraction of vague sentences in document i, $\Lambda(.)$ is the logistic function and x_i refers to the explanatory variables of document i, a $1 \times K + 1$ vector for K independent variables. β is a $K + 1 \times 1$ vector of coefficients. The quasi-likelihood is the same as the Bernoulli log-likelihood used in the ordinary logistic regression case, with the individual contribution given by:

$$l_i(\beta) = y_i log\big[\Lambda(x_i\beta)\big] + (1 - y_i) log\big[1 - \Lambda(x_i\beta)\big] \tag{4.2}$$

Papke and Wooldridge (1996) show that variance misspecification can be an issue when estimating a fractional logit regression. In my application, such a misspecification could arise if the number of sentences in a decision document and some of the covariates are not independent. The usual standard errors from ordinary logistic regression are then misleading. As a fix for this, Papke and Wooldridge (1996) propose to use robust standard errors based on the well-known sandwich estimator (Papke and Wooldridge, 1996, 622). I follow this estimation strategy in the main analyses, but provide additional analyses based on bootstrapping as an alternative way to obtain estimates and standard errors in the robustness section.

In order to test H1 and H2 in the German application, the following model specification is estimated ($\Lambda(.)$ is always the logistic function):

$$\begin{aligned} E(ProportionVagueSentences_i|x_i) = &\Lambda(\beta_1 + \beta_2 IdeologicalDistance_i \\ &+ \beta_3 CaseComplexity_i \\ &+ \beta_4 SecondSenate_i \\ &+ \beta_5 OralHearing_i \\ &+ \beta_6 RiskNoncompliance_i) \end{aligned}$$

[17] For the French application, it is the proportion of vague words in a decision document.

Testing H3a requires an interaction because the behavior of the GFCC in cases of high risk of noncompliance is argued to be conditional on the ideological distance to legislator. To test H3 in the German application, the following model is estimated:

$$\begin{aligned} E(ProportionVagueSentences_i|x_i) = \Lambda(&\beta_1 + \beta_2 IdeologicalDistance_i \\ &+ \beta_3 CaseComplexity_i \\ &+ \beta_4 SecondSenate_i \\ &+ \beta_5 OralHearing_i \\ &+ \beta_6 RiskNoncompliance_i \\ &+ \beta_7 IdeologicalDistance_i \times RiskNoncompliance_i) \end{aligned}$$

For the French application, the following models are estimated for H1 and H2:

$$\begin{aligned} E(ProportionVagueWords_i|x_i) = \Lambda(&\beta_1 + \beta_2 IdeologicalDistance_i \\ &+ \beta_3 CaseComplexity_i \\ &+ \beta_4 RiskNoncompliance_i) \end{aligned}$$

The model for H3b in the French applications again contains an interaction term, because the behavior of the court in cases of a high perceived risk of noncompliance is conditional on the ideological distance between court and legislator. The following model is estimated:

$$\begin{aligned} E(ProportionVagueWords_i|x_i) = \Lambda(&\beta_1 + \beta_2 IdeologicalDistance_i \\ &+ \beta_3 CaseComplexity_i \\ &+ \beta_4 RiskNoncompliance_i \\ &+ \beta_5 IdeologicalDistance_i \times RiskNoncompliance_i) \end{aligned}$$

Generally, positive coefficients suggest that a court writes more vague decisions and negative coefficients suggest that a court writes less vague (ergo, specific) decisions. Because the coefficients of fractional logit models are difficult to interpret, I use simulations to produce quantities of interest for sensible scenarios (King, Tomz and Wittenberg, 2000).

In the simulations, I use the so-called "*observed value*" approach (Hanmer and Ozan Kalkan, 2013). In this approach, only the variable(s) of interest are varied and each of the other independent variables are hold at their observed values for each observation in

the data. Then, the relevant quantity of interest is calculated for each observation, and finally averaged over all observations (Hanmer and Ozan Kalkan, 2013, 264). Hanmer and Ozan Kalkan (2013) argue that the observed value approach has multiple advantages compared with the usually-used "*average-case*" approach[18], most importantly that the obtained results better represent the collected data and that the findings are more robust to model misspecification. It needs to be emphasized that the expected values of these simulations are not the predicted probabilities of writing a vague decision, but the expected proportion of vague sentences in a decision vagueness. I also want to stress that the dependent variable in the German and French analysis is not the same: in Germany, it is the proportion of *vague sentences*, in France it is the proportion of *vague words* in a decision. Thus, the expected values are not directly comparable.

4.5 Results

This section shows the results of the fractional logit models for the Judicial Uncertainty Hypothesis (H1), the Preference Divergence Hypothesis (H2) and the two hypotheses related to the perceived risk of legislative non-compliance for both analysis in Germany (Pressure Hypothesis (H3a)) and France (Defensive Mechanism Hypothesis (H3b)). I only show simulation results in this section, but provide the corresponding regression tables in the Appendix in Table C.2.

Results of the Judicial Policy Uncertainty Hypothesis

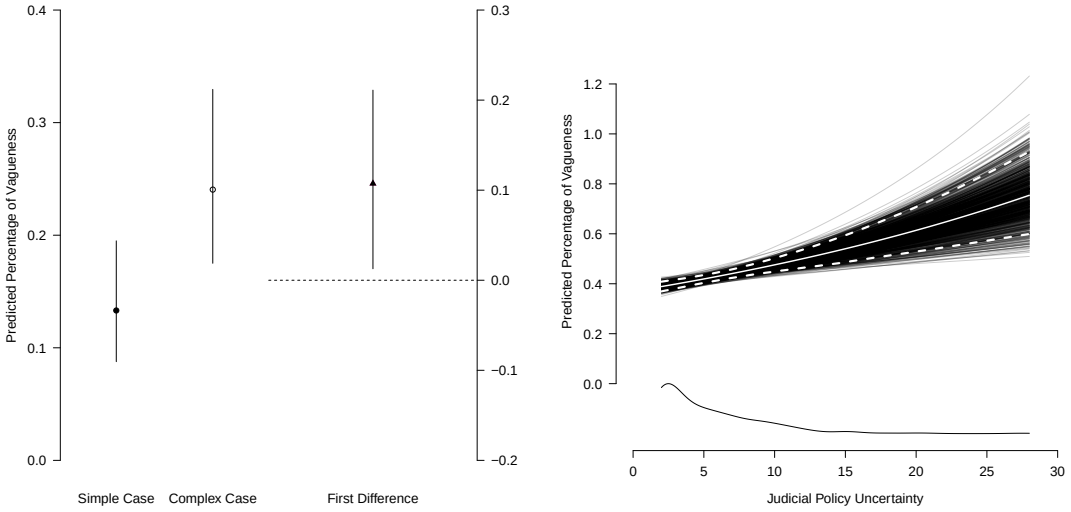
The formal model predicts that decision vagueness will increase as judicial policy uncertainty increases, because courts want to give discretion to the legislator if they face decisions which required specialized knowledge they do not possess. Figure 4.1 plots the effect of judicial policy uncertainty on decision vagueness for Germany (left side) and France (right side). For all graphs, I scaled the Y-axis such that it ranges from 0 - 100, so that the numbers can be directly interpreted as percentages.[19] For continuous variables, I visualize the inferential uncertainty in a "spaghetti"-style plot instead of directly summarizing it in confidence intervals. Each line in this plot represents the expected values based on one draw of the simulation. Confidence intervals are displayed using dashed white lines. The distribution of the independent variable of interest is shown by a density plot on the x-axis. This is because the usually used rugs do not reveal the number of observations in the data represented by each rug.

For both countries, we observe a positive and statistically significant effect of judicial policy uncertainty on decision vagueness. Just as the formal model predicts, this means

[18] In the average-case approach, the variable(s) of interest are varied and the values of the other independent variables are usually set at their mean or median.

[19] This simply means that the fractions are multiplied by 100.

Figure 4.1 – Effect of Judicial Uncertainty on Decision Vagueness, GFCC and CC



(a) GFCC analysis *(b) CC analysis*

Note: Left Side: Expected percentage of decision vagueness for simple and complex cases, with the corresponding first difference using simulations. Simulations are carried out using $N = 1,000$ draws using Model 1 of Table C.3. For the simple and complex case scenario, the *complex case* dummy was set to zero and one, respectively. The points represent the point estimates and the bars represent 90% confidence intervals.
Right side: Expected percentage of decision vagueness over a range of judicial uncertainty, including 90% confidence intervals using Model 1 of Table C.4. Judicial uncertainty is measured by the number of legal doctrines considered in a decision.

that the higher the judicial policy uncertainty of the judges, the higher the decision vagueness. In the German analysis, decision vagueness (percentage of vague sentences) increases by 0.1 percentage points when judges face a complex case compared with a simple case. This is a small but significant effect (the corresponding first difference is statistically significant at the 90% level).

In the French analysis, decision vagueness increases by about 0.36 percentage points from 0.39 percent (minimum judicial uncertainty) to 0.75 percent (maximum judicial uncertainty) decision vagueness (vague words per decision) over the range of judicial uncertainty (measured by the number of legal doctrines examined in a decision). This is, on average, around a 1.5 standard deviation increase in the predicted percentage of decision vagueness. Both analyses thus support the predictions of the formal model of Staton and Vanberg (2008) with respect to the Judicial Policy Uncertainty Hypothesis.

Results of the Preference Divergence Hypothesis

With respect to preference divergence, the formal model predicts that higher preference divergence, namely less ideological agreement between court and legislator, will lead to writing less vague decisions. This is because the courts are aware that if they give the legislator discretion by writing vague decisions, the final policy outcome will less strongly represent the preferences of the court and more the preferences of the legislator.

While this is not a problem when the court and legislator share the same preferences, it is undesirable from the court's perspective to write vague decisions in cases of a large preference divergence with the legislator.

Figure 4.2 plots the effect of preference divergence on decision vagueness. We observe different effects when looking at the GFCC (left side) and the CC (right side). In the analysis of the GFCC, there is, as expected, a negative effect of preference divergence on decision vagueness. This means that the higher the ideological distance between court and legislator, the less vague the decisions become. This is to ensure that the final policy outcome is not too distant to the GFCC's ideal point. Therefore, the German analysis supports the formal model's predictions. Please note that although the confidence intervals at the upper and lower end of the curve (minimum and maximum preference divergence) overlap, the corresponding first difference is statistically significant at the 90% level (see Appendix C.3).

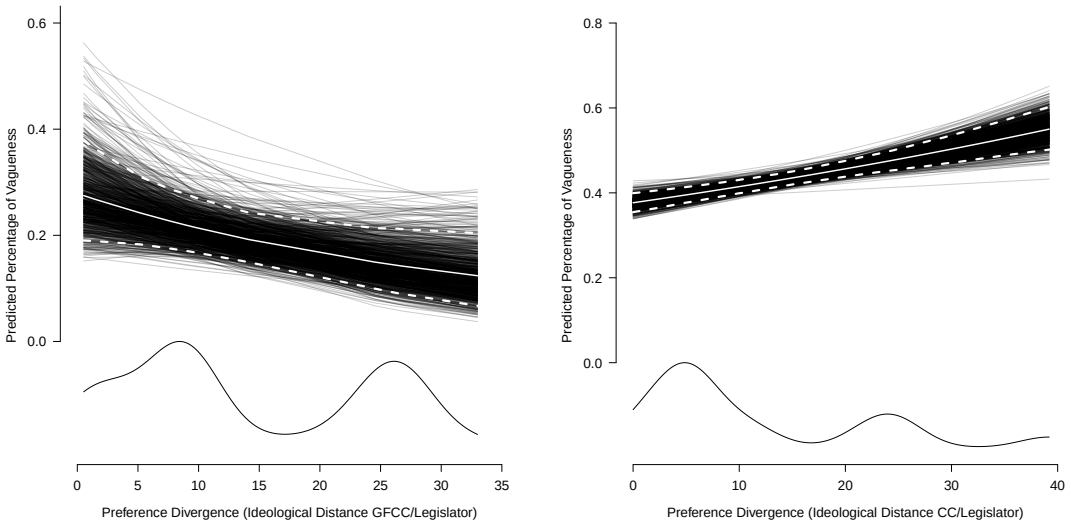
However, when looking at the French court, we observe the opposite effect than in Germany: an increase in preference divergence leads the French judges to write increasingly vague decisions. This effect is statistically significant at the 95% level. This is against the formal model's prediction. One reason for this finding might be that the French judges are willing to make more concessions to the legislator because of the highly political appointment procedure in France (Hönnige, 2009). Another explanation might be measurement error: the CMP scores used to measure the ideological distance between court and legislator are criticized with regard to their spatial and temporal comparability (Lowe et al., 2011; König, Marbach and Osnabrügge, 2013). I will have a closer look at this possibility in the robustness section. In summary, my analyses of the Preference Divergence Hypothesis shows that the formal model's prediction is only confirmed when looking at the GFCC, but not for the CC.

4.5.1 Results of the Non-Compliance Risk Hypotheses

Both hypotheses with respect to noncompliance argue that public support is a decisive factor that determines how constitutional courts deal with potential non-compliance of the legislator. Here, we expect the two constitutional courts to behave differently. A popular court such as the GFCC is predicted to use decision vagueness to pressure the government when it is ideologically distant to the government and fears legislative noncompliance (Pressure Hypothesis). By contrast, the CC is expected to use vagueness as a "defensive mechanism": when it is ideologically distant to the legislator and the risk of non-compliance is high, the CC is expected to mask its lack of enforcement power with vague language (Defensive Mechanism Hypothesis).

Figure 4.3 plots the effect of preference divergence conditional on the risk of legislative noncompliance on decision vagueness. When looking at the results for the GFCC, we observe that the German court writes less vague (-0.14 percentage points) decisions when facing a high risk of noncompliance in cases where preference divergence is large

Figure 4.2 – Effect of Preference Divergence on Decision Vagueness, GFCC and CC



(a) GFCC analysis *(b) CC analysis*

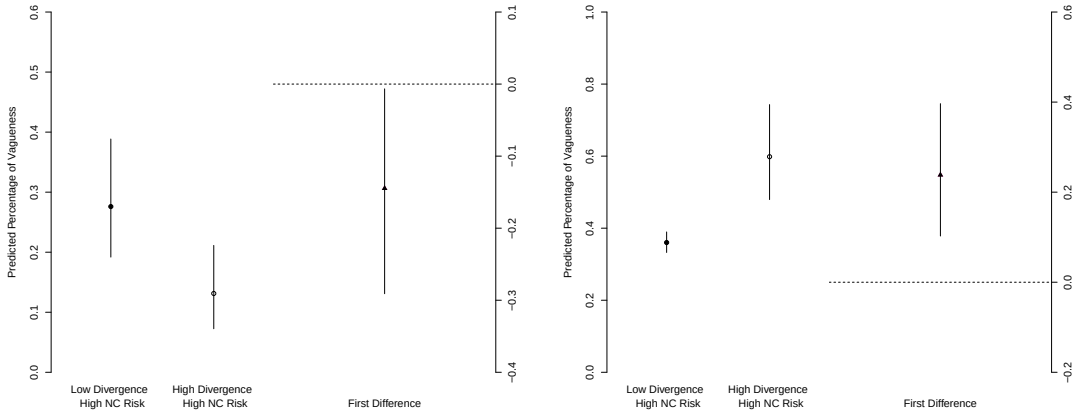
Note: Left Side: Expected percentage of decision vagueness over the range of preference divergence, measured by the ideological distance between court and legislator. 90% confidence intervals are used. Simulations are carried out using $N = 1,000$ draws using Model 1 of Table C.3 in Appendix C.2.
Right side: Expected percentage of decision vagueness over the range of preference divergence, measured by the ideological distance between court and legislator. 90% confidence intervals are used. Simulations are carried out using $N = 1,000$ draws using Model 1 of Table C.4 in Appendix C.2.

compared with cases in which preference divergence is low. The corresponding first difference is statistically significant at the 90% level (see Appendix C.3). This is exactly in line what the formal model predicts: If the legislature's ideal point is already close to the one of the GFCC, it is better from the GFCC's perspective to hide the remaining differences behind vague language than to try to pressure the legislature. However, when both preferences diverge considerably, the court writes less vague decisions to apply high pressure for compliance.

If we look at the results for the CC, we observe the opposite behavior. Again, in line with the formal model, the CC writes more vague (+0.24 percentage points more vague words) decisions in cases with a high risk of noncompliance and large preference divergence than in cases with a high risk of noncompliance but low preference divergence. This is because the Conseil is more concerned about its own institutional reputation (which could be harmed if legislative evasion becomes evident to the public) in cases of noncompliance than about pressuring the government with specific language (which in turn also increases the chances that the public takes notice of an evasion). Hence, the Pressure Hypothesis and the Defensive Mechanism Hypothesis are confirmed in the empirical analysis.

In summary, the presented evidence on both courts mainly supports the theory of Staton and Vanberg (2008). Just as the formal model suggests, my comparative

Figure 4.3 – Conditional Effect of Preference Divergence and Non-Compliance Risk on Decision Vagueness, GFCC and CC



(a) GFCC analysis *(b) CC analysis*

Note: This figure shows the conditional effect of preference divergence and the risk of non-compliance (NC) on decision vagueness. The black points represent the expected decision vagueness for a case with a high perceived risk of non-compliance and low preference divergence. The white points represent the expected decision vagueness for a case with a low perceived risk of non-compliance and high preference divergence. The first difference between these two scenarios is displayed on the right of each graph. The bars represent 90% confidence intervals. For the scenarios, the risk of non-compliance variable is set to 1 and ideological distance is set to the minimum (low divergence) and maximum (high divergence). Estimates are based on Model 2 in Table C.3 and Model 2 in Table C.4 in Appendix C.2.

results confirm that judges strategically vary their decision vagueness depending on the amount of policy uncertainty they face. However, the second hypothesis can only be confirmed with respect to the GFCC: German judges write increasingly less vague decisions the more ideologically distant that they are from the legislator. This relationship cannot be confirmed when looking at the French CC, where we observe the opposite effect.

More important, in line with Staton and Vanberg (2008)'s argument, I find that both courts anticipate the potential future legislative behavior and adapt the vagueness of their decision accordingly. The GFCC uses extra-specific language to pressure the legislator to follow its rulings, while the CC uses extra-vague language to mask potential attempts of evasion.

4.5.2 Robustness Analyses

I check the robustness of my findings when using different measurements for policy divergence and judicial uncertainty and an alternative dependent variable. A detailed report of all robustness analyses is in Appendix C.4.

First, a central variable in my analyses is the measurement of the degree of policy divergence between court and legislator. For the main analysis, I used the scores from the Comparative Manifesto Project to calculate the ideological position of court and

legislator. These scores are increasingly criticized regarding their spatial and temporal comparability (Lowe et al., 2011; König, Marbach and Osnabrügge, 2013). In order to check whether my findings are robust to the measurement of ideological distance, I replicate my analyses but use the Manifesto Common Space Scores (MCSS) of König, Marbach and Osnabrügge (2013) instead of the original CMP scores. My findings remain robust to the usage of this different ideology measure.

The second robustness check concerns the measurement of judicial policy uncertainty. For the German analysis, instead of using the complexity of the topic of a decision, I rely on another complexity measure first proposed by Wittig (2016) that measures the length of the case facts of a decision. The case facts can be found at the beginning of a decision of the GFCC and describe the case, the plaintiff's arguments and other decision-relevant context. Her measure takes into account how long this section is by counting the number of paragraphs. Wittig (2016) argues that the longer the case facts, the more complex the content of a decision is (Wittig, 2016, 102). I repeat my analyses of the GFCC by using the length of case facts as the judicial uncertainty measure.[20] I find that while the sign of the respective coefficient is in the expected direction (the longer the case facts, the higher the vagueness of a decision), the coefficient itself is not statistically significant ($p > 0.1$). One possible reason for this could be that the length of case facts is strongly correlated with the overall length of a decision (Pearson's r is 0.64, $p < 0.01$). Thus, endogeneity could be an issue since the dependent variable is a function of decision length and the number of vague sentences.

For the French analysis, I re-run the main analyses but this time I use a dummy variable indicating whether a case is complex or not (using the same Comparative Agenda Project coding scheme than in the German analysis) instead of the count of the numbers of legal doctrines examined as a measure for judicial uncertainty.[21] The results remain unchanged when using this alternative measure.

The third robustness test replicates German main analysis but uses a different measurement of the dependent variable. Instead of using the proportion of vague *sentences* in a decision body (the dependent variable based on the CNN classifier), I employ the proportion of vague *words* (measured by the expanded dictionary established in Chapter 3). For all hypotheses, the coefficients are in the expected directions, but not statistically significant. One reason for this could be that the validation of this measure in Chapter 3.5 showed that the expanded dictionary produces a number of false positives. These, in turn, could affect the statistical analysis since some decisions might be declared as vaguer than they actually are.

[20] Like Wittig (2016), I cut off the variable at 200 because of an extreme low density above (Wittig, 2016, 104).

[21] I did not use this measure in the main analysis because not for all decision in my data the CAP coding is provided (only until 2007), and therefore the overall N is smaller.

For the fourth robustness check, I apply repeated random sub-sampling validation to account for unobserved heterogeneity. I re-run my analyses twenty times using only a two-third subset of the data, collect the results and combine the estimates as outlined in King et al. (2001). In the French analysis, the results remain unchanged over the different subsets with respect to size, sign and significance. In the German analysis, I find that the direction and size of the estimates are robust across different subsets, but that the combined estimates are no longer statistically significant, as indicated by larger standard errors. However, this is a result of the rather small N of the German data ($N = 372$ decisions), so that the larger standard errors are a result of sampling variability.

For the last robustness check, I use bootstrapping as an alternative way to obtain estimates and standard errors (Efron, 1979; Efron and Tibshirani, 1986). In the main analysis, the robust variance estimator outlined in Papke and Wooldridge (1996) was used. In this robustness check, I replicate all main analysis and simulations using bootstrapping. Bootstrapping is a straightforward and easy-to-implement computational procedure to derive estimates of means and standard errors. The strength of bootstrapping is that the sampling distribution of a quantity is approximated by repeatedly taking n samples *with replacement* from the original data. Like this, less distributional assumptions are required, because mean and standard error are directly calculated from the sampling distribution. In my robustness analysis, I use $n = 1,000$ where the size of each bootstrap sample is identical to the original data. These bootstrapped samples are also directly used as the sampling distributions for the simulations. The coefficients and standard errors obtained via the bootstrapping procedure are similar the ones from the main analysis. Also, all simulation results mirror the findings from the main analysis (see Appendix C.5 for a detailed analysis): how courts use vague language is a function of their (limited) policy expertise, the preference divergence with the legislator and the public support of the court. In summary, the evidence of the robustness analysis mostly supports the findings of the main analysis. Future work, though, must assess the robustness of the German analysis using alternative measures and more data in further detail.

4.6 Conclusion

In this chapter, I examined the indirect influence of public support on the choices that judges make. In particular, this chapter is the first one to test the empirical implications of the game-theoretic model of Staton and Vanberg (2008). In essence, this formal model argues that courts strategically vary the vagueness of their decisions as a function of their policy preferences, judicial uncertainty and their fear of potential legislative noncompliance. Nonetheless, the court's ability to use vagueness ultimately depends on their level of public support. Therefore, this chapter considers both the varying

degrees of diffuse support of different courts and the perception of courts as active actors that use the institutional tools at their disposal to manage legislative resistance.

In a comparative study of two constitutional courts in Germany and France using the judicial policy implementation vagueness measure established in Chapter 3, I find mostly support for the central implications of the formal model. Both courts use vague decision language to give discretion to the better-informed legislator. The German judges refrain from doing so if the legislator is ideologically distant, whereas the French judges behave in contrast to the formal model in this regard and write more vague decisions. I also show that courts use decision language differently depending on their public support: The German Federal Constitutional Court, a court with high public support, writes extra-specific decision in the face of potential noncompliance to strategically pressurize the legislator and to induce higher costs for legislative noncompliance. By contrast, the French Conseil Constitutionnel, a court with low public support, uses vagueness as a "defensive mechanism" to hide noncompliance from public view.

These findings have larger implications for the study of judicial politics. If judges strategically write vague decisions to manage the challenges of judicial policy-making, then empirical tests of common separation-of-powers models should take this into account. Studies only using binary measures of decision outcomes (e.g. a law is declared as unconstitutional or not), as currently undertaken in many studies, are likely to underestimate the actual degree of strategic judicial behavior. Therefore, my findings suggest that judicial politics must overcome the binary coding of judicial outcomes by using richer and more fine-grained measures of strategic judicial behavior.[22] In addition, considering the language of judicial decisions is an important step to reconcile legal scholarship with political science, since it takes factors into account that most often are ignored in quantitative judicial politics.

My results are in line with a growing part of the judicial literature which suggests that courts and judges take active measures to prevail in the strategic interaction between court, government, and the public (Staton, 2006, 2010; Krehbiel, 2016; Engst, 2018). It shows that courts use the institutional tools at their disposal to deal with potential legislative noncompliance. Finally, because the implications of Staton and Vanberg's (2008) model can be applied to many other delegation relationships, the findings of this chapter have implications beyond judicial politics and open new avenues for further research. Central banks and other non-majoritarian institutions face the same delegation problems as constitutional courts, but also lack proper enforcement mechanisms. Further research could thus investigate whether these institutions strategically use vague language in a similar way as suggested in this chapter.

[22] A remarkable exception of conceptualizing judicial choices as binary choices is Engst (2018).

CHAPTER 5

How to Forecast Constitutional Court Decisions? Legal and Political Context in a Machine Learning Application

5.1 Introduction

Is it possible to correctly predict decisions of the GFCC with an algorithm? And which factors are important for the prediction: legal context or political context factors? Algorithmic forecasting of court decisions is relatively new to the field. However, building upon recent efforts in applied machine learning, several studies already achieve impressive forecasting performances predicting US Supreme Court decision-making (Ruger et al., 2004; Guimera and Sales-Pardo, 2011; Katz, Bommarito and Blackman, 2017*b*). Nonetheless, these studies have two important limitations. First, they exclusively focus on the US Supreme Court, which raises concerns about the applicability of these forecasts to other courts, and thus, the external validity of their findings. Second, none of the existing studies explicitly tests the relative contribution of legal context versus political context factors for the forecast of court decisions. There is a long-standing debate about which factors influence judicial decision-making. On the one hand, traditional legal scholars emphasize the importance of the legal and procedural context of a decision, while social scientists on the other hand also acknowledge the importance of the political context of a decision. Teasing out the relative importance of these factors improves our understanding of court decision-making from a predictive perspective.

The contribution of this chapter is to address these two limitations. First, I investigate *whether it is possible to correctly predict the decision-making of the GFCC using a machine*

learning algorithm? I find that with a widely-used machine learning approach (random forests), on average it is possible to correctly predict 76.40 percent of the outcomes of over 2,900 proceedings decided by the GFCC between 1972 and 2010 using out-of-sample prediction. I also address the second limitation by explicitly teasing out the importance of variables associated with the legal and the political context of a decision. The key argument here is that if traditional legal scholars are right, then the legal and procedural context of a decision should be a sufficient predictor of court decision-making. However, if social scientists have a point, then including political context into the forecasting model should increase its predictive performance. For this reason, in this chapter I also analyze *whether political context factors, including public opinion, contribute to the prediction of court decision-making on top of legal context factors*? The results of my prediction show that the legal context alone is already a good predictor of court outcomes. However, I find that forecasting performance can be further improved when the political context of a decision is additionally considered. I conclude that the ensemble of both legal and political factors is needed to characterize court decision-making. The results of this chapter therefore shed light at the importance of public opinion (and other political context factors) for the relationship between *court-public* and have important implications beyond the application to the GFCC which I discuss at the end of this chapter: the value of predictive modeling for the field of social science.

5.2 Existing Approaches to Forecast Court Decision-Making

Forecasting the outcome of a court decision is a long-standing idea which originates from the very early stages of judicial politics research. "Legal prophecy", how Holmes (1897) termed it, has drawn considerable interest of scholars from various fields. Legal academics and political scientists have long scrutinized judicial decisions to understand what motivates courts and judges and how they arrive at a given outcome. These studies often look at past decisions and historical facts, e.g. individual judges' voting patterns, to *explain* why a certain court decided in a certain way. Most often, the goal is not to predict the outcome itself, but to use the causal connection between certain aspects of judicial decision-making to assess the consistency of some explanatory theory. Most of these studies use classic hypothesis testing, and are not interested in whether a model correctly predicts the outcome, but rather if certain estimates are statistically significant or not.

However, with the rise of artificial intelligence over the last decade, a new sub-field has emerged in judicial politics: the field of *quantitative legal prediction*[1] (Katz, 2013).

[1]This term was first introduced by Katz (2013). Quantitative legal prediction can be understood as an umbrella term for all different kinds of non-inferential, predictive approaches that aim at analyzing or predicting legal outcomes.

In contrast to traditional causal inference approaches that make – at best – theory driven predictions about future outcomes, quantitative legal prediction focus entirely on the forecasting enterprise. Often, machine learning is the preferred method this. Machine learning in general is defined as "a subfield of computer science concerned with computer programs that are able to learn from experience and thus improve their performance over time" (Russel and Norvig, 2016, 693). The main purpose of machine learning is to detect patterns and correlations in data and derive predictions about future outcomes. Not the explanatory but rather the predictive power of a variable is important here.

Over recent years, there has been a sharp increase in studies predicting the outcome of court decision-making with machine learning. In what follows, I will discuss the most prominent approaches. However, I will narrow my discussion only to approaches that actually employ an *ex-ante* forecasting approach. Ex-ante prediction is defined here as any prediction performed using information that is available prior to a judicial decision. Studies that use the texts of a decision to arrive at their predictions (e.g. Sulea et al., 2017; Medvedeva, Masha and Vols, Michel and Wieling, 2018) are excluded, since decision texts are typically not available in advance of a decision.

One of the first attempts to use machine learning to make ex-ante predictions about judicial outcomes dates back to 2004. In a seminal study, Ruger et al. (2004) held a prediction tournament in which known legal experts competed against a simple machine learning algorithm, a classification and regression tree (Breiman et al., 1984).[2] The goal of their work was straightforward: predict the votes of individual judges as well as the final decision outcome of cases referred by lower courts to the US Supreme Court in advance of the release of the Supreme Court's decision. Their machine learning model only relied on observable case characteristics such as the type of respondent, the type of petitioner, or the issue area of a case. Their model was trained on data from the "Rehnquist Court" (1994 to 2002), and then the predictive performance was tested on the October 2002 term. Known legal experts have also attempted to predict the same outcomes. The result of this prediction tournament is impressive: the simple machine learning algorithm already outperforms the legal experts by correctly forecasting 75% of all outcomes, while the human experts only forecasted 59% correctly. With respect to individual judges' votes, the model was correct in 66.7% of the cases wile human experts correctly predicted 67.9%.

As a follow up of this work, Guimera and Sales-Pardo (2011) investigate whether it is possible to make predictions of a justice's vote based on the other justices' votes in the same case by analyzing the voting behavior of each natural court between 1953 and

[2] A similar approach, but in a much richer setting, is currently undertaken by Katz, Bommarito and Blackman (2017*a*), where the authors test the predictive ability of a large crowd (a large group of humans) compared with experts and algorithms.

2004. They use the votes of all judges in all previous cases, and the votes of the eight other judges in the current case to predict the vote of the ninth judge in the same case. They do not include any variables in their model, but solely rely on voting patterns. Their approach predicts 83% of the individual justice's votes correctly, but does not forecast the case level outcomes directly.

The work of Katz, Bommarito and Blackman (2017*b*) presents a major advance with respect to court prediction. The authors predict Supreme Court decisions over almost two centuries (1816-2015), forecasting 28,000 cases outcomes and more than 240,000 individual justice votes. Using random forests, a popular ensemble machine learning method and only relying on data available prior to the date of decision, Katz, Bommarito and Blackman's (2017*b*) model correctly predicts 70.2% of the court's overall affirm/reverse decisions and correctly forecasts 71.9% at the individual justice vote level. A recent study builds on their efforts and improves the prediction to about 75%, leveraging an even more powerful algorithm (AdaBoosted decision trees) for the prediction (Kaufman, Kraft and Sen, 2019).

5.3 Limitations of Existing Forecasting Approaches

All of these studies provide important insights about the predictability of court decision-making. However, I argue that existing forecasting approaches have two major limitations. First, existing ex-ante prediction models exclusively analyze and predict the US Supreme Court decision-making. This raises concerns about the external validity of previous work, and whether a similar prediction model could also be successfully applied to Kelsenian constitutional courts. In this regard, there are two issues. First, the US common-law system is guided by the norm of stare decisis, under which judges are supposed to decide cases based on similar precedents in the past. This leads to the expectation that just by how the legal system is constructed, there is supposed to be a high consistency between certain case-fact patterns. This "path-dependency" potentially facilitates the forecast, and might explain why even simple machine learning approaches (such as classification trees) already reach a high prediction accuracy (Ruger et al., 2004; Kastellec, 2010). As most European constitutional courts are under the civil-law system, there is no such thing as the norm of stare decisis. In other words, a European constitutional court judge is formally less bound to past case outcomes when making her decision in a current case. This absence of "path-dependency" should make it potentially harder for machine learning algorithms to detect and identify patterns between certain factors and outcomes. Second, some of the previous studies use the past voting behavior of individual judges to obtain predictions (e.g. Guimera and Sales-Pardo, 2011). Unfortunately, this rich source of information cannot be leveraged for most European constitutional courts due to the non-disclosure of individual judges' votes. Both points raise concerns whether legal prediction models can also be success-

fully applied to European constitutional courts. The first question this chapter will answer is, therefore, *whether similar predictive approaches already successfully applied to the Supreme Court also work in the European court setting*?

Second, none of the existing studies have explicitly evaluated the relative importance of the predictors, namely the variables used for the prediction. There is a long-standing debate about which factors influence judicial decision-making, and thus assist its prediction. Although nowadays, the traditional divide between the two "camps" of legalists on the one hand and realists on the other hand is less clear and not as stark as it has been before, there remains considerable disagreement on which factors exactly are important for legal prediction. Traditional legal scholarship still emphasizes the important role of jurisprudence and legal doctrine, and tends to downplay the role of non-legal factors. According to this notion, judges find the solution to a legal question or the case outcome by neutrally applying law through legal reasoning and interpretative methods. To exaggerate, in this regard law works as a set of static, natural, apolitical rules that can be mechanically applied to decisions. Or, as Dyevre (2008) characterizes it: "rules + facts = decision" (Dyevre, 2008, 27). This traditional legal perspective remains strong in the European constitutional court context. The German legal scholar Ossenbühl (1998) for instance states that the jurisdiction of the FCC is a decision of dispute by means and guided by methods of law not political judgment (Ossenbühl, 1998, 85).

By contrast, legal realists and political scientists argue that legal factors alone are not sufficient to fully explain and predict judicial decision-making. Attitudinalists for instance argue that judges are single-minded political actors whose decisions reflect their unconstrained policy preferences (Segal and Cover, 1989; Segal et al., 1995; Segal and Spaeth, 2002; Baum, 2009). Related, strategic accounts of judicial decision-making claim that judges are strategic actors who originally pursue policy-goals, but must adapt their behavior to external and internal constraints from other actors from time to time. Such constraints are, for instance, following public opinion to maintain their public support (e.g. Vanberg, 2005; Hall, 2014), or a strategic restraint from their own policy preferences in a separation-of-powers framework (e.g. Epstein, Knight and Martin, 2001; Bailey and Maltzman, 2011).

In this chapter, I do not aim to enter this (sometimes still) stylized debate. Instead, I want to explicitly test and tease out which factors actually contribute to the prediction of court decision-making. This idea is already noted in Martin et al. (2004), who write that "the best test of an explanatory theory is its ability to predict future events. To the extent that social science and legal scholarship seeks to explain court behavior, they ought to test their theories not only against cases already decided, but against future outcomes as well" (Martin et al., 2004, 761). Nonetheless, none of the existing ex-ante prediction models have explicitly teased out and quantified the contribution

of variables belonging to different strands of argument.[3] In this context, predictive modeling offers an excellent possibility to compare theories that seek to predict the same outcome (Cranmer and Desmarais, 2017, 149). In particular, predictive modeling "is an exceedingly simple means to highlight the extent to which the theoretically informed models anticipate reality, and which among those models does a better job of it" (Cranmer and Desmarais, 2017, 149).

For this reason, I will tease out the relative contribution of both political context factors and legal context factors for the prediction of court outcomes. I conceptualize political context factors as all factors that relate to the political aspects of court decision-making. Political context factors include the ideological position of the court, the public support for the court, the public opinion towards a certain issue upon which the court will decide, or any other political factor which social scientists have carved out in their work on judicial decision-making (see above). By contrast, legal context factors describe all non-political case characteristics associated with a case. These factors include the issue area of a case, the type of legal question that is raised, or the type of plaintiff or respondent. In other words, the legal context is rather understood as the legal baseline of a case in the absence of political factors.

Evaluating the contribution of political and legal context factors for the prediction of court decision-making can thus help us to gain a better understanding of court decision-making. The key argument here is that if traditional legal scholars are right, then the legal context of a decision should already be sufficient to predict a substantial part of court decision-making. Therefore, according to the pure legalist view, adding political context to the prediction should not improve the predictive performance of a forecasting model. However, if legal realists and social scientists have a point, the observable implication is that including political context into the prediction should increase the predictive power of the forecast. The second question this chapter thus addresses is *whether political context factors contribute to the prediction of court decision-making compared with legal context factors*?

To sum up, in this section I have discussed several existing ex-ante prediction models that use machine learning to forecast court decision-making. I have argued that there are two limitations in prior work: *a)* the exclusive focus on the US Supreme Court which raises concerns about the external validity of previous findings; and *b)* the lack of evidence that explicitly tests the contribution of both legal context and political context variables to the prediction of court decision-making. In the next section I present a research design that addresses these two limitations.

[3] Katz, Bommarito and Blackman (2017*b*) use variables belonging to legal and political context, but neither map their variables to these dimensions nor do they compare the variable's contribution.

5.4 An Ex-Ante Prediction Model for GFCC Decisions

In this section, I present a research design for an ex ante prediction of the decision-making of the GFCC that is able to carve out the relative contribution of legal context and political context factors for the prediction. My design addresses the two limitations outlined before. I discuss why the German Federal Constitutional Court is an appropriate study object as a European constitutional court, which data and variables I use to capture the legal context and political context of a case, and why I use the random forests algorithm for the prediction.

5.4.1 Case Selection: The German Federal Constitutional Court

The purpose of this chapter is to develop a forecasting model that *a)* predicts the decision-making of a constitutional court outside the US and *b)* compares the predictive contribution of legal and political context factors for the prediction. Here, the GFCC is analyzed. The case selection is motivated by three reasons. First, the GFCC is the archetype of the European Kelsenian constitutional court type and is considered as being one of the most powerful and influential constitutional courts world wide. It has served as a model for many newly established constitutional courts, e.g. in Eastern Europe. A prediction model that is suitable for the GFCC could also work as a blueprint for prediction models of these other courts. Moreover, the GFCC operates in a civil law system, and the individual votes of judges are mostly confidential. This means that one cannot simply predict individual judges' votes and aggregate them to make case outcome predictions. On these grounds, the GFCC represents a meaningful yet challenging study object from a predictive perspective. Third, the institutional power of the German court provides it with a strong institutional independence of other political actors, for instance with an appointment process of judges which requires a broad inter-party agreement. This makes it a hard-case scenario to test the importance of political context for the prediction: if we find evidence that political context matters for the GFCC, it presumably also matters for constitutional courts where the nomination procedure is more politicized (for instance, in France).

5.4.2 Data and Analytical Approach

The data used in this study were compiled as part of the Constitutional Court Database (CCDB) (Hönnige et al., 2015). The CCDB features 38 years (1972-2010) of data on decisions of the GFCC. Here, I use 2,910 proceedings (referrals) decided in this time frame. The court often bundles multiple proceedings in one main decision but decides on each of them individually (Wittig, 2016, 27). Thus, although being reviewed in the same main decision, the proceeding of petitioner A can be successful while the proceeding of petitioner B is not. I therefore follow common practice and treat the

proceedings and their respective outcomes as the level of analysis (Hönnige, 2009; Sternberg et al., 2015; Krehbiel, 2016).

The GFCC knows over 21 different proceeding types, which differ in the actors entitled to file an appeal, the possible causes of action, and also in their political importance and societal relevance. In my analysis, I concentrate on four proceeding types: *constitutional complaints*, *concrete reviews of statutes*, *abstract judicial reviews of statutes* and *Organstreit proceedings*. These proceeding types account for 98 percent of all proceedings decided by the GFCC. The proceeding types left out appear only rarely or are not a proceeding in the classic sense. Such proceedings include e.g. the procedure to impeach the Federal President.[4]

Constitutional complaints are the most common proceeding type (1941 proceedings in my data) accounting for around two-thirds of the observations in my data. Constitutional complaints allow citizens to assert their freedoms that are guaranteed by the constitution vis-à-vis the state, and can be filed by any person directly affected by a public law or act (after all other legal remedies are used). *Specific judicial reviews* are the second most common proceeding type (760 proceedings in my data). They can be filed by regular lower courts to review laws or statues if they are unsure whether this law is unconstitutional or not. *Abstract judicial reviews*[5] are typically filed by political actors such as the parliamentary opposition, often challenging governmental laws or statutes. Although abstract reviews are relatively rare (121 proceedings in my data), they often concern matters of political nature and hold great political and societal importance (Kranenpohl, 2010, 260). These type of proceedings are also called the "sword" of the opposition (Schneider, 1974, 222). Finally, *Organstreit proceedings* (88 proceedings in my data) may be filed if high state organs, or actors that are equivalent to such organs, disagree on their respective rights and obligations under the Basic Law. Similar to abstract reviews, they often raise questions of fundamental political issues that are relevant for the political system. Because abstract reviews and Organstreit proceedings only appear relative rarely, but are both considered as rather political proceeding types, I group them together in the analysis. Therefore, the final data contains three distinct data sets for each constitutional complaints, concrete reviews and abstract reviews/Organstreit proceedings.

Based on this data, I develop a separate prediction model for each of the three proceeding types, but with the same fixed set of predictors. This strategy is different to other court prediction models that rely on only one general model (Ruger et al., 2004; Katz, Bommarito and Blackman, 2017*b*) for all different kinds of decision types.

[4]Official annual statistics provided by the GFCC can be found at `https://www.bundesverfassungsgericht.de/SharedDocs/Downloads/EN/Statistik/statistics_2018.pdf?__blob=publicationFile&v=4`, accessed 12.04.2019.

[5]In line with Hönnige (2009), I also code Bund-Länder-Streits, a vertical conflict of competence between the federal and the state governments, as abstract reviews due to their equivalence as regards content.

However, I argue that my approach has several advantages. First, using different proceeding types but the same fixed set of predictors allows me to compare the models with respect to their predictive performance and the contribution of the same variables in a different proceeding context. It is thus possible to test whether, for instance, political context variables contribute considerably more for the prediction of political proceeding types than for the prediction of proceedings without a political context. Second, developing one prediction model for all distinct proceeding types requires the assumption that the data generating process is the same across all types. This would be a strong (and potentially incorrect) assumption, given that the proceeding types strongly differ in their character. Finally, using one general model for all proceeding types would result in a heavy bias towards predictors that best explain constitutional complaints, as this type account for the majority of the data. The final model would hence not be a general model for all different proceeding types, but a model that is good for predicting constitutional complaints. In turn, this would not be beneficial to tease out the relative contribution of legal and political context factors across different proceeding types.

5.4.3 Outcome Variable

The outcome variable (dependent variable) is a binary variable indicating the individual outcome of each proceeding. This variable is coded as a one if the GFCC decided in favor of the plaintiff and it is coded as zero if it decides against it. In other words, it indicates whether the plaintiff was successful or not. Following common practice, I consider a partial success to be a ruling in favor of the petitioner (Hönnige, 2007; Vanberg, 2005; Hönnige, 2009; Sternberg et al., 2015; Krehbiel, 2016, 2019). This binary coding scheme also allows me to compare my results with the findings of existing studies later.

In order to predict the outcome of a proceeding, I employ a number of predictors which represent the legal and political context of a proceeding. All of these variables can be used for an ex-ante prediction, since they all can be obtain *a priori* to a GFCC decision, and are thus exogenous to the final outcome. In fact, all information used for the prediction are publicly available the same day the plaintiff decides to submit the proceeding to the court. The model thus provides a substantial lead time.[6]

5.4.4 Legal Context Variables

I conceptualize the legal context of a decision as non-political, legal case characteristics associated with a decision. In other words, these factors should represent the "legal" or

[6]Lead time can be defined as the amount of time between a forecast is released and the actual occurrence of the event or outcome that is predicted.

"procedural" baseline of a case. This baseline can then be used to compare the predictive power of political context in the subsequent assessment. Representing the legal context of a proceeding, I include the following variables: the *decision type*, the *issue* area a proceeding, the *Senate* who is supposed to adjudicate, the *legal area* a proceeding is concerned with and whether proceedings are *grouped together* or not. The decision type describes whether the decision is, for instance, a main decision or a provisional order. The issue variable describes the topic of a decision and is coded according to the Comparative Agenda Coding scheme (e.g. macroeconomic issues, social insurance). The *legal area* of a proceeding describes the legal doctrine a decision is related to, for instance family law or asylum law. However, it does not contain information on the exact legal norms the court examines in a decision, because from an ex-ante perspective this information is not available in advance of a court decision. All of those variables are taken from the CCDB. I did not include information on the petitioner type or respondent type of a proceeding, because this information is already mostly covered by the proceeding type itself.[7] Table 5.1 provides a more detailed description of these variables with examples.

5.4.5 Political Context Variables

Political context is conceptualized as all factors that relate to the political aspects of court decision-making. The following predictors are used to represent the political context of a proceeding: the *ideological position* of the GFCC, the *salience* of a proceeding, the *popularity* of the opposition at the time of a decision, and a measure for the *perceived state of the economy* by German citizens as a measure of public economic mood. These political context factors are included because prior research of political scientists have found them to be important for the decision-making of the GFCC (Hönnige, 2007, 2009; Sternberg et al., 2015).

The ideological direction of the GFCC is measured on a common left-right scale using the Manifesto Common Space Scores (MCSS) (König, Marbach and Osnabrügge, 2013). To calculate the position of the court, I use the same measurement approach already explained in the previous chapter (Chapter 4.4.2). The importance of the GFCC's ideological position is demonstrated in previous work by Hönnige (2007, 2009). The following predictors are all related to public opinion and public support, and thus represent the direct effect that the public has on the decision-making of the GFCC. The salience of a proceeding, namely its importance for the public is measured by a binary variable indicating whether a proceeding is accompanied by an oral hearing or not. Vanberg (2005) uses this variable as a proxy measure for the degree of the

[7] These variables are not supposed to represent *all* factors that legal scholars or legal traditionalist consider as being the most important factors of court decision-making. Rather, the legal context factors should serve as a baseline to compare the political context with.

public awareness of a case, because "cases involving oral arguments are usually cases of great significance" (Vanberg, 2005, 103). Therefore, this variable is included as a political context factor because several studies demonstrate that the decision-making of the GFCC is affected by a proceeding's salience (Vanberg, 2005; Krehbiel, 2016, 2019). The popularity of the opposition captures the difference in the opposition's popularity relative to the popularity of the governments. This variable is included as political context variable because there is evidence that popular oppositions win their cases more often than oppositions with little public support (Sternberg et al., 2015). The data for this variable is taken from the German Politbarometer survey (Forschungsgruppe Wahlen, 2019). Finally, I capture the economic mood of the German public by measuring the perceived state of the economy. Evidence from the US Supreme Court shows that its decision-making is shaped by the economic state of the country (Brennan, Epstein and Staudt, 2009; Staudt and He, 2010). This could also be the case for the decision-making of the GFCC, although this causal relationship has not yet been tested. The economic mood variable is also part of the Politbarometer survey.

As an important note, I want to stress that model building and model specification is undertaken differently in predictive modeling than compared with classical inferential modeling. In predictive modeling, the inclusion of certain variables into a model is not guided by theory or expected causal relationships between the outcome variable and the predictors. Instead, generally all available information that could be somehow relevant for the prediction is included into a model, and the predictors are only modified to obtain a better prediction (to avoid over-fitting, for instance) or reduce computational burden (this process is called feature engineering in the machine learning literature (Hastie, Tibshirani and Friedman, 2009)). Following this, I did not make a specific effort to reduce the list of legal or political context variables, and that there is no doubt that some of them are correlated. Nonetheless, this (some would call it "kitchen sink") approach is not problematic for my analysis. The machine learning method I use does not suffer from the same problems that conventional regression analysis has with correlated predictors. Therefore, I am rather over-inclusive in adding predictors to the model. All variables are once more summarized in Table 5.1.

Table 5.1 – Legal and Political Context Variables Used for the Forecast

Legal Context	**Description**	**Example**
Decision Type	The type of the decision	Main decision, preliminary ruling
Issue	Issue area (Comparative Agenda Coding Scheme)	Macroeconomic Issues
Senate	Senate dealing with a proceeding	Senate I or II
Legal Area	Legal area a proceeding is concerned with	Labor law
Grouped	Whether a proceeding is grouped with others or not	0 = not grouped, 1 = grouped
Political Context		
Salience	Whether there was an oral hearing before the proceeding	0 = no oral hearing, 1 = oral hearing
Popularity Opposition	Difference in popularity of opposition relative to government	1 = very unpopular, 11 = very popular
Economic Perception	Perceived state of the economy	1 = very good, 5 = very bad
Ideological Direction	Ideological direction of the Court (MCSS scores)	-1 = left, 1 = conservative

5.4.6 Method

To build my prediction models I rely on random forests (Breiman, 2001). Random forests is a popular ensemble classifier and is among the most commonly used machine learning algorithms for supervised learning (Hastie, Tibshirani and Friedman, 2009). Although random forests and similar tree-based methods were long neglected by the field, they become increasingly used in the social science context (e.g. Green and Kern, 2012; Beauchamp, 2017; Montgomery and Olivella, 2018; Jones and Lupu, 2018; Bonica, 2018; Kaufman, Kraft and Sen, 2019). In what follows, I give a brief introduction to random forests. For a recent, non-technical introduction of tree-based methods for political scientists see Montgomery and Olivella (2018).

A random forest uses an ensemble of classification and regression trees (CART). CART is a supervised machine learning algorithm that iteratively divides the outcome variable observations into increasingly homogeneous groups using the predictor variables through binary splits (this is called recursive partitioning). CARTs are known to be

notoriously unstable, meaning that already small changes in the data can lead to completely different splits. They also tend to be biased towards continuous covariates (Hothorn, Hornik and Zeileis, 2006). A random forest overcomes these limitations by using an ensemble of many randomized trees that leverage two forms of randomness: *bagging* – short for *b*ootstrap *agg*regation – (Breiman, 1996) and *random substrates* of the predictor variables. The underlying idea is that many uncorrelated trees are constructed and then aggregated. The procedure to construct one (out of many, typically between 500 and 1,000) tree in random forest is as follows.

First take a random sample with replacement, typically containing about two-third of the observations, while the remaining (one-third) of the observations are hold "*out-of-bag*" (*oob*). On the bootstrapped sample, construct a decision tree. At each node of the tree, randomly select m out of p predictors, where m is a hyper-parameter and is typically chosen by the researcher. Out of these m randomly selected predictors (random substrates), the one that gives the best classification at this node is used to partition the data. This process is repeated at each subsequent node, such that at each node a random substrate of m predictors is chosen. The random selection of splitting variables allows predictors that were otherwise outplayed by their competitors to enter the ensemble. This has the benefit of obtaining less correlated and thus, more robust trees. The model then averages predictions over all trees, whereby the predicted class of an observation is calculated by majority voting of the *oob*-predictions for that observation. In Appendix D.1 I outline the random forest algorithm in further detail.

There are four reasons to use random forests and not another machine learning classifier. First, random forests has proved to be a strong learner in a comparable study (Katz, Bommarito and Blackman, 2017*b*). Second, in an analysis of judicial decisions and legal rules using a single decision tree, Kastellec (2010) finds that the tree structure actually mirrors the "hierarchical and dichotomous structure that often seems apparent in judicial opinions" (Kastellec, 2010, 210). Third, an experiment using several popular classification algorithms shows that random forests outperforms other algorithms.[8] Fourth, random forests is very efficient in detecting non-linearities in the data without requiring the specification of any functional form and also provides built-in estimates of variable importance. All of these aspects make random forests the optimal method choice for my prediction task.

[8]I test the predictive performance of Classification and Regression Tree (CART), Random Forests, Support Vector Machines, k-nearest neighbors, extreme gradient boosted trees and regularized logistic regression on the constitutional complaints data. Predictive performance was assesses using 10-fold cross-validation without hyper-parameter tuning. Cross-validation was performed such that every algorithm received exactly the same data slices, to make the model comparison as fair as possible. This constitutional complaints data set is used because it has the largest N. The classification results are found in the Appendix D.2.

5.5 Results

In this section I present the results of the ex-ante prediction of proceedings decided by the GFCC. The section is divided into two parts. In the first part, I use random forests to predict the outcomes of each proceeding type in my data using the same fixed set of input variables. I show that a combined model consisting of legal and political context variables yields to a higher predictive performance than a model using legal context factors alone. Moreover, I conduct a simulation that shows that the increase in predictive power is not just an artifact of adding more variables to the model. In the second part, I open the black-box of the prediction model by comparing the predictive importance of the predictors across the proceeding types. The section concludes with a discussion of the ability of random forests to detect interesting non-linearities in the data that conventional regression analysis might have overlooked.

5.5.1 Predicting Proceeding Outcomes of the GFCC

In order to tease out the relative importance of legal and political context for the prediction of GFCC decision-making, I run a series of experiments. For each of the three proceeding type data sets, two different random forests are developed: a *legal model* only featuring the legal context variables, and a *combined model* featuring the legal context *and* political context variables. The legal model here serves as a "legal" baseline and is used to evaluate the predictive performance one can expect by just using the legal and procedural context of a given proceeding. The combined model is used to assess whether and to what extent political context can improve the model's predictive power. To repeat, the observable implication with respect to this comparison is that if legal realists and political scientist are right by arguing that political context matters, then the inclusion of this context into the prediction model should increase its predictive capability. If political context is irrelevant for the prediction, then its inclusion should not change model performance. At this point I want to highlight again that my analysis does not seek to disentangle the causal effect of legal and political context on judicial behavior, nor to test whether political context outweighs legal context.

For a fair model comparison, a robust model performance evaluation is of crucial importance. In predictive modeling, the goal is to obtain an estimate of *true error* (also known as *generalization error*). True error is a measure of how well a model can predict outcomes of previously unseen data (Efron and Hastie, 2016; Cranmer and Desmarais, 2017). An estimate of true error is important in practice, as it allows one to check whether a model generalizes well to unseen data or just memorizes the patterns in the training data (i.e. over-fitting).

With this in mind, I provide two performance evaluations of the models. In the first performance evaluation, I report the model's performance based on their aggregated

cross-validation score *without hyper-parameter tuning*. Cross-validation, when correctly applied, can be used to obtain an almost unbiased method of true error without setting aside additional test data (see Cawley and Talbot, 2010; Efron and Hastie, 2016). However, note that combining cross-validation for model tuning and to estimate true error at the same time leads to serious misreporting of performance measures (Neunhoeffer and Sternberg, 2019).

In my experiments, on each of the three data sets[9], I perform (stratified) 10-fold cross-validation and hold only the hype-parameter of random forests fix at $m = \sqrt{p}$, where p is the number of predictors and m is the number of random substrates. This value is recommended by Hastie, Tibshirani and Friedman (2009) for classification problems using random forests (Hastie, Tibshirani and Friedman, 2009, 592). In short, cross-validation refers to randomly dividing a data set into K about equally sized folds, where each fold contains about $\frac{N}{K}$ observations. A random forest classifier[10] is then trained K times on all but the kth fold, where k runs from 1 to K. In every iteration, a performance measure is used to evaluated the model performance on the kth fold (holdout/test fold) that was not part of the training. Finally, the average (across the K folds) of a performance measure is reported, which is the aggregated cross-validation score. However, as Cawley and Talbot (2010) show, even if cross-validation is applied correctly, the variability of such hold-out methods can lead to over-fitting in a finite sample nonetheless (Cawley and Talbot, 2010, 2084-2086). This, in turn, would lead to reporting an overly optimistic model performance.

For this reason, I report the results of an *out-of-sample* prediction as a second evaluation. Out-of-sample prediction is considered as the gold standard to obtain an unbiased estimate of true error (Hastie, Tibshirani and Friedman, 2009, 220). In out-of-sample prediction, a model is trained on a training set and then used to predict the observations of a test set (the out-of-sample data). During the training process, hyper-parameter tuning can be performed. This is because due to the strict split between training set and test set, the final model evaluation cannot suffer from over-fitting since the test set never occurred in the model building process.[11] For each of the three proceeding data sets, I randomly divide the data into a training set, containing 75 percent of the

[9] I only use the training data sets (see next paragraph) to obtain the cross-validated performance scores. This ensures that each model only has access to exactly the same amount of information. Taking the cross-validation scores of the whole data set would constitute an unfair model comparison, because then models of the cross-validation procedure would have seen more data than the models of the out-of-sample evaluation.

[10] The random forests are estimated using the *R* packages *caret* (Kuhn, 2008) and *ranger*, a fast (parallel) implementation of random forests (Wright and Ziegler, 2015). For each random forests, $1,000$ trees ($ntree = 1,000$) are grown because simulation studies suggest that smaller values can result in unstable estimates under certain circumstances (Strobl et al., 2007; Strobl, Hothorn and Zeileis, 2009).

[11] Of course, a model can over-fit the training data, although the over-fit will lead to a poor out-of-sample prediction.

observations, and a test (out-of-sample) set with the remaining 25 percent.[12] On each of these training data sets, I train two random forests models: one using only the legal context variables, and one using both. Tuning is performed to find the best set of hyper-parameters using five-fold cross-validation and random grid-search. These models are then used to predict the outcomes of the observations in the test set.

As performance metrics, I report the accuracy and Cohen's *Kappa* (Cohen, 1960). Accuracy is simply defined as the sum of true positives and true negatives divided by the overall number of observations. The *Kappa* metric takes into account the class distributions and is based on the observed accuracy (accuracy of the classifier) and the expected accuracy (expected accuracy of a random classifier). In Appendix D.3, I report additionally the receiver operating characteristic area under the curve (ROC AUC) and the precision recall area under the curve (PR AUC).[13] In order to calculate the accuracy, the conventional threshold of 0.5 is used for positive predictions. The majority class (baseline) is also reported to compare the performance of the random forest with respect to a naive learner. A naive learner is defined here as a classifier who always assigns the majority (most frequently occurring) category of the training set.[14]

Table 5.2 reports the model evaluations based on the aggregated cross-validation scores across the three different data sets. All columns labeled as "legal" report the performance of the legal model and all columns labeled as "combined" report the performance of the combined model. The corresponding confusion matrices of each model are provided in Appendix D.4. The best models according to the respective performance measure are highlighted in bold. We see that the legal model itself is already sufficiently good to predict a substantial part of all decisions correctly, outperforming the baseline for all proceeding types. The weighted accuracy across all proceeding types is 62.55 percent.[15] Using the weighted accuracy is important to obtain the overall percentage of correctly-predicted proceedings, since the proceeding data sets are of different sizes.

However, and this is the important observation, we also see that for all proceeding types, the model performance is improved when the political context variables are added (combined model). Across all proceeding types, the weighted accuracy improves to 72.16 percent. This means that using the combined model, it is possible to correctly forecast approximately three out of four outcomes. On average, across all proceeding

[12]The randomly created training and tests sets are of the following sizes (N of training set, N of test set): Constitutional complaints 1,455, 486; concrete reviews 570, 190; Abstract reviews and Organstreit proceedings 156, 53.

[13]The calculation of all performance metrics is defined in Appendix B.3.1.

[14]Note that it only makes sense to report the baseline for accuracy. This is because the *kappa* measure already takes into account the majority class in its calculation (*kappa* = 0 means that majority voting takes place).

[15]Calculated by weighting the accuracy of the respective proceeding type with the number of observations of this type.

Table 5.2 – Model Evaluation Based on Aggregated Cross-Validation Scores

	Accuracy			Kappa	
	Legal	Combined	Baseline	Legal	Combined
Constitutional Complaints	60.14	**68.93**	53.47	0.20	**0.37**
Concrete Review	68.42	**80.18**	67.02	0.08	**0.50**
Abstract Review/Organstreit	63.54	**73.04**	60.26	0.19	**0.41**
Weighted Performance	62.55	**72.16**	57.50	0.17	**0.41**

Note: Model performances of the legal model and the combined model based on the aggregated 10-fold cross-validation scores. The random forests were built with a fixed m. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier that always votes the majority category of the training set. The best performances are highlighted in bold.

types, adding the political variables to the classifier increases the predictive performance by about 9.61 percentage points in terms of weighted accuracy and 0.24 in terms of *Kappa*. The higher *Kappa* values of also indicate that the better performance of the combined model is robust when considering the class distributions. The largest performance increase is for concrete reviews, where the addition of political context improves the predictive performance by +11.76 percentage points in accuracy. We can also see that the performance is considerably increased for the political proceeding types (abstract review/Organstreit proceedings): here, the addition of the political context variables improves the prediction from approximately two out of three to correctly predicting around three out of four outcomes (+9.5 percentage points). This finding makes intuitively sense from a political science perspective: these proceeding types often deal with political matters, so that the potential influence of political context is expected to be strong here.

These results are also confirmed when looking at model evaluation using out-of-sample prediction in Table 5.3. We, again, observe that for all different proceeding types, the combined model has a higher predictive power than the legal model. Across all proceeding types, the weighted accuracy of the combined model is 76.41 percent, and thus about +7.94 percentage points better than compared with the legal model (68.47 percent weighted accuracy). Here, the performance improvement is the highest for the political proceeding types (+16.98 in accuracy). Note that both model performances (the legal and the combined model) have higher scores when using the out-of-sample prediction evaluation. This is the case because although I split the data randomly into training set and test set for the out-of-sample prediction, due to random chance we observe differences between the cross-validation scores and the out-of-sample scores (different splits of training and test set might result in different scores). These differences are so strong because the overall N of the data sets is not very large (the abstract review/Organstreit proceedings data set only contains 209 observations overall). At this point, I want to emphasize that one should not over-interpret the

Table 5.3 – Model Evaluation Based on Out-of-Sample Prediction

	Accuracy			Kappa	
	Legal	Combined	Baseline	Legal	Combined
Constitutional Complaint	66.67	**74.49**	52.67	0.33	**0.49**
Concrete Reviews	75.26	**81.05**	65.79	0.41	**0.57**
Abstract Reviews/Organstreit	60.38	**77.36**	58.49	0.17	**0.52**
Weighted Performance	68.47	**76.41**	56.52	0.34	**0.51**

Note: Model performances of the legal model and the combined model based on out-of-sample prediction. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

exact performance scores, but that my findings rather demonstrate a general tendency independent of the performance evaluation approach: on average, adding information about the political context of a proceeding improves the prediction.

5.5.2 The Predictive Power of the Combined Model Versus White Noise

A critical reader might wonder whether the improvement of predictive performance that we observe when adding the political context variables to the legal model is due the predictive power of these variables or due to simply adding more variables (like the expected increase in R^2 in the regression context). In order to convince such critical voices and demonstrate that the political context variables actually improve the predictions because they are related to the outcome of a decision, I conduct an additional experiment. In this experiment, I substitute the four political context variables with randomly drawn variables unrelated to the outcome. I refer to these randomly drawn variables as "*noise features*" in the following.

The noise features are constructed by randomly sampling from a multivariate normal distribution with means equal to the means of the original variables and the corresponding variance-covariance matrix to capture the structure of the variables to each other. Accordingly, these randomly-sampled variables mirror the distribution and correlation structure of the original variables, but are not correlated with the other features or the outcome. For each proceeding type data set, I remove the original political context variables and replace them by noise features. The final data sets thus only include the legal context variables and the four noise features. On each of these data sets, I then run random forests models using the 10-fold cross-validation procedure without hyper-parameter tuning which has already been used to obtain the aggregated cross-validation scores reported in Table 5.2. I call these models the "*random models*", due to the four randomly created noise features in it. The observable implication is that if the performance of the combined models in the main analysis just improves because more variables are added, then we should also observe an increase in the predictive performance of the random models, although the noise features should have

Table 5.4 – Model Evaluation of Legal, Combined and Random Model based on aggregated cross-validation scores

	Accuracy			Kappa		
	Legal	Combined	Random	Legal	Combined	Random
Constitutional Complaints	60.14	**68.93**	62.75	0.20	**0.37**	0.24
Concrete Review	68.42	**80.18**	68.06	0.08	**0.50**	0.09
Abstract Review/Organstreit	63.54	**73.04**	66.58	0.19	**0.41**	0.26

Note: Model performances of the legal, combined and random model based on aggregated cross-validation scores. The legal and combined models are the same as in Table 5.2. The "random model" only contains the legal context variables plus four randomly created noise features. The best performances are highlighted in bold.

no predictive power by construction. However, if the political context variables actually contribute to the prediction, we should observe the combined model to perform better than the random model.

Table 5.4 reports the result of this experiment. We can observe that the combined model still performs better than the other two models. In fact, the performance scores of the legal model and the random model are about the same for constitutional complaints and concrete reviews. Interestingly, for abstract reviews and Organstreit proceedings, the random model is about three percentage points better than the legal model. For these proceedings, adding "white noise" improves the prediction, although not as much as the original political context variables. A possible explanation for this is provided by Bishop (1995), who shows that adding noise to data can have a similar effect like l_2 regularization if the predictive method is over-fitting. However, interestingly Bishop (1995) describes that the random noise is added by "adding a random vector onto each input pattern" (Bishop, 1995, 109). In simple terms, this means that for each individual data point of some features X_1, X_2, X_3, random noise is added like $X_1 + z, X_2 + z, X_3 + z$, where X represents the original predictors and z is a random noise vector. By contrast, what I do is adding extra noise features, so that the data used for the prediction then is like $X_1, X_2, X_3, Z_1, Z_2, Z_3$, where each Z represents a randomly created noise feature. While being beyond the scope of this dissertation, this phenomenon is, to the best of my knowledge, not well known in the field of political science, and needs to be examined in further detail in a follow-up study.

In fact, the improvement of prediction by adding white noise for the political proceedings data might be a hint that the legal model in Table 5.2 over-fits, and therefore the addition of random noise makes it harder for the random forests to over-fit the data. I obtain similar findings when using out-of-sample evaluation instead of the aggregated cross-validation scores (Appendix Table D.5) and when replacing the draws from a multivariate normal distribution with draws from a standard normal distribution (such that the four added randomly sampled noise features are not related to each other at all).

5.5.3 An Alternative Out-of-Sample Prediction

In order to further demonstrate the robustness of my findings, I provide an additional out-of-sample prediction in Appendix D.6 where I take into account the time dimension of the data. Randomly dividing the data into training set and test set requires assuming that the data is *iid* (independent and identically distributed). The *iid* assumption might be violated using data with a clear time dimension (the data set covers 1972 to 2010). For this reason, I split the data into a training set and a test set where all observations before 2005 are assigned to the training set and all observations after 2005 are assigned to the test set. This test set is then used for the out-of-sample prediction. I did not use this split approach in the main analysis because splitting by an (arbitrary) point in time results in different train/test set size ratios. To illustrate, due to the split in 2005, the test set of abstract reviews/Organstreit proceedings contains around 19 percent of all observations (33 observations of 209), while the test set of the constitutional complaints contains only 8 percent of all of the observations (150 observations out of 1,941). This is because the number of proceedings decided by the GFCC is not equally distributed over time. Accordingly, a fair model comparison is difficult because the information each classifier has access to differs in terms of percentage of the overall data. Nonetheless, using the additional out-of-sample prediction the patterns of the main analysis are confirmed: adding political context improves the prediction of GFCC decision-making.

The results of this section lead to several conclusions. First, the findings for the US Supreme Court – that a machine learning model can successfully predict judicial decision outcomes – can be generalized at least to the German Constitutional Court, an archetype of the Kelsenian European Constitutional Courts. Similar machine learning approaches can reach similar accuracies. Across all proceeding types, the weighted accuracy of the combined model is 76.41 percent (out-of-sample prediction) and 72.16 percent (aggregated cross-validation scores). This is very close to the achieved performances of Ruger et al. (2004) with 78% and better than the achieved 70% of Katz, Bommarito and Blackman (2017*b*), who use over 95 predictors and heavy feature engineering. The first research question of this chapter – whether a machine learning classifier can correctly predict GFCC decision outcomes – is thus to be answered with a clear yes.

Second, I also find evidence that political context (including public opinion) improves the prediction of all proceeding types, and thus support for the second research question – whether political context factors contribute to the prediction of court decision-making compared with legal context factors. This is a strong and interesting finding, because a part of the German legal scholarship still considers the GFCC's decision-making as totally apolitical. (Böckenförde, 1976; Ossenbühl, 1998). I want to emphasize again that this does not mean that political context outweighs the importance procedural characteristics or other legal aspects of a proceeding. Instead, just the ensemble of legal

and political variables collectively contributes to the prediction in the combined model. To further investigate the role of legal and political context, I look at each variable's importance for the forecast in the next section.

5.5.4 The Importance of Legal Context and Political Context

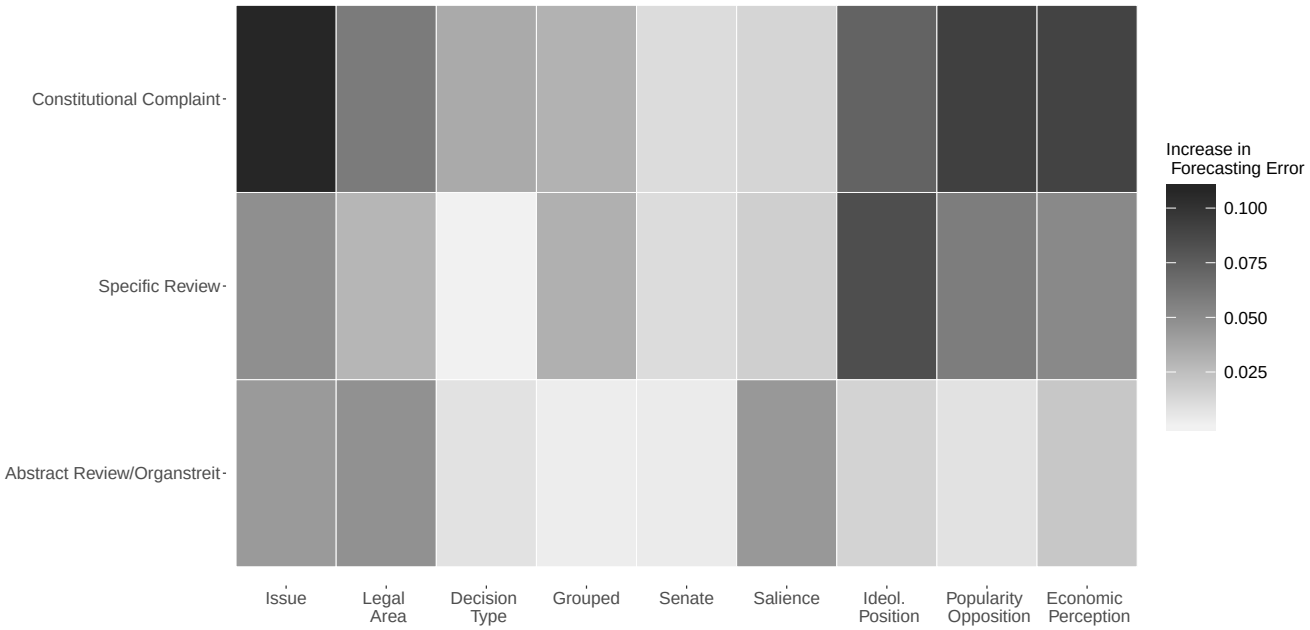
Which of the variables contribute to the prediction? Is there any variation in their importance across proceeding types? The importance of a variable in random forests can be obtained via its *variable importance*. Variable importance (also known as permutation importance) is a measure for the mean increase in the oob error if the values of a given predictor are randomly permuted. The idea behind this is straight forward: If the values of a predictor are randomly permuted and the oob error remains constant, the predictor is regarded as unimportant. By contrast, the larger the increase in oob error when a predictor has been permuted, the more important this predictor is for the forecast (Hastie, Tibshirani and Friedman, 2009, 593). Figure 5.1 shows the variable importance of all predictor variables on the horizontal axis with the respective proceeding type on the vertical axis of the heatmap. The darker a cell in a heatmap, the higher the variable importance of the given predictor for the respective proceeding type. The forecasting error of constitutional complaints increases, for instance, by about six percent if the values of the issue variable are randomly permuted, and thus withheld from the prediction.[16]

Figure 5.1 shows a considerable variation in the predictor's importance across the proceeding types. There is not a single predictor that is of equally strong importance for all proceeding types. The issue of a decision is a important predictor for constitutional complaints and concrete reviews, but not so much for abstract reviews/Organstreit proceedings. Some issues seem to be especially important in this regard. Not knowing whether the issue "education" or "law and crime" is present in a constitutional complaint proceeding, for instance, increases the forecasting error by about 1.8% and 2.1%, respectively (not shown in the graph). Interestingly, the ideological position of the GFCC is the most important predictor for concrete reviews, but not so important for the political proceeding types. In line with what we would expect theoretically, we also observe that the political context variables contribute more than the legal context factors to the prediction of these political proceeding types. Again, I want to highlight that variable importance is not equivalent to a causal relationship between a predictor and the outcome variable.[17] Nonetheless, it can help us to gain a deeper understanding

[16]For the sake of terminology, it is important to note that oob variable importance does not measure the increase in forecasting error if a certain predictor is excluded from the model. This is because if the model was rebuild without this predictor, the model could put more emphasis on other predictors, which then became surrogates (Hastie, Tibshirani and Friedman, 2009, 593).

[17]In addition, some of the predictors are correlated which can complicate the interpretation of the variable importance (Strobl et al., 2007, 2008).

Figure 5.1 – Heatmap of variable importance per proceeding type



Note: The different predictors are displayed on the horizontal axis. The different types of proceedings are shown on the vertical axis. Darker fields indicate a higher importance of the respective predictor for the respective proceeding type. The variable importance is obtained from the combined models from Table 5.3 to enable a comparison of legal context and political context predictors.

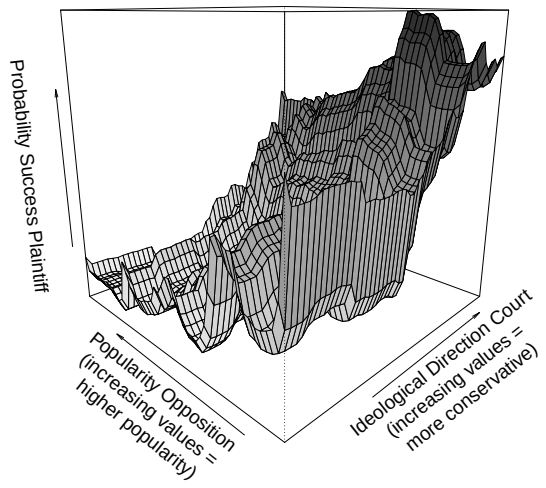
of the factors which drive the prediction, and can hint towards interesting relationships. In the next section, I will look at how certain predictors increase the winning chances of the plaintiff, which is something we cannot infer from variable importance plots. This information is contained in partial dependence plots.

5.5.5 Partial Dependencies and Non-Linear Relationships in the Data

Partial dependence plots are a method to visualize the partial relationship between predictors and the outcome in forecasting models. In short, such plots give a graphical representation of the marginal effect of a variable on the predicted outcome, after accounting for the average effects of the other predictor variables (Hastie, Tibshirani and Friedman, 2009, 369).

Figure 5.2 shows the partial dependence plot for the interaction between the ideological direction of the GFCC and the popularity of the opposition on the probability of a petitioner success. I focus at this interaction because the variable importance plot in Figure 5.1 shows that these variables are important predictors of concrete reviews. Moreover, these are political context variables that hold the most importance for predicting a rather apolitical proceeding type. Thus, it is a surprising finding that warrants further investigation.

Figure 5.2 – Partial dependence plot for ideological direction of the GFCC conditional on the popularity opposition/government on the plaintiff's success probability for concrete reviews



Note: Partial dependence plot for the interaction between the popularity of opposition/government and the ideological direction of the GFCC on the probability of plaintiff success in concrete reviews. The combined model for concrete reviews from Table 5.2 was used for the calculation. The graph shows a clear non-linear relationship between the outcome and the two predictors.

We can draw several conclusions from the partial dependence plot. First, there is a negative association between oppositional popularity and petitioner success, indicated by the flat surface in lower left part of the figure. However, this effect is conditional on the ideological direction of the GFCC: the more conservative the GFCC, the higher the likelihood of a petitioner's success (indicated by the sharp rise in the upper right). In other words, the winning chances of a petitioner in this scenario are the lowest if the opposition is very unpopular and the GFCC is rather left, whereas the winning chances are the highest if public support for the opposition is low and the court is rather conservative. This is an important observation, because these results suggest that the rather apolitical proceeding types such as concrete reviews might not be per se as apolitical as one thinks. Second, and more important, the effect between the two predictors on the outcome is clearly non-linear. This non-linearity would not be captured by conventional approaches such as logistic regression, at least not without specifically specifying the functional form of this relationship in the systematic component. Machine learning approaches such as random forests learn these non-linearities in the data without the need to be pre-specified by the researcher.

Partial dependence plots of other predictors show that the directions of how these variables are related to other variables or the outcome are largely as one would expect. The salience of a proceeding, for instance, strengthens the effect of other political context predictors such as the ideological position of the GFCC. This is in line with

existing political science research showing that judges behave differently in salient than in non-salient cases (Vanberg, 2005). Furthermore, the perception of the current state of the economy by the public plays a greater role if the main issue and sub issue of a case is an economic one. This is a relationship that makes intuitively sense. One of the important lessons of this chapter is that predictive modeling can help researchers to find (non-linear) relationships which conventional methodological approaches might have overlooked. In fact, most of the relationships between inputs and outcome do not display the typical *S*-shaped curve of e.g. logistic regression models, the model which is most often used to analyze binary outcomes. Machine learning approaches are, therefore, a fruitful approach to identify interaction effects or other non-linearities in the data.

5.6 Conclusions and Implications

In this chapter, I highlighted the ability of machine learning to ex-ante forecast decisions of the GFCC. I demonstrated that it is possible to correctly predict 76.40 percent of all outcomes of over 2,900 GFCC proceedings decided between 1972 and 2010 using only data that is available prior to a proceeding. In particular, I did not use any information which stems from decision texts, court statements or press releases or any other source that only becomes available after the actual decision outcome is released. Such a forecasting model is a novelty in European court research, and does not yet exist for the GFCC or any other European constitutional court.

I make two contributions. First, I confirm the external validity of similar work on the US Supreme Court and show that the decision-making of a European Kelsenian Court type can also be correctly forecasted by means of an algorithm. This is an important result, because the predictive setting for most of the European courts is more challenging since no individual voting records of justices are available. Second, and this is unique to my analysis, I explicitly test the predictive contribution of legal context and political context variables to the forecast. I find that legal context is, on average, a relatively good predictor proceeding outcomes. Moreover, I find that the predictive performance is improved when the political context of a decision is leveraged. Constitutional court decision-making is thus best characterized by the ensemble of legal and political context factors.

Beyond the application to the GFCC, my findings have other important implications with respect to legal philosophy and the value of machine learning approaches for the field of judicial politics and political science in general. What does it mean for our understanding of law and judicial decision-making if a relatively simple machine learning algorithm can correctly predict a substantial number of judicial outcomes? While this might appear alarming first, I argue that in fact, this is a sign of consistent judicial decision-making of the GFCC. If an algorithm can correctly predict outcomes,

it means that on average, similar proceedings with similar case characteristics are decided in a similar way. This consistency in judicial decision-making is important for the basic functioning of the rule of law. Therefore, for the sake of legal certainty, it is desirable that cases with the same context lead to the same judicial outcomes on average. Moreover, no algorithm could in any way substitute for the important work that judges do in their reasonings.

My findings have another implication for an important group beyond academia: the world of plaintiffs before the GFCC. For lawyers, politicians or ordinary citizens, the expected outcome of a case, namely the (perceived) probability of winning or losing, plays a crucial role in a plaintiff's decision to appeal or not. Given that a predictive model of GFCC decision-making can be improved over time and with more and possibly richer data, my results are beneficial for practicing attorneys and their clients likewise. In fact, such a model would also have consequences for the political system: for instance, the opposition would not only consider political factors in their decision to appeal to the GFCC or not, but would also be able to refrain from appealing cases where the success probability is low.

Finally, my analyses demonstrate the value of predictive modeling for social science: machine learning can help to identify patterns which conventional methodological approaches might overlook. This is especially important with respect to non-linearities in the data. Thus, even when the goal is causal inference, such forecasting approaches can help to identify undiscovered patterns in the data and therefore, can lead to new research questions. What is the causal mechanism that links the perception of the economic shape in Germany to its outcome? Why is the ideological position of the GFCC the most important predictor for concrete reviews, a proceeding type most often only dealing indirectly with political matters. While I do not argue that machine learning will replace conventional statistical social science methods, algorithmic procedures will become increasingly common as a supplementary tool in the tool box of quantitative social scientists.

CHAPTER 6

Conclusion

6.1 Summary, Implications and Answers

This dissertation began by stating that constitutional review is a key feature of modern liberal democracy. Nonetheless, despite constitutional review is a hallmark of democratic governance in both established and newly-formed democracies, legislative compliance with judicial decisions cannot be taken for granted due to the inherent institutional weakness of constitutional courts.

The aim of this dissertation was to understand the importance of public support for effectiveness of constitutional review in context of the relation between *court-government* and *court-public*. I drew on the comparative judicial politics literature on separation of powers, public support and legislative noncompliance and extended existing theory in two regards. First, I argued that not all courts possess about the sufficient level of public support that is necessary to ensure legislative compliance. Varying degrees of public support strongly affect the leverage that courts posses in judicial-legislative and judicial-public relationships. Second, I argued that courts will take active measures in the form of the institutional tools at their disposal when they expect legislative noncompliance. One such tool is decision language, whose strategic usage allows judges to pressure the government or hide likely noncompliance from public view if necessary.

I test these arguments empirically by combining classical inferential methods such as survey experiments with novel data on court decision-making and methodologies from the field of machine learning and computational linguistic. Throughout the chapters, I employ a comparative perspective and test my arguments using data on the German

Federal Constitutional Court, a court with strong and robust levels of public support, and the less popular French Conseil Constitutionnel.

My empirical evidence shows that considering varying degrees of public support and the institutional tools of judges is indeed beneficial to obtain a more accurate understanding of how judges behave in judicial-legislative and judicial-public interactions. I draw three main conclusions. First, court decisions can legitimize public policies, albeit only if the court itself is perceived as a legitimate institution. Second, courts are more attentive to the political environment of a decision than previously thought: depending on their degree of public support, they actively adapt the language of their decisions as a function of the risk of legislative noncompliance and their institutional support. Third, public support (and other political context factors) is important for judicial decision-making not only from a causal, but also from a predictive perspective.

These findings emphasize that the institutional setting of courts and country-specific context are important aspects that must be considered theoretically and empirically. In what follows, I will provide a brief overview of the examined research questions and the central findings of each chapter. I also highlight how these are embedded in the general framework of courts, governments and the public in the context of constitutional review.

Chapter 2: Constitutional courts as opinion leaders. In Chapter 2, I directly investigated the impact of public support on institutional legitimacy. The research question in this regard was *to what extent constitutional courts can shape public opinion on governmental policies*. Using a comparative survey experiment conducted in France and Germany, I found that public opinion, even amongst those with strong prior attitudes, is affected by court decisions. Court decisions can shift public opinion, but only if the diffuse support, and thus the institutional legitimacy of a court is sufficiently high. The main implication of this chapter is thus that varying degrees of public support are crucial for understanding the legitimacy-conferring ability of courts. This chapter, therefore, sheds light on the *court-public* relationship and the court's power to affect public opinion that comes along with institutional legitimacy.

Chapter 3: The automatic detection of vague language in constitutional court decisions. *How can we automatically measure vague language in written court decisions?* Chapter 3 was devoted to the computer-assisted measurement of vague language in court decisions. I argued that currently used measurement approaches are insufficient because they ignore the specificity of judicial texts. To overcome the limitation of existing work, I developed the concept of judicial policy implementation vagueness as a particular form of vagueness unique to judicial decisions. Based on recent advances in computational linguistic, I demonstrated that this concept can be measured with a dictionary-based approach using word embeddings and a machine learning classifier trained and benchmarked on a novel large self-collected data set. Because both mea-

surements are tailored to the judicial domain, they outperformed existing measures of vague language. This chapter indirectly contributes to the study of the relationship between *court-government* because it is the central dependent variable for the following chapter. The main message of this chapter is that researchers must carefully consider whether methodological approaches developed for one domain can also be applied to another without domain adaption, and that often interdisciplinary work is promising to overcome such methodological obstacles.

Chapter 4: Why do courts craft vague decisions? In Chapter 4, I focus on the relationship between *court-government* and *shed light at the ability of courts to strategically manipulate the amount of vagueness in their decision texts*. The main theoretical argument with respect to the indirect effect of public support in this relationship is that courts actively use the institutional tools at their disposal to maximize their utility when threatened with legislative noncompliance, but how they use these tools ultimately depends on their public support. I used my novel measures for the judicial policy implementation vagueness of a decision and found that courts, depending on whether they are popular or not, either write specific decisions to pressure the legislator or write vague rulings to hide noncompliance from public view. This chapter thus demonstrated the importance of my two fundamental arguments: varying degrees of public support matter, and courts are active actors who use the institutional tools at their disposal to deal with likely noncompliance.

Chapter 5: How to forecast constitutional court decisions? Chapter 5 left behind the field of traditional inferential statistics and entered the world of predictive modeling. The research question answered with this chapter was *to what extent political context factors (including public opinion) contribute to the prediction of court decision-making*. This question is relevant to the study of the relationship between *court-public* because I test the direct effect of public opinion on judicial decision-making. I evaluated whether it is possible to correctly forecast the decision-making of the GFCC using a machine learning algorithm, and showed that it is possible to correctly predict, on average, three out of four proceeding outcomes correctly. What is more, I explicitly teased out the predictive contribution of legal and political context in the forecasting framework and find that the predictive performance of the algorithm is considerably improved when the political context of a proceeding is considered. This chapter therefore supports the view of a multifaceted decision-making of constitutional courts which is best characterized by the ensemble of both legal and political context. The main implication of this chapter is that the given predictability, and thus consistency of GFCC decision-making, is actually a good sign for the basic functioning of the rule of law in Germany, because it means that on average, similar proceedings with similar case characteristics are decided in a similar way.

6.2 Contributions, Implications and Avenues for Further Research

The findings of this dissertation were only possible due to several theoretical and methodological innovations. In the following, I wish to highlight the contributions and central implications of my study, as well as how they open new avenues for future research.

6.2.1 Contributions and Central Implications

In this section, I wish to briefly recap the main contributions of this dissertation and give a summary of its implications. I would also like to emphasize how they advance our understanding of judicial behavior and contribute to the field of political science in general.

Varying Levels of Public Support

My first theoretical contribution is that I relaxed the common assumption that constitutional courts possess about a consistently high level of diffuse support as it is assumed in many theories (e.g. Vanberg, 2001, 2005; Krehbiel, 2016, 2019). Instead, I argued that the degree of public support varies across countries, and showed that this has severe consequences for court's ability to effectively exercise constitutional review and their institutional legitimacy. While this argument is not entirely new to the literature (e.g. Staton and Vanberg, 2008), this is the first time that it has been tested in a comparative research design in several applications. Moreover, I showed that the varying degrees of public support not only affect the ability of constitutional courts to exercise constitutional review in the judicial-legislative relation, but also that it shapes the extent to which judicial decisions can move public opinion.

Courts Are Not Helpless When Facing Noncompliance

My second theoretical contribution is to demonstrate that courts are not entirely helpless when they expect legislative noncompliance. Besides other institutional tools, they can use the language of their decision to maximize their utility when strategically interacting with the legislature and the public. This is relevant in order to understand why judges behave the way in which they do. A primary implication of traditional separation of power models is that courts must strategically uphold governmental laws when they expect legislative noncompliance (Rogers, 2001; Vanberg, 2001, 2005; Staton, 2006). However, my findings show that they have yet another option available: they can strike down a governmental policy, but can write a decision where the judicial policy implementation vagueness is sufficient such that the government can uphold the status quo policy at little cost. This is an important observation, because standard

data sets using a binary coding practice would have coded the decision of the judges to reject a law as a loss of the government. Thus, the binary coding would lead to an under-estimation of the real extent of strategic judicial behavior.

A Novel Measure for Judicial Vagueness

A further contribution is made with respect developing of a novel measure for judicial vagueness. I argued that the current usage of dictionary-based methods to measure latent concepts in political texts is difficult and potentially misleading, because serious errors can occur when dictionaries from one domain are directly applied without any adaption to another (see Grimmer and Stewart, 2013). I overcome this problematic practice by providing two state-of-the-art measurement strategies to automatically detect vague language in judicial decisions. I rely on recent advances in computational linguistic to demonstrate *a)* how a general dictionary can be expanded to a specific domain using word embeddings and *b)* how to develop and benchmark state-of-the-art supervised machine learning classifiers. I also provide an exhaustive validation of both measures, and demonstrate that my measures outperform commonly-used dictionaries such as the Linguistic Inquiry and Word Count dictionary.

More importantly, at the conceptional level my work provides a blueprint and practical guidance for other researchers in social science who face the sample hurdles, i.e. data scarcity and the lack of prior research in a certain linguistic domain. I highlighted that the concept of the phenomenon desired to be measured must be carefully developed. I have demonstrated that social scientists and computational linguistics mean different things when they both speak about linguistic vagueness. With this in mind, I developed the notion of judicial policy implementation vagueness and defined it as a particular form of vagueness unique to the context of judicial decisions. Moreover, I illustrated that which of my methods should be applied is problem-specific. For researchers who can rely on an existing general dictionary, I show how word embeddings can be used to expand this dictionary to a specific domain. For researchers without such an existing dictionary, I demonstrate how a state-of-the-art machine learning classifier can be developed to solve a certain classification problem.

Researchers should keep in mind that there is always a trade-off along the two dimensions of accuracy/quality of a classification approach and the required resources. An expanded dictionary does not reach the same precision as a carefully-developed and tuned algorithm, although it is easy to implement and low in costs with respect to computational resources and required time. By contrast, a machine learning classifier often outperforms a dictionary in terms of accuracy, but requires extensive resources to gather the necessary training data and algorithmic fine-tuning. The contribution of my work thus extends beyond the methodological application on the judicial decisions, but teaches us some important lessons about the general feasibility and applicability of natural language processing approaches to political science problems.

The Value of Predictive Modeling for the Field of Social Science

In this dissertation, I also demonstrated the value of predictive modeling for the field of social science. Social scientists from various disciplines still prioritize inferential modeling over predictive modeling, with the common perception that not much can be learned from the process of predicting a certain outcome from an inferential perspective. In this dissertation, I challenge this claim and show with my application on the prediction of the decision-making of the GFCC that predictive modeling is an excellent way to compare the predictive power of competing theories and the contribution of certain variables for the prediction.

I make three contributions. First, predictive modeling allows me to test whether variables associated with the legal context of a decision are sufficient to predict court decision-making, or whether political context adds to the prediction. This, in turn, has implications for our understanding of the predictability of judicial decision-making. Second, I show that machine learning methods are helpful to discover non-linearities in the data that conventional regression models often do not detect. Third, my findings raise new research questions that require novel or refined theories, and allow me to create the first benchmark of predictive accuracy for decisions of the German Federal Constitutional Court. I hope to have convinced the reader that even if the ultimate goal is to make statements of causal nature, predictive modeling is beneficial to gather new insights from another perspective on a certain research problem. While I do not argue that machine learning will replace conventional statistical social science methods, one should be aware that algorithmic procedures will become an increasingly-used methodological tool in the toolbox of quantitative social scientists.

Beyond the US Supreme Court: A Comparative Perspective on Germany and France

Finally, in the course of my study, I extend beyond analyzing a single constitutional court. My theoretical arguments were constantly tested in a comparative setting using two European constitutional courts, namely the German and the French constitutional court. This design allows me to compare the importance of varying degrees of institutional support across countries without the need to rely on scarce survey data. Furthermore, my comparative results have higher external validity compared with studies that only focus on a single court. This is especially important with respect to the generalization of my findings of this study. Therefore, my work contributes to the still under-researched area of comparative research on European constitutional courts. Moreover, many constitutional courts established after the third wave of democratization are mainly modeled along the lines of the German constitutional court and the French Conseil Constitutionnel (Hönnige, 2011, 347). With a similar institutional design and equipped with the right of constitutional review, my conclusions also hold implications for these courts.

6.2.2 Implications and Avenues for Future Research

The findings of my dissertation open up new avenues for future research. In particular, I wish to highlight five aspects that I consider to hold important implications for a better understanding of the relationship between the court, government and the public, as well as how future work should proceed.

The Promises of Computer-assisted Text Analysis for Judicial Politics

Chapter 3 and 4 demonstrated that text contains a lot of useful information about the motives of courts and their strategic behavior. To arrive at this conclusion, I developed novel measures of judicial policy implementation vagueness. Although my vagueness measures outperform commonly-used measurement approaches such as LIWC, there remains still potential for improvement. Other researchers can use my annotated data set as a benchmark to test different algorithms and explore more powerful model architectures. In this regard, recent advances in semi-supervised text classification (the combination of adversarial neural networks to create more training data and supervised classification) have shown promising results, especially in situations where training data is scarce but the classification problem is challenging (e.g. Aghakhani et al., 2018; Chen and Cardie, 2018).

My application was only a brief glance at the rich source of information that is available in judicial texts. There are many judicial texts that remain unexplored but relevant for research on judicial behavior, especially with respect to European courts. Whereas the document bodies of opinions of the US Supreme Court have already been subject to scholarly attention (Owens and Wedeking, 2011, 2012; Cross and Pennebaker, 2014; Black et al., 2016*b*; Clark and Lauderdale, 2010; Lauderdale and Clark, 2014; Bonica and Sen, 2017; Wedeking and Zilis, 2018), only little work has been done on judicial texts of European courts (Dyevre, 2015). Building on these efforts, future studies could try to use the polarity of citations to either locate decisions in a doctrine space or combine information on the citations and other available information to create a common space for the court and political elites (e.g. Clark and Lauderdale, 2010; Lauderdale and Clark, 2014). One has to keep in mind that unfortunately much less information is available in the European court context compared with that of the Supreme Court, where scholars can combine various sources of information to develop elaborate measures of preferences; for instance, the individual writings of judges and their corresponding votes.

Moreover, vagueness is only one aspect of language. For instance, research from the US Supreme Court shows that Supreme Court justices strategically vary opinion clarity when a case's outcome contradicts popular sentiment. The basic argument in this regard is that courts that rule against the prevailing public opinion will try to explain the reasons of their ruling with clear and thus more "readable" language to persuade

the public of their ruling. Since German judges, for instance, are also affected by the specific support for their decisions (Sternberg et al., 2015), such a mechanism could also work with respect to the German court. It would be possible that German judges also try to use more persuasive language when issuing unpopular rulings, because they also have to find ways to foster their diffuse support.[1] In summary, analyzing other aspects of judicial language is a promising path for future research to gain a better understanding of why judges write decisions the way in which they do.

Improving the Prediction of GFCC Decision-making

In Chapter 5, I demonstrated that it is possible to correctly predict a substantial number of GFCC decision outcomes only using information that is available in advances. This holds at least three implications for further research. First, my prediction of the GFCC decisions can be used as a benchmark for future studies to incorporate a much richer set of predictors. Second, I will start to publish my forecasts on outcomes of the GFCC in real time on Twitter. This will constitute the reality check regarding whether an algorithmic procedure is indeed capable of predicting constitutional court outcomes. Third, it would be interesting to see how the algorithmic forecast performs against human legal experts; for instance, in the form of a prediction tournament where my algorithm competes against a crowd of law professors or even a large crowd of interested novices. Such a tournament was successfully carried out in the work of Ruger et al. (2004) and recently by Katz, Bommarito and Blackman (2017*a*). Such an interdisciplinary approach could help to shed light on the advantages and disadvantages of human- and computer-based predictive modeling, and could therefore be beneficial to evaluate the usefulness of crowd-sourced prediction in general. This is not only relevant for the prediction of court decisions but also for other forecasting problems, such as the prediction of electoral outcomes.

Advanced Formal Models: Incorporating the Notion of Active Courts and Varying Public Support

One obvious implication of this dissertation is that future game-theoretic models must integrate the central findings of this dissertation: not all courts are equally popular, and courts are more active than previously assumed. This has consequences for the judicial-legislative strategic interaction, which is central to many models that formalize the relationship between courts and governments in a separation of powers framework. Courts like the French Conseil Constitutionnel or the Russian constitutional court (Trochev, 2002, 2008) cannot rely on the public to threaten the legislator in case of

[1] A first preliminary analysis shows that there is indeed substantial variation in the textual readability of separate opinions of GFCC judges. This variation could be systematic because I, for instance, find that the higher the number of judges writing such an opinion together, the lower the textual readability of these texts (along the line of the notion that "too many cooks spoil the broth").

governmental resistance. However, they might have other tools at their disposal (like vague language) that help them to deal with this situation. These two aspects in turn affect the utility functions and pay-offs of the actors involved in these games.

One possible starting point for a formalization of these aspects could be the game-theoretic model of Krehbiel (2016), which already incorporates a belief about the probability of the public becoming aware of noncompliance and another belief about whether an oral hearing increases this probability. In an extension, the court could hold a third belief about the level of diffuse support that it enjoys, or a second model could be introduced for unpopular courts that explicitly relaxes the assumption that a court enjoys high levels of diffuse support. The fact that such a formalization is in principle possible and useful has already been demonstrated by the model of Staton and Vanberg (2008), which I empirically tested in Chapter 4.

Disentangling the Role of Constitutional Courts as Opinion Leaders

Chapter 2 showed that if equipped with sufficient institutional legitimacy, constitutional courts act as opinion leaders and affect public opinion on governmental policies. Nonetheless, more work must be done to disentangle the causal mechanisms of how public support and institutional legitimacy translate into the legitimacy-conferring capacity of (European) constitutional courts. For example, with respect to the external validity of my findings, follow-up work could use survey panels to measure public attitudes before and after landmark decisions to assess potential changes in public mood in a more realistic setting. Furthermore, future studies should take into account different salient and non-salient public policies, the role of the media as a mediator in terms of how the public learns about a decision or how individuals form their attitudes when they have access to competing arguments. Such work would help to better understand public opinion formation in the context of court decisions, and thus shed more light on the legitimacy-conferring capacity of courts.

The Value of Vagueness in Different Contexts

In Chapter 4, I examined the strategic use of vague language in the context of judicial decision-making. However, the central arguments of Staton and Vanberg's (2008) model and its implications also apply to other delegation relationships. In the context of the European Union, for instance, the European Commission must monitor the faithful compliance of member states with EU law. Here, similar dynamics as in the judicial-legislative bargaining between national high courts and national governments play a role in terms of whether the Commission starts an infringement procedure (Steunenberg, 2010; König and Mäder, 2014; Fjelstul and Carrubba, 2018). Moreover, akin to constitutional courts, other non-majoritarian institutions such as central banks are in a similar principal-agent relationship and must rely on other actors for the implementation of their instructions. Future research could therefore directly apply

my measurement approaches proposed in Chapter 3 to measure the amount of vague language in statements or other official records of the European Commission or central banks, and connect this data with information on the preferences of other actors to study noncompliance in the context of the European Union or central banks.

6.3 Concluding Thoughts

If courts cannot effectively exercise constitutional review, does this threat the simple model of checks and balances in liberal democracy? The power of constitutional courts to constrain legislative majorities is a central hallmark of liberal democracy and a key feature of the system of checks and balances of modern democracies. The findings of this dissertation indicate that the public plays a critical role for the efficacy of constitutional review: in the absence of diffuse support, courts must constrain themselves when legislative noncompliance is likely. However, if sufficient public support is given, they are in a much more powerful position and can strategically pressure the government to force compliance.

The decisive role of the public is also acknowledged by the courts. In Germany, we can currently observe that the German Federal Constitutional Court shows a tendency to intensify its work on public relations and its political communication in general. For instance, the President of the GFCC, Andreas Voßkuhle, has started to give regular talks about the "GFCC as a citizen's court".[2] As the first member of the GFCC ever, he also took part of the Federal Press Conference and publicly gave his legal assessment to political and societal questions. Voßkuhle also publicly criticized the rhetoric of the Christian Social Union with respect to their asylum policy as "unacceptable" in an interview. The German court has also started to translate important decisions into English to make them available to a wider audience. All of these developments can be interpreted as an effort to move the GFCC into the center of public attention and as a strategic signal to politicians that the court is not only aware of its power, but also willing to use it if necessary.

This strategy is not without risk. On the one hand, these activities increase the public awareness for the institutional work of the court and draws the attention of the public to certain rulings or upcoming decisions. Thus, it strengthens the position of the court in judicial-legislative conflicts. On the other hand, the legitimacy of courts and their diffuse support is largely based on their perception as an impartial and apolitical actor. It remains unclear whether the GFCC's strategy will be beneficial for the court. Nonetheless, the German court has recognized that public support is a crucial factor for its ability to effectively exercise constitutional review. It will be very interesting to see how the upcoming new President of the GFCC, Stephan Harbath, will interpret his

[2]German original: "Das Bundesverfassungsgericht als Bürgergericht".

role as the head of the court given that he is directly appointed from the Bundestag and has a pronounced partisan background.

The recent developments in Poland, Hungary as well as in the United States draw attention to the area of conflict between courts and governments and the role of the public. What is the appropriate strategy for courts to deal with legislative noncompliance? Should constitutional courts become more offensive and actively use their institutional tools to pressure the other branches of the government for compliance, if their level of public support allows them to do so? What does this imply for their institutional legitimacy if they slowly turn to yet another political actor? Finding answers to these questions is not only relevant from an academic perspective, but it is also crucial for the efficacy of constitutional review in a constitutional state, and thus the sustainability of liberal democracy.

Bibliography

Aghakhani, Hojjat, Aravind MacHiry, Shirin Nilizadeh, Christopher Kruegel and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*. pp. 89–95.

Alesina, Alberto and Alex Cukierman. 1990. "The politics of ambiguity." *The Quarterly Journal of Economics* 105(4):829–850.

Allaire, J.J. and Francois Chollet. 2017. *Deep Learning with R*. Manning Publications.

Baas, Larry R. and Dan Thomas. 1984. "The Supreme Court and Policy Legitimation. Epxerimental Tests." *American Politics Quarterly* 12(3):335–360.

Bailey, Michael A. and Forrest Maltzman. 2011. *The Constrained Court: Law, Politics, and the Decisions Justices Make*. Princeton University Press.

Banea, Carmen, Rada Mihalcea, Janyce Wiebe and Samer Hassan. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. pp. 127–135.

Bartels, Brandon L. and Christopher D. Johnston. 2013. "On the Ideological Foundations of Supreme Court Legitimacy in the American Public." *American Journal of Political Science* 57(1):184–199.

Bartels, Brandon L and Diana C Mutz. 2009. "Explaining processes of institutional opinion leadership." *Journal of Politics* 71(1):249–261.

Bartels, M. 2002. "Beyond the running tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Baum, Lawrence. 2009. *The puzzle of judicial behavior*. University of Michigan Press.

Bawn, Kathleen. 1995. "Political Control Versus Expertise: Congressional Choices about Administrative Procedures." *American Political Science Review* 89(01):62–73.

Beauchamp, Nicholas. 2017. "Predicting and Interpolating State-Level Polls Using Twitter Textual Data." *American Journal of Political Science* 61(2):490–503.

Bengio, Yoshua, Patrice Simard and Paolo Frasconi. 1994. "Learning Long-Term Dependencies with Gradient Descent is Difficult." *IEEE Transactions on Neural Networks* 5(2):157–166.

Bishop, Christopher M. 1995. "Training with noise is equivalent to Tikhonov regularization." *Neural Computation* 7(1):108–116.

Black, Ryan C., Ryan J. Owens, Justin Wedeking and Patrick C. Wohlfarth. 2016*a*. "The Influence of Public Sentiment on Supreme Court Opinion Clarity." *Law and Society Review* 50(3):703–732.

Black, Ryan C., Ryan J. Owens, Justin Wedeking and Patrick C Wohlfarth. 2016*b*. *U.S. Supreme Court Opinions and Their Audiences*. Cambridge University Press.

Blom, Annelies G., Christina Gathmann and Ulrich Krieger. 2015. "Setting Up an Online Panel Representative of the General Population: The German Internet Panel." *Field Methods* 27(4):391–408.

Böckenförde, Ernst-Wolfgang. 1976. "Die Methoden der Verfassungsinterpretation: Bestandsaufnahme und Kritik." *Neue Juristische Wochenschrift* 29(46):2089–2144.

Bonica, Adam. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science* 62(4):830–848.

Bonica, Adam and Maya Sen. 2017. "A Common-Space Scaling of the American Judiciary and Legal Profession." *Political Analysis* 25(1):114–121.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24(421):123–140.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Breiman, Leo, Jerome H. Friedman Friedman, Richard A. Olshen and Charles L. Stone Stone. 1984. *Classification and Regression Trees*. Belmont: CA: Wadsworth International Group.

Brennan, Thomas, Lee Epstein and Nancy Staudt. 2009. "Economic Trends and Judicial Outcomes: A Macrotheory of the Court." *Duke Law Journal* 58(7):1191–1230.

Brickman, Danette and David A M Peterson. 2006. "Public Opinion Reaction to Repeated Events: Citizen Response to Multiple Supreme Court." *Political Behavior* 28(1):87–112.

Brouard, Sylvain. 2008. The Constitutional Council: The Rising Regulator of French politics Despite Continued Politicization. In *Beyond Stereotypes, French Fifth Republic at Fifty*, ed. Sylvain Brouard, Andrew M. Appleton and Amy G. Mazur. Palgrave pp. 99–150.

Brouard, Sylvain. 2009. "The Politics of Constitutional Veto in France: Constitutional Council, Legislative Majority and Electoral Competition." *West European Politics* 32(2):384–403.

Brouard, Sylvain and Christoph Hönnige. 2017. "Constitutional courts as veto players: Lessons from the United States, France and Germany." *European Journal of Political Research* 56(3):529–552.

Caldeira, Gregory A. and James L. Gibson. 1992. "The Etiology of Public Support for the Supreme Court." *American Journal of Political Science* 36(3):635–664.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.

Carrubba, Clifford J. 2005. "Courts and Compliance in International Regulatory Regimes." *Journal of Politics* 67(3):669–689.

Carrubba, Clifford J. and Christopher Zorn. 2010. "Executive Discretion, Judicial Decision Making, and Separation of Powers in the United States." *The Journal of Politics* 72(03):812–824.

Carrubba, Clifford J., Matthew Gabel and Charles Hankla. 2008. "Judicial Behavior under Political Constraints: Evidence from the European Court of Justice." *American Political Science Review* 102(4):435–452.

Casillas, Christopher J., Peter K. Enns and Patrick C. Wohlfarth. 2011. "How Public Opinion Constrains the U.S. Supreme Court." *American Journal of Political Science* 55(1):74–88.

Cawley, Gavin C. and Nicola L. C. Talbot. 2010. "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *Journal of Machine Learning Research* 11:2079–2107.

Chen, Xilun and Claire Cardie. 2018. "Multinomial Adversarial Networks for Multi-Domain Text Classification." *arXiv preprint arXiv:1802.05694* .

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." *arXiv preprint arXiv:1406.1078* .

Chortareas, Georgios, David Stasavage and Gabriel Sterne. 2002. "Does it pay to be transparent ? International evidence from central bank forecasts." *Review-Federal Reserve Bank of Saint Louis* 84(4):99–118.

Christenson, Dino P. and David M. Glick. 2015. "Issue-specific opinion change: The Supreme Court and health care reform." *Public Opinion Quarterly* 79(4):881–905.

Christenson, Dino P. and David M. Glick. 2018. "Reassessing the Supreme Court: How Decisions and Negativity Bias Affect Legitimacy." *Political Research Quarterly* Online Fir.

Chung, Junyoung, Caglar Gulcehre, Kyunghyun Cho and Yoshua Bengio. 2014. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." *arXiv preprint arXiv:1412.3555* .

Clark, Tom S. and Benjamin E. Lauderdale. 2010. "Locating Supreme Court Opinions in Doctrine Space." *American Journal of Political Science* 54(4):871–890.

Clark, Tom S. and Jonathan P. Kastellec. 2015. "Source Cues and Public Support for the Supreme Court." *American Politics Research* 43(3):504–535.

Clawson, Rosalee A and Elizabeth R Kegler. 2001. "The Legitimacy-Conferring Authority of the U.S. Supreme Court. An Experimental Design." *American Politics Research* 29(6):566–591.

Cohen, Jacob. 1960. "A coefficient of agreement for nominal scales." *Educational and psychological measurement* 20(1):37–46.

Corley, Pamela C. and Justin Wedeking. 2014. "The (dis)advantage of certainty: The importance of certainty in language." *Law and Society Review* 48(1):35–62.

Cortes, Corinna, Vladimir Vapnik and Lorenza Saitta. 1995. "Support-vector networks." *Machine Leaming* 20(3):273–297.

Cranmer, Skyler J. and Bruce A. Desmarais. 2017. "What Can We Learn from Predictive Modeling?" *Political Analysis* 25(2):145–166.

Cross, Frank B and James W Pennebaker. 2014. "The language of the Roberts court." *Mich. St. L. Rev* 853:854–894.

Dahl, Robert A. 1957. "Decision-making in a democracy: The Supreme Court as a national policy-maker." *Journal of Public Law* 6:279–295.

Dauphin, Yann, Angela Fan, Michael Auli and David Grangier. 2017. "Language Modeling with Gated Convolutional Networks Yann." *Proceedings of the 34th International Conference on Machine Learning* 70:933–941.

De-Veaux, Amelia and Oliver Roeder. 2018. "Is The Supreme Court Facing A Legitimacy Crisis?" *FiveThirtyEight* 1th October.

de Vries, Erik, Martijn Schoonvelde and Gijs Schumacher. 2018. "No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications." *Political Analysis* 26(4):417–430.

Denny, Matthew James and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26:168–189.

Dietrich, Bryce J, Ryan D Enos and Maya Sen. 2018. "Emotional Arousal Predicts Voting on the U.S. Supreme Court." *Political Analysis* 27:237–243.

Dolderer, Winfried. 2018. "Im permanenten Spannungsfeld." *Das Parlament* 40(1th October).

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65(2):135–150.

Durr, Robert H., Andrew D. Martin and Christina Wolbrecht. 2000. "Ideological Divergence and public support for the Supreme Court." *American Journal of Political Science* 44(4):768–776.

Dusserre, Emmanuelle and Muntsa Padró. 2017. "Bigger does not mean better! We prefer specificity." *Iwcs 2017—12th international conference on computational semantics—short papers* .

Dyevre, Arthur. 2008. "Making sense of judicial lawmaking: A theory of theories of adjudication." *EUI Working Papers* (9):1–57.

Dyevre, Arthur. 2015. "The Promise and Pitfalls of Text-Scaling Techniques for the Analysis of Judicial Opinions." *Working Paper Available at SSRN 2626370* pp. 1–33.

Easton, David. 1965. *A systems analysis of political life*. New York: Wiley.

Efron, Bradley. 1979. "Bootstrap method: another look at the jackknife." *The Annals of Statistics* 7(1):1–26.

Efron, Bradley and Robert Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 1(1):54–77.

Efron, Bradley and Trevor Hastie. 2016. *Computer age statistical inference*. Vol. 5 Cambridge University Press.

Eichorst, Jason and Nick Lin. 2019. "Resist to Commit: Concrete Campaign Statements and the Need to Clarify a Partisan Reputation." *Journal of Politics* 81(1):15–32.

Engst, Benjamin G. 2018. The Two Faces of Judicial Power. The Dynamics of Judicial-Political Bargaining PhD thesis.

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

Epstein, Lee, Jack Knight and Andrew D. Martin. 2001. "The Supreme Court as a strategic national policymaker." *Emory Law Journal* 50:583–611.

Farkas, Richárd, Veronika Vincze, György Móra, János Csirik and György Szarvas. 2010. "The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text." *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task* pp. 1–12.

Federal Constitutional Court Act in the version of 11 August 1993 (Federal Law Gazette I p. 1473), last amended by Article 2 of the Act of 8 October 2017 (Federal Law Gazette I p. 3546). 2017.

Fjelstul, Joshua C. and Clifford J. Carrubba. 2018. "The Politics of International Oversight: Strategic Monitoring and Legal Compliance in the European Union." *American Political Science Review* 112(03):429–445.

Fleiss, Joseph L., Bruce Levin and Myunghee Cho Paik. 1981. "The measurement of interrater agreement." *Statistical methods for rates and proportions* 2(212-236):22–23.

Forschungsgruppe Wahlen, Mannheim. 2019. "Partial Cumulation of Politbarometers 1977-2017. GESIS Data Archive, Cologne. ZA2391 Data file Version 10.0.0.".

Fraser, Bruce. 2010. "Hedging in political discourse." *OKULSKA, U., CAP, P., Perspectives in Politics and Discourse, Capitolo* 8.

Gal, Yarin. 2016. "Uncertainty in Deep Learning." *PhD Thesis, University of Cambridge* .

Ganter, Viola and Michael Strube. 2009. "Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features." *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* pp. 173–176.

Gibson, James L. and Gregory A. Caldeira. 2009. "Confirmation Politics and The Legitimacy of the U.S. Supreme Court: Institutional Loyalty, Positivity Bias, and the Alito Nomination." *American Journal of Political Science* 53(1):139–155.

Gibson, James L., Gregory A. Caldeira and Lester Kenyatta Spence. 2003*a*. "Measuring Attitudes toward the United States Supreme Court." *American Journal of Political Science* 47(2):354–367.

Gibson, James L., Gregory A. Caldeira and Lester Kenyatta Spence. 2003*b*. "The Supreme Court and the US Presidential Election of 2000: Wounds, Self-Inflicted or Otherwise?" *British Journal of Political Science* 33(4):535–556.

Gibson, James L, Gregory A Caldeira and Vanessa A Baird. 1998. "On the Legitimacy of National High Courts." *The American Political Science Review* 92(2):343–358.

Gibson, James L. and Michael J. Nelson. 2014. "The Legitimacy of the US Supreme Court: Conventional Wisdoms and Recent Challenges Thereto." *Annual Review of Law and Social Science* 10(1):201–219.

Gibson, James L. and Michael J. Nelson. 2015. "Is the U.S. Supreme Court's Legitimacy Grounded in Performance Satisfaction and Ideology?" *American Journal of Political Science* 59(1):162–174.

Gibson, James L. and Michael Nelson. 2018. "De-Conferring Judicial Legitimacy." *Working Paper, available at SSRN 3096019* .

Giles, Micheal W, Bethany Blackstone and Richard L Vining. 2008. "The supreme court in american democracy: Unraveling the linkages between public opinion and judicial decision making." *Journal of Politics* 70(2):293–306.

Ginsburg, Tom and Mila Versteeg. 2014. "Why do countries adopt constitutional review?" *Journal of Law, Economics, and Organization* 30(3):587–622.

Giuseppina Scotto di Carlo. 2013. *Vagueness as a political strategy: Weasel words in security council resolutions relating to the second gulf war*. Cambridge Scholars Publishing.

Green, Donald P. and Holger L. Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees." *Public Opinion Quarterly* 76(3):491–511.

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

Gruber, Helmut. 1993. "Political language and textual vagueness." *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* 3(1):1–28.

Guimera, Roger and Marta Sales-Pardo. 2011. "Justice blocks and predictability of us supreme court votes." *PLoS ONE* 6(11):e27188.

Hall, Matthew E. K. 2014. "The Semiconstrained Court: Public Opinion, the Separation of Powers, and the U.S. Supreme Court's Fear of Nonimplementation." *American Journal of Political Science* 58(2):352–366.

Hamilton, Alexander, James Madison and John Jay. 1996. Federalist Paper No 78. In *The Federalist Papers*, ed. Benjamin Wright. New York: Barnes and Noble Books.

Hanley, John. 2008. The death penalty. In *Public Opinion and Constitutional Controversy*, ed. Nathaniel Persily, Jack Citrin and Patrick J. Egan. New York: Oxford University Press pp. 108–138.

Hanmer, Michael J. and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Second ed. New York: Springer series in statistics.

Hlavac, Marek. 2018. "stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer.".

Hochreiter, Sepp. 1998. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06(02):107–116.

Hoekstra, Valerie J. 1995. "An Experimental Study of the Court ' s Ability to Change Opinion." *American Politics Quarterly* 23(1):109–129.

Hoekstra, Valerie J. 2000. "The Supreme Court and Local Public Opinion." *American Political Science Review* 94(1):89–100.

Hoekstra, Valerie J and Jeffrey A Segal. 1996. "The Shepherding of Local Public Opinion : The Supreme Court and Lamb's Chapel." *The Journal of Politics* 58(4):1079–1102.

Holmes, Oliver Wendell. 1897. "The Path of Law." *10 Harvard Law Review* 457.

Hönnige, Christoph. 2007. *Verfassungsgericht, Regierung und Opposition: Die vergleichende Analyse eines Spannungsdreiecks*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Hönnige, Christoph. 2009. "The Electoral Connection: How the Pivotal Judge Affects Oppositional Success at European Constitutional Courts." *West European Politics* 32(5):963–984.

Hönnige, Christoph. 2011. "Beyond Judicialization: Why We Need More Comparative Research About Constitutional Courts." *European Political Science* 10(3):346–358.

Hönnige, Christoph, Thomas Gschwend, Caroline Wittig and Benjamin Engst. 2015. "Constitutional Court Database (CCDB), V17.01 [Mar.].".

Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical Statistics* 15(3):651–674.

Hyland, Ken. 1998. *Hedging in scientific research articles*. Vol. 54 John Benjamins Publishing.

Jastrzebski, Stanisław, Damian Leśniak and Wojciech Marian Czarnecki. 2017. "How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks." *arXiv preprint arXiv:1702.02170* .

Johnson, Timothy R and Andrew D Martin. 1998. "The Public's Conditional Response to Supreme Court Decisions." *The American Political Science Review* 92(2):299–309.

Jones, Zachary M. and Yonatan Lupu. 2018. "Is There More Violence in the Middle?" *American Journal of Political Science* 62(3):652–667.

Jónsson, Ólafur Páll. 2009. "Vagueness, interpretation, and the law." *Legal Theory* 15(3):193–214.

Jozefowicz, Rafal, Wojciech Zaremba and Ilya Sutskever. 2015. "An Empirical Exploration of Recurrent Network Architectures Rafal." *Proceedings of the 32nd International Conference on Machine Learning* 37:2342–2350.

Kastellec, Jonathan P. 2010. "The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees." *Journal of Empirical Legal Studies* 7(2):202–230.

Katz, Daniel Martin. 2013. "Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry." *Emory L. J.* 62(July 2011):909–966.

Katz, Daniel Martin, Michael James Bommarito and Josh Blackman. 2017*a*. "Crowdsourcing Accurately and Robustly Predicts Supreme Court Decisions." *Available at SSRN 3085710* pp. 1–11.

Katz, Martin Daniel, Michael J. Bommarito and Josh Blackman. 2017*b*. "A general approach for predicting the behavior of the Supreme Court of the United States." *PLoS ONE* 12(4):e0174698.

Kaufman, Aaron Russell, Peter Kraft and Maya Sen. 2019. "Improving Supreme Court Forecasting Using Boosted Decision Trees." *Political Analysis* Online fir:1–7.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." *arXiv preprint arXiv:1408.5882* .

King, Gary, James Honaker, Anne Joseph, Kenneth Scheve and Joe Schafer. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *The American Political Science Review* 95(1):49–69.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347.

König, Thomas and Lars Mäder. 2014. "The Strategic Nature of Compliance: An Empirical Evaluation of Law Implementation in the Central Monitoring System of the European Union." *American Journal of Political Science* 58(1):246–263.

König, Thomas, Moritz Marbach and Moritz Osnabrügge. 2013. "Estimating Party Positions across Countries and Time–A Dynamic Latent Variable Model for Manifesto Data." *Political Analysis* 21(4):468–491.

Kranenpohl, Uwe. 2010. *Hinter dem Schleier des Beratungsgeheimnisses. Der Willensbildungs- und Entscheidungsprozess des Bundesverfassungsgerichts*. 1 ed. Wiesbaden: VS Verlag für Sozialwissenschaften.

Krehbiel, Jay N. 2016. "The Politics of Judicial Procedures: The Role of Public Oral Hearings in the German Constitutional Court." *American Journal of Political Science* 60(4):990–1005.

Krehbiel, Jay N. 2019. "Elections, Public Awareness and the Efficacy of Constitutional Review." *Journal of Law and Courts* 7(1):53–79.

Kuhn, Max. 2008. "Building Predictive Models in R Using the caret Package." *Journal Of Statistical Software* 28(5):1–26.

Lai, Siwei, Kang Liu, Liheng Xu and Jun Zhao. 2016. "How to Generate a Good Word Embedding?" *IEEE Intelligent Systems* 31(6):5–14.

Lakoff, George. 1973. "Hedges: A Study In Meaning Criteria And The Logic Of Fuzzy Concepts." *Journal of Philosophical Logic* 2(4):458–508.

Landis, J. Richard and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1):159.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling Meaningful Political Dimensions." *American Journal of Political Science* 58(3):754–771.

Laver, Michael and Ian Budge. 1992. Measuring Policy Distances and Modelling Coalition Formation. In *Party policy and government coalitions*. New York: St. Martin's Press pp. 15–40.

Lecun, Yann, Yoshua Bengio and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521(7553):436–444.

Lee, Ji Young and Franck Dernoncourt. 2016. "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks." *arXiv preprint arXiv:1603.03827* .

Light, Marc, Xin Ying Qiu and Padmini Srinivasan. 2004. "The Language of Bioscience : Facts , Speculations , and Statements in Between." *BioLink 2004 – Proceedings of the Workshop on Linking Biological Literature, Ontologies and Databases* pp. 17–24.

Linos, Katerina and Kimberly Twist. 2016. "The Supreme Court, the Media, and Public Opinion: Comparing Experimental and Observational Methods." *The Journal of Legal Studies* 45(2):223–254.

Liu, Fei, Nicole Lee Fella and Kexin Liao. 2016. "Modeling language vagueness in privacy policies using deep neural networks." *AAAI Fall Symposium Series* pp. 257–263.

Lodge, Milton and Ruth Hamill. 1986. "A Partisan Schema for Political Information Processing." *American Political Science Review* 80(02):505–519.

Lopez, Marc Moreno and Jugal Kalita. 2017. "Deep Learning applied to NLP." *arXiv preprint arXiv:1703.03091* .

Loughran, Tim and Bill Mcdonald. 2011. "When is a Liability not a Liability? Textual Analysis, Distionaries, and 10-Ks." *Journal of Finance* 66(1):35–65.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36(1):123–155.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Marschall, Melissa J. and Robert J. McKee. 2002. "From campaign promises to presidential policy: Education reform in the 2000 election." *Educational Policy* 16(1):96–117.

Marshall, Thomas R. 1987. "The Supreme Court as an Opinion Leader. Decisions and the Mass Public." *American Politics Quarterly* 15(1):147–168.

Martin, Andrew D., Kevin M. Quinn, Theodore W. Ruger and Pauline T. Kim. 2004. "Competing Approaches to Predicting Supreme Court Decision Making." *Symposium: Forecasting U.S. Supreme Court Decisions* 2(4):761–767.

McGuire, Kevin T. and James A. Stimson. 2008. "The Least Dangerous Branch Revisited: New Evidence on Supreme Court Responsiveness to Public Preferences." *The Journal of Politics* 66(4):1018–1035.

Medlock, Ben and Ted Briscoe. 2007. "Weakly supervised learning for hedge classification in scientific literature." *Annual Meeting of Assosiation of Computational Linguistics* 45(June):992.

Medvedeva, Masha and Vols, Michel and Wieling, Martijn. 2018. "Judicial decisions of the European Court of Human Rights: looking into the crystall ball." *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*. pp. 1–24.

Meirowitz, Adam. 2005. "Informational party primaries and strategic ambiguity." *Journal of Theoretical Politics* 17(1):107–136.

Mikolov, Tomas, Wen-Tau Yih and Geoffrey Zweig. 2013. "Linguistic regularities in continuous space word representations." *Proceedings of NAACL-HLT* (June):746–751.

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Neunhoeffer, Marcel and Sebastian Sternberg. 2019. "How cross-validation can go wrong and what to do about it." *Political Analysis* 27(1):101–106.

Newton, Kenneth. and Jan W. Van Deth. 2012. *Foundations of Comparative Politics*. Cambridge University Press.

O'Rourke, Anthony. 2017. "Semantic Vagueness and Extrajudicial Constitutional Decisionmaking." *William & Mary Bill of Rights Journal* 25(4):1301.

Ossenbühl, Fritz. 1998. Verfassungsgerichtsbarkeit und Gesetzgebung. In *Verfassungsgerichtsbarkeit und Gesetzgebung. Symposion aus Anlass des 70. Geburtstages von Peter Lerche*, ed. Peter Badura, Peter Scholz and Rupert Scholz. München: Beck pp. 75–99.

Owens, Ryan J. and Justin P. Wedeking. 2011. "Justices and legal clarity: Analyzing the complexity of U.S. Supreme Court opinions." *Law and Society Review* 45(4):1027–1061.

Owens, Ryan J. and Justin Wedeking. 2012. "Predicting drift on politically insulated institutions: A study of ideological drift on the United States Supreme Court." *Journal of Politics* 74(2):487–500.

Owens, Ryan J., Justin Wedeking and Patrick C. Wohlfarth. 2013. "How the Supreme Court Alters Opinion Language to Evade Congressional Review." *Journal of Law and Courts* 1(1):35–59.

Papke, Leslie E. and Jeffrey M. Wooldridge. 1996. "Econometric methods for fractional response variables with an application to 401(k) plan participation rates." *Journal of Applied Econometrics* 11(6):619–632.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, A M. Perrot and E. Duchesnay. 2011. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, James W. and Laura A. King. 1999. "Linguistic styles: Language use as an individual difference." *Journal of personality and social psychology* 77(6):1296–1312.

Pennebaker, James W., Ryan L. Boyed, Kayla Jordan and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. Technical report University of Texas.

Pennington, Jeffrey, Richard Socher and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543.

Piolat, A., R. J. Booth, C. K. Chung, M. Davids and James W. Pennebaker. 2011. "La version franaise du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation." *Psychologie Francaise* 56(3):145–159.

Porter, Martin F. 1980. "An algorithm for suffix stripping." *Program* 14(3):130–137.

Poscher, Ralf. 2012. Ambiguity and vagueness in legal interpretation. In *The Oxford Handbook of Language and Law*, ed. Peter Solan, Lawrence Tiersma. Oxford Unitversity Press.

Post, Robert C. 1994. "Reconceptualizing Vagueness: Legal Rules and Social Orders." *California Law Review* 82(3):491.

Przeworski, Adam and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.

Rabab, Ghaleb and Ronza Abu Rumman. 2015. "Hedging in Political Discourse : Evidence from the Speeches of King Abdullah II of Jordan." *Prague Journal of English Studies* 4(1):157–185.

Rahn, Wendy M. 1995. "The Role of Partisan Stereotypes in Information Processing about Political Candidates." *American Political Science Review* 37(2):472–496.

Reidenberg, Joel R, Travis D Breaux and Thomas B Norton. 2016. "Automated Comparisons of Ambiguity in Privacy Policies and the Impact of Regulation." *Journal of Legal Studies* 1(9):1–29.

Rheault, Ludovic and Christopher Cochrane. 2018. Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora. In *35th annual meeting of the Society for Political Methodology*.

Rogers, James R. 2001. "Information and judicial review: A signaling game of legislative-judicial interaction." *American Journal of Political Science* 45(1):84–99.

Rong, Xin. 2014. "word2vec Parameter Learning Explained." *arXiv preprint arXiv:1411.2738* pp. 1–21.

Rosenthal, Sara, Noura Farra and Preslav Nakov. 2017. "SemEval-2017 task 4: Sentiment analysis in Twitter." *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* pp. 502–518.

Rudzio, Wolfgang. 2006. *Das politische System der Bundesrepublik Deutschland*. Vol. 7 7. ed. Wiesbaden: VS Verlag für Sozialwissenschaften.

Ruger, Theodore W., Pauline T. Kim, Andrew D. Martin and Kevin M. Quinn. 2004. "The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking." *Columbia Law Review* 104(4):1150–1210.

Russel, Stuart J. and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited.

Santora, Marc. 2018. "Polish Crisis Deepens as Judges Condemn Their Own Court." *New York Times* 5th July.

Schaal, Gary S. 2007. Crisis! What Crisis? Der „Kruzifix-Beschluss" und seine Folgen. In *Das Bundesverfassungsgericht im politischen System*, ed. Robert Chr. van Ooyen and Martin H. W. Möllers. Wiesbaden: VS Verlag für Sozialwissenschaften pp. 175–186.

Schnabel, Tobias, Igor Labutov, David Mimno and Thorsten Joachims. 2015. "Evaluation methods for unsupervised word embeddings." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (September):298–307.

Schneider, Hans-Peter. 1974. *Die parlamentarische Opposition im Verfassungsrecht der Bundesrepublik Deutschland*. Vittorio Klostermann.

Segal, Jeffrey A. and Albert D. Cover. 1989. "Ideological Values and the Votes of U.S. Supreme Court Justices." *American Political Science Review* 83(2):557–565.

Segal, Jeffrey A and Harold J Spaeth. 2002. *The Supreme Court and the attitudinal model revisited*. Cambridge: Cambridge University Press.

Segal, Jeffrey A., Lee Epstein, Charles M. Cameron and Harold J. Spaeth. 1995. "Ideological Values and the Votes of US Supreme-Court-Justices Revisited." *Journal of Politics* 57(3):812–823.

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations." *Political Research Quarterly* 70(2):374–393.

Setiawan, Erwin B., Dwi H. Widyantoro and Kridanto Surendro. 2017. "Feature expansion using word embedding for tweet topic classification." *10th International Conference on Telecommunication Systems Services and Applications (TSSA). IEEE* (2011).

Shmueli, Galit. 2010. "To Explain or To Predict?" *Statistical science* 25(3):289–310.

Sieberer, Ulrich. 2006. "Strategische Zurückhaltung von Verfassungsgerichten. Gewaltenteilungsvorstellungen und die Grenzen der Justizialisierung." *Zeitschrift für Politikwissenschaft* 16(4):1299–1323.

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

Štajner, Sanja, Goran Glavas, Simone Paolo Ponzetto and Heiner Stuckenschmidt. 2017. Domain Adaptation for Automatic Detection of Speculative Sentences. In *Proceedings - IEEE 11th International Conference on Semantic Computing, ICSC 2017*. pp. 164–171.

Štajner, Sanja, Nicole Baerg, Simone Paolo Ponzetto and Heiner Stuckenschmidt. 2016. "Automatic detection of speculation in policy statements." *Workshops on Natural Language Processing and Computational Social Science. Association for Computing Machinery* .

Staton, Jeffrey K. 2006. "Constitutional Review and the Selective Promotion of Case Results." *American Journal of Political Science* 50(1):98–112.

Staton, Jeffrey K. 2010. *Judicial power and strategic communication in Mexico*. Cambridge University Press.

Staton, Jeffrey K. and Georg Vanberg. 2008. "The value of vagueness: delegation, defiance, and judicial opinions." *American Journal of Political Science* 52(3):504–519.

Staudt, Nancy and Yilei He. 2010. "The Macroeconomic Court: Rhetoric and Implications of New Deal Decision-Making." *Northwestern Journal of Law and Social Policy* 5(5).

Sternberg, Sebastian, Thomas Gschwend, Caroline E. Wittig and Benjamin G. Engst. 2015. "Zum Einfluss der öffentlichen Meinung auf Entscheidungen des Bundesverfassungsgerichts: Eine Analyse von abstrakten Normenkontrollen sowie Bund-Länder-Streitigkeiten 1974 - 2010." *Politische Vierteljahresschrift* 56(4):570–598.

Steunenberg, Bernard. 2010. "Is big brother watching? Commission oversight of the national implementation of EU directives." *European Union Politics* 11(3):359–380.

Stone, Alec. 1992. "Where Judicial Politics Are Legislative Politics: The French Constitutional Council." *West European Politics* 15(3):29–49.

Stoutenborough, James W., Donald P. Haider-Markel and Mahalley D. Allen. 2006. "Reassessing the Impact of Supreme Court Decisions on Public Opinion : Gay Civil Rights Cases." *Society* 59(3):419–433.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 2007. "Bias in random forest variable importance measures: illustrations, sources and a solution." *BMC Bioinformatics* 8(1):25.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. "Conditional variable importance for random forests." *BMC bioinformatics* 9(23):307.

Strobl, Carolin, Torsten Hothorn and Achim Zeileis. 2009. "Party on!" *R Journal* 1(2):14–17.

Sulea, Octavia Maria, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu and Josef Van Genabith. 2017. "Exploring the use of text classification in the legal domain." *arXiv preprint arXiv:1710.09306* .

Surden, Harry. 2014. "Machine Learning and Law." *Washington Law Review* 89.

Szarvas, György. 2008. "Hedge classification in biomedical texts with a weakly supervised selection of keywords." *Proceedings of ACL-08: HLT* pp. 281–289.

Tang, Duyu, Bing Qin and Ting Liu. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 1422–1432.

Tankard, Margaret E. and Elizabeth Levy Paluck. 2017. "The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes." *Psychological Science* 28(9):1334–1344.

Tausczik, Yla R. and James W. Pennebaker. 2010. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of Language and Social Psychology* 29(1):24–54.

Theil, Christoph Kilian, Sanja Štajner and Heiner Stuckenschmidt. 2018*a*. "Exploiting Word Embeddings for Industry-Specific Dictionary Expansions." *Working Paper* .

Theil, Christoph Kilian, Sanja Štajner and Heiner Stuckenschmidt. 2018*b*. Word Embeddings-Based Uncertainty Detection in Financial Disclosures. In *Economics and Natural Language Processing - proceedings of the First workshop (ECONLP 2018)*. pp. 32–37.

Theil, Christoph Kilian, Sanja Štajner, Heiner Stuckenschmidt and Simone Paolo Ponzetto. 2018. Automatic Detection of Uncertain Statements in the Financial Domain. In *Computational Linguistics and Intelligent Text Processing : 18th International Conference, CICLing 2017*. pp. 17–23.

Trochev, Alexei. 2002. "Implementing Russian Constitutional Court Desicions." *East European Constitutional Review* 11(12):1–11.

Trochev, Alexei. 2008. *Judging Russia: the role of the constitutional court in Russian politics 1990 - 2006*. Cambridge University Press.

Tsai, Ming-Feng and Chuan-Ju Wang. 2014. "Financial Keyword Expansion via Continuous Word Vector Representations." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1453–1458.

Tsebelis, George. 2002. *Veto Player: How Political Institutions Work*. Princeton University Press: Russell Sage Foundation.

van der Maaten, Laurens and Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." *Journal of Machine Learning Research* 1:1–48.

Vanberg, Georg. 2001. "Legislative-Judicial Relations: A Game-Theoretic Approach to Constitutional Review." *American Journal of Political Science* 45(2):346–361.

Vanberg, Georg. 2005. *The Politics of Constitutional Review in Germany*. Cambridge: Cambridge University Press.

Venables, William N. and Brian D. Ripley. 2011. oprobit: Ordinal Probit Regression for Ordered Categorical Dependent Variables. In *Zelig: Everyone's Statistical Software*, ed. Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King and Olivia Lau.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. In *BMC bioinformatics*. Vol. 9 pp. 1–9.

Volcansek, Mary L. 1991. Judicial Activism in Italy. In *Judicial activism in comparative perspective*, ed. Kenneth Holland. New York: St. Martin's Press pp. 117—-132.

Vorländer, Hans. 2006. Deutungsmacht – Die Macht der Verfassungsgerichtsbarkeit. In *Die Deutungsmacht der Verfassungsgerichtsbarkeit*, ed. Hans Vorländer. 1 ed. Vol. 6 Wiesbaden: VS Verlag für Sozialwissenschaften pp. 9–33.

Vorländer, Hans and André Brodocz. 2006. Das Vertrauen in das Bundesverfassungsgericht. Ergebnisse einer repräsentativen Bevölkerungsumfrage. In *Die Deutungsmacht der Verfassungsgerichtsbarkeit*, ed. Hans Vorländer. 1 ed. Wiesbaden: VS Verlag für Sozialwissenschaften pp. 259–295.

Vorländer, Hans and Gary S. Schaal. 2002. Integration durch Institutionenvertrauen? Das Bundesverfassungsgericht und die Akzeptanz seiner Rechtsprechung. In *Integration durch Verfassung*, ed. Hans Vorländer. 1 ed. Wiesbaden: VS Verlag für Sozialwissenschaften pp. 343–374.

Waldron, Jeremy. 1994. "Vagueness in Law and Language: Some Philosophical Issues." *California Law Review* 82(3):509.

Wedeking, Justin and Michael A. Zilis. 2018. "Disagreeable Rhetoric and the Prospect of Public Opposition: Opinion Moderation on the U.S. Supreme Court." *Political Research Quarterly* 71(2):380–394.

Wefing, Heinrich. 2016. "Der andere Präsident." *Zeit Magazin* 12.

Whissell, Cynthia, Michael Fournier, Rene Pelland, Deborah Weir and K. Makarec. 1986. "A dictionary of affect in language: IV. Reliability, validity, and applications." *Perceptual and Motor Skills* 62(3):875–888.

Wittig, Caroline E. 2016. *The Occurrence of Separate Opinions at the Federal Constitutional Court. An Analysis with a Novel Database*. Berlin: Logos Verlag Berlin GmbH.

Wolf, Markus, Andrea B. Horn, Matthias R. Mehl, Severin Haug, James W. Pennebaker and Hans Kordy. 2008. "Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count." *Diagnostica* 54(2):85–98.

Wright, Marvin N and Andreas Ziegler. 2015. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77(i01).

Wu, Zhizheng, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng and Haizhou Li. 2015. "Exemplar-based unit selection for voice conversion utilizing temporal Information." *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* pp. 3057–3061.

Yin, Wenpeng, Katharina Kann, Mo Yu and Hinrich Schütze. 2017. "Comparative Study of CNN and RNN for Natural Language Processing." *arXiv preprint arXiv:1702.01923* .

Young, Lori and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29(2):205–231.

Young, Tom, Devamanyu Hazarika, Soujanya Poria and Erik Cambria. 2018. "Recent trends in deep learning based natural language processing." *ieee Computational intelligence magazine* 13(3):55–75.

Zaller, John R. 1992. *The nature and origins of mass opinion*. Cambridge University Press.

Zhang, Ye and Byron Wallace. 2015. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." *arXiv preprint arXiv:1510.0382* .

Zilis, Michael. 2014. "Investigating the Effects of Judicial Legitimacy on Micro-Level Opinion." *APSA 2014 Annual Meeting Paper* .

APPENDIX A

Chapter 2

A.1 Question Wording of Institutional Trust Questions GIP and ENEF

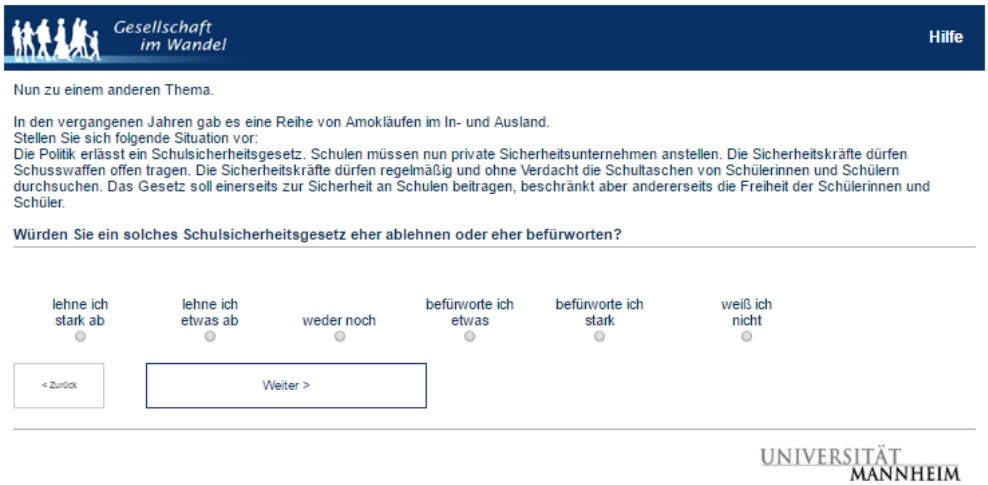
- GIP (asked in Wave 26), exact wording (in German): Geben Sie bitte bei jeder Einrichtung oder Organisation an, wie groß das Vertrauen ist, das Sie ihr entgegenbringen. Benutzen Sie dazu bitte diese Skala: 1 bedeutet, dass Sie ihr überhaupt kein Vertrauen entgegenbringen; 7 bedeutet, dass Sie ihr sehr groß es Vertrauen entgegenbringen. Mit den Zahlen dazwischen können Sie Ihre Meinung wiederum abstufen. Wie ist das mit dem Bundesverfassungsgericht?
- ENEF (asked in Wave 17), exact wording (in French): Sur une échelle de 0 à 10, avez-vous confiance ou pas dans chacune des institutions suivantes: Le Conseil Constitutionnel?

In the ENEF survey, the trust in the CC was originally asked on a 10-point scale. For a more convenient visualization and to enable a comparative analysis, the values were recoded the same 7-point scale as used in the GIP. Findings are robust to using the original 10-point scale instead of the recoded 7-point scale.

A.2 Original Screenshots of Survey Experiments in the GIP

Figure A.1 shows an original screen-shot of the survey experiment implemented in Wave 26 and Wave 27 of the German Internet Panel.

Figure A.1 – Actual Screenshot (in German) of the Survey Experiment Implemented in German Internet Panel



The wording of the French experiment is as follows:

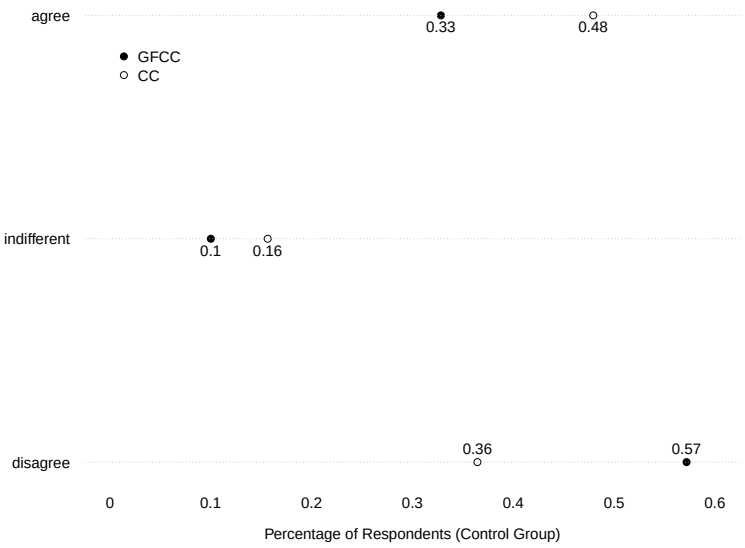
D'après le Ministère de l'Education nationale, la violence scolaire en France est en constante augmentation ces dernières années. Imaginez le scénario suivant : le Parlement adopte une nouvelle loi sur la sécurité dans les écoles. Les écoles doivent maintenant garantir un accès permanent de la police dans les établissements scolaires. Les policiers sont autorisés à porter des armes à feu et à contrôler régulièrement les cartables des élèves sans suspicion particulière. Cette loi a pour objectif une augmentation de la sécurité dans les écoles en restreignant la liberté des élèves.

After this text, the treatment was added in form of: *Le [CC, Haut Conseil de l'Éducation] émet un avis [positif, négatif] sur la loi.* More information on the French implementation are available at `https://www.enef.fr/données-et-résultats/`.

A.3 Distribution of Attitudes towards School Security Law across Germany and France

Figure A.2 shows the distribution of attitudes towards the school security law across the German and French respondents. Only respondents of the control group are used, so that the experimental treatment does not affect the respondents' answers. Whereas in Germany the majority of the respondents (57%) disagree with the school security

Figure A.2 – Distribution of Attitudes towards School Security Law across Germany and France



Note: Comparison of the attitudes of German and French respondents towards the school security law. Only respondents in the control group are used for the evaluation.

law and only 33% agree with it, the majority of the French respondents supports such a law (48%) and only 36% oppose it.

A.4 Estimation Strategy for Ordered Probit Models

Let Y_i be the ordered categorical dependent variable for observation i that takes one of the integer values from 1 to J where J is the total number of categories. In our case, $J = 3$ as respondents rate the school security law in three ordered categories: *disagree*, *indifferent* and *agree*.

- The stochastic component is given by an latent continuous variable Y_i^*, defined by a normal distribution with mean μ_i and unit variance:

$$Y_i^* \sim N(\mu_i, 1).$$

We cannot observe Y_i^*, but instead we can only observe the categories of the response given by the some cut-points τ_j:

$$Y_i = \begin{cases} 1 = agree & \text{if } (\tau_0 = -\infty) \leq y_i^* < \tau_1 \\ 2 = indifferent & \text{if } \tau_1 \leq y_i^* < \tau_2 \\ 3 = disagree & \text{if } \tau_2 \leq y_i^* < (\tau_3 = +\infty) \end{cases}$$

In simple words, this means that we observe a particular ordinal response depending upon what part of the latent distribution a respondent is in. The next "higher" ordered response is only observed when a respondent crosses that particular cut-point. More formally, the probability to observe each category is given by:

$$\Pr(Y_i = j) = \Phi(\tau_j \mid \mu_i) - \Phi(\tau_{j-1} \mid \mu_i) \quad \text{for} \quad j = 1, \ldots, J$$

where $\Phi(\mu_i)$ is the cdf for the normal distribution with mean μ_i and unit variance.

- The systematic component is given by:

$$\mu_i = x_i \beta$$

where x_i is the vector of independent variables and β is the vector of coefficients. In the baseline model, the independent variables are dummy variables indicating whether a respondent belongs to either of the treatment groups (the dummy for the control group is omitted as the reference category).

A.5 Ordered Probit Regression Tables of Baseline Analysis in Germany and France

Table A.1 – Results of Ordinal Probit Regression, GIP Survey Germany

	DV: 3-Scale Rating of Likeability School Security Law
Data Security Official Approves	0.079
	(0.064)
Data Security Official Disapproves	−0.095
	(0.066)
GFCC Approves	0.126**
	(0.054)
GFCC Disapproves	−0.140***
	(0.053)
Conference of the Ministers of Education Approves	−0.076
	(0.065)
Conference of the Ministers of Education Disapproves	−0.178***
	(0.066)
disagree | indifferent	0.195***
	(0.040)
indifferent | agree	0.426***
	(0.040)
N	5208

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment). Standard errors are clustered by respondent because of panel design. All regression tables in this project are created with the *stargazer* package (Hlavac, 2018).

Table A.2 – Results of Ordinal Probit Regression, ENEF Survey France

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	0.093
	(0.074)
CC Disapproves	−0.118
	(0.074)
Haut Conseil de l'Éducation Approves	0.179**
	(0.074)
Haut Conseil de l'Éducation Disapproves	−0.093
	(0.074)
disagree | indifferent	−0.329***
	(0.053)
indifferent | agree	0.037
	(0.053)
N	2566

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment).

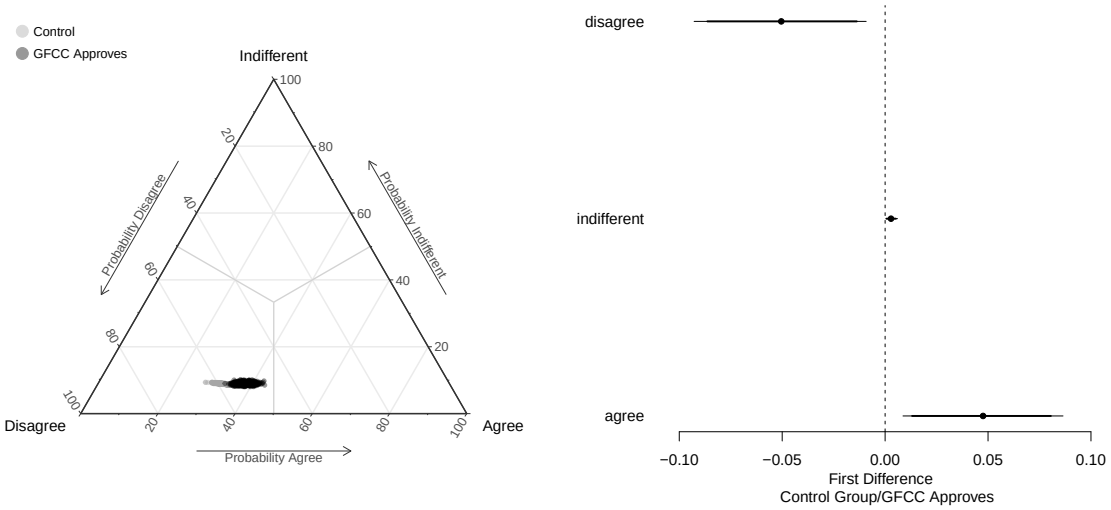
A.6 Ternary Plots of Simulation Results

In this section I provide an alternative visualization of the simulation results of the main analysis using "ternary plots" (King, Tomz and Wittenberg, 2000, 358). Ternary plots are triangular plots, and each of the three sides represents the probability of one of the three-categorical outcome. The tick marks at each side allows reading the probability of a simulated outcome being either indifferent, or agreeing or disagreeing with the schools security law. For instance, the right side of the ternary plot provides the probability of disagreeing for a respondent. The probability of that outcome would be at 100 percent in the left vertex of the triangle. The closer that the predicted probabilities are to one of the outcomes at the vertex, the higher the chance that a respondents favors the respective choice. If a simulated respondent's choice would be located at the centroid of the ternary plot, the probabilities would be equal for all three outcomes. I decided to use the parallel coordinate plots in the main text because they exhibit a better data-ink ration than the ternary plots. This is because the effect magnitudes of the experimental manipulations are rather small, and therefore two-third of the ternary plots are "empty".

The ternary plot on the left side of Figure A.3 shows 1,000 simulated expected values for a respondent in the control group (gray dots) and a respondent who received the endorsement that the GFCC approves the school security law (black dots). For the simulations, only the respective dummy (control group, GFCC approves) was set to one and all other dummies of the model were set to zero. The expected values are predicted probabilities across the three outcome categories (which sum to one).

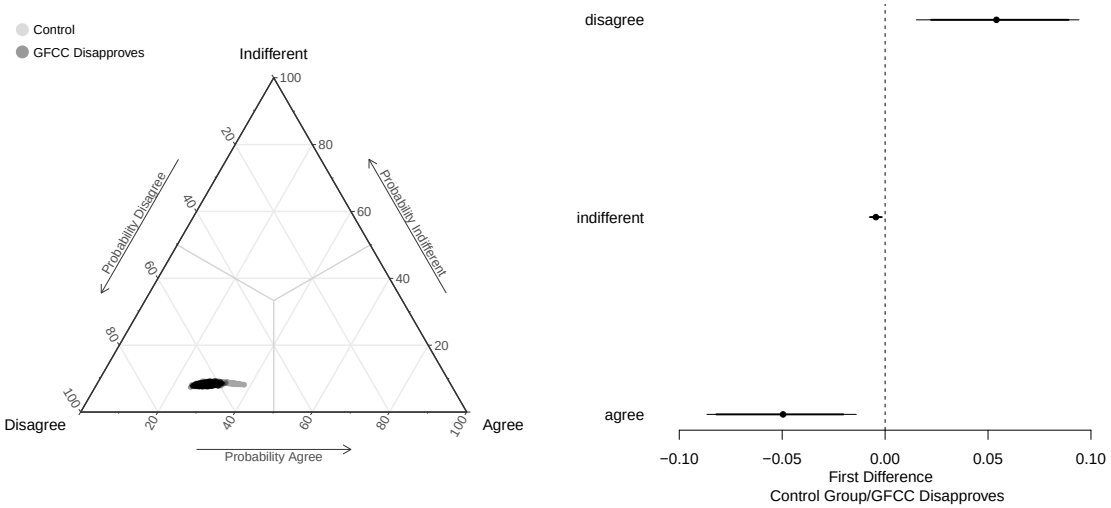
The simulation results show that even if respondents receive the treatment that the GFCC approves the school security law, the probability of disagreeing remains high (the black dots are still located on the left side of the vertex). However, we can also observe an increase in the probability of agreeing, as the black dots are located to the right side of the gray control group points. To illustrate, respondents in the control group have a 57% predicted probability of disagreeing with the school security law, a 9% probability of being indifferent and a 34% probability of agreeing on average. By contrast, respondents who received the GFCC approves endorsement have on average a probability of 52% of disagreeing, 9% of being indifferent and 39% of agreeing. The right side of the figure shows the corresponding first differences (same as in the main analysis). Figure A.4 provides similar plots for the GFCC disapproves treatment effect.

Figure A.3 – Pred. Prob. and First Difference of Control Group and GFCC Approves Endorsement



Note: Left Side: Ternary plot of the simulated predicted probabilities for the school security law. N of simulation $= 1,000$. The probabilities are calculated by setting the respective dummy (control group and GFCC approves) of the model to one and all other dummies to zero.
Right side: First differences between the predicted probabilities of the control group and the GFCC approves treatment group from the same simulation. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals.

Figure A.4 – Predicted probabilities and first differences of Control and GFCC Disapproves Treatment



Note: Left Side: Ternary plot of the simulated predicted probabilities for the school security law. N of simulation $= 1,000$. Expected values are calculated by setting the respective dummy (control group and GFCC disapproves) of the model to one and all other dummies to zero.
Right side: First differences between the expected probabilities between the control group and the GFCC approves treatment group from the same simulation. The points represent the first difference point estimates and the thin and thick bars represent 95% and 90% confidence intervals.

A.7 Ordered Probit for Party Affiliation in Germany and France

Table A.3 – Results of Ordinal Probit Regression for AfD Voters, GIP

	DV: 3-Scale Rating of Likeability School Security Law
Data Security Official Approves	0.187
	(0.140)
Data Security Official Disapproves	−0.199
	(0.150)
GFCC Approves	0.253**
	(0.117)
GFCC Disapproves	−0.197
	(0.121)
Conference of the Ministers of Education Approves	0.067
	(0.144)
Conference of the Ministers of Education Disapproves	−0.206
	(0.150)
AfD Voter (=1)	1.597***
	(0.239)
Data Security Official Approves X AfD-Voter	−0.476
	(0.398)
Data Security Official Disapproves X AfD-Voter	0.022
	(0.379)
GFCC Approves X AfD-Voter	−0.169
	(0.322)
GFCC Disapproves X AfD-Voter	−0.502
	(0.309)
Conference of the Ministers of Education Approves X AfD-Voter	−0.807**
	(0.402)
Conference of the Ministers of Education Disapproves X AfD-Voter	−0.083
	(0.376)
disagree | indifferent	0.620***
	(0.086)
indifferent | agree	0.991***
	(0.088)
N	3427

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment group and for the voter affiliation. The baseline category is the control group (no treatment). Standard errors are clustered by respondents. Smaller *N* than main analysis because of missing values in the party affiliation question variable.

Table A.4 – Results of Ordinal Probit Regression for Green-supporters using GIP data

	DV: 3-Scale Rating of Likeability School Security Law
Data Security Official Approves	0.097
	(0.137)
Data Security Official Disapproves	−0.144
	(0.145)
GFCC Approves	0.213*
	(0.115)
GFCC Disapproves	−0.285**
	(0.114)
Conference of the Min. of Educ. Approves	−0.143
	(0.145)
Conference of the Min. of Educ. Disapproves	−0.271*
	(0.142)
Green Voter (=1)	−1.188***
	(0.259)
Data Security Official Approves X Green-Voter	0.034
	(0.433)
Data Security Official Disapproves X Green-Voter	0.034
	(0.414)
GFCC Approves X Green-Voter	0.309
	(0.327)
GFCC Disapproves X Green-Voter	0.392
	(0.359)
Conference of the Min. of Educ. Approves X Green-Voter	0.775**
	(0.378)
Conference of the Min. of Educ. Disapproves X Green-Voter	0.433
	(0.440)
disagree | indifferent	0.240***
	(0.084)
indifferent | agree	0.598***
	(0.085)
N	3427

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment group and for the voter affiliation. The baseline category is the control group (no treatment). Standard errors are clustered by respondents. Lower *N* than main analysis because of missing values in the party affiliation question variable.

Table A.5 – Results of Ordinal Probit Regression for Front National-voters, ENEF

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	−0.015
	(0.093)
CC Disapproves	−0.188*
	(0.097)
Haut Conseil de l'Éducation Approves	0.140
	(0.095)
Haut Conseil de l'Éducation Disapproves	−0.082
	(0.093)
FN-Voter (=1)	0.659***
	(0.191)
CC Approves × FN-Voter	0.285
	(0.283)
CC_Disapproves × FN-Voter	−0.110
	(0.254)
Haut Conseil d'Education × FN-Voter	−0.023
	(0.270)
Haut Conseil d'Education × FN-voter	−0.103
	(0.273)
disagree | indifferent	−0.262***
	(0.067)
indifferent | agree	0.059
	(0.067)
N	1830

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment groups and for the voter affiliation. The baseline category is the control group (no treatment).

Table A.6 – Results of Ordinal Probit Regression for Parti Socialiste-voters, ENEF

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	−0.003
	(0.096)
CC Disapproves	−0.246**
	(0.096)
Haut Conseil de l'Éducation Approves	0.076
	(0.098)
Haut Conseil de l'Éducation Disapproves	−0.140
	(0.096)
PS-Voter (=1)	−0.682***
	(0.165)
CC Approves × PS-Voter	0.037
	(0.237)
CC_Disapproves × PS-Voter	0.263
	(0.270)
Haut Conseil d'Education × PS-Voter	0.405*
	(0.234)
Haut Conseil d'Education Disapproves × PS-Voter	0.263
	(0.229)
disagree | indifferent	−0.463***
	(0.069)
indifferent | agree	−0.145**
	(0.068)
N	1830

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment groups and for the voter affiliation. The baseline category is the control group (no treatment).

A.8 High Trust and Low Trust in Germany and France

Table A.7 – Ordinal Probit Regression for Trust Interaction, Germany

	DV: 3-Scale Rating of Likeability School Security Law
Data Security Official Approves	−0.001
	(0.398)
Data Security Official Disapproves	0.850**
	(0.428)
GFCC Approves	0.462
	(0.351)
GFCC Disapproves	0.136
	(0.328)
Conference of the Min. of Educ. Approves	−0.812**
	(0.400)
Conference of the Min. of Educ. Disapproves	0.747*
	(0.419)
Institutional Trust GFCC	−0.188***
	(0.047)
Data Security Official Approves × Institutional Trust	0.026
	(0.075)
Data Security Official Disapproves × Institutional Trust	−0.189**
	(0.081)
GFCC Approves × Institutional Trust	−0.044
	(0.066)
GFCC Disapproves × Institutional Trust	−0.070
	(0.062)
Conference of the Min. of Educ. Approves × Institutional Trust	0.123
	(0.075)
Conference of the Min. of Educ. Disapproves × Institutional Trust	−0.201**
	(0.081)
disagree | indifferent	−0.619**
	(0.251)
indifferent | agree	−0.253
	(0.251)
N	4760

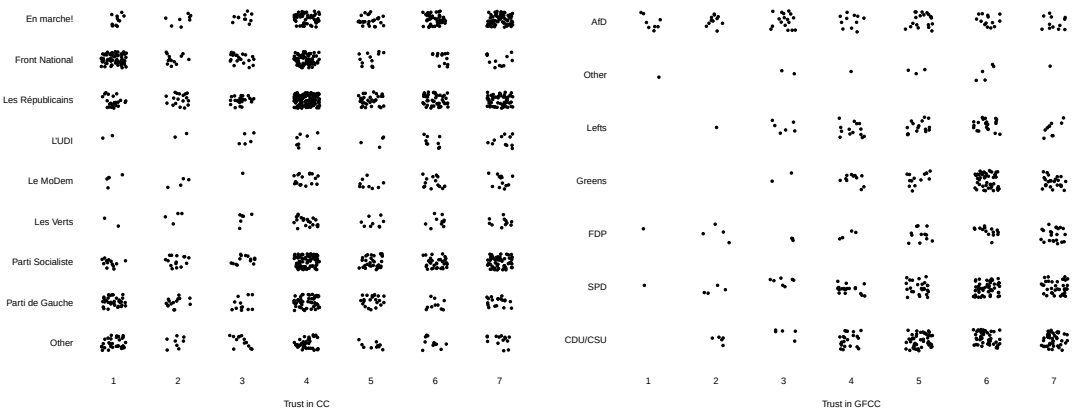
***p < .01; **p < .05; *p < .1. The baseline category is the control group (no treatment). Trust is a numeric variable ranging from 1 to 7.

Table A.8 – Ordinal Probit Regression for Trust Interaction, ENEF France

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	0.081
	(0.203)
CC Disapproves	−0.239
	(0.200)
Haut Conseil d'Education Approves	−0.104
	(0.205)
Haut Conseil d'Education Approves	−0.299
	(0.212)
Institutional Trust in CC	−0.078**
	(0.031)
CC Approves × Institutional Trust	−0.009
	(0.044)
CC Disapproves × Institutional Trust	0.025
	(0.044)
Haut Conseil d'Education × Institutional Trust	0.062
	(0.045)
Haut Conseil d'Education Disapproves × Institutional Trust	0.055
	(0.045)
disagree | indifferent	−0.675***
	(0.145)
indifferent | agree	−0.309**
	(0.145)
N	1943

***p < .01; **p < .05; *p < .1. The baseline category is the control group (no treatment). Trust is a numeric variable ranging from 1 to 7.

Figure A.5 – Distribution of Institutional Trust in the GFCC and the CC over Party Support



(a) Trust over Party Identification, CC *(b) Trust over Party Identification, GFCC*

Note: The trust question in France was asked one wave after the experiment took place. Therefore, I used all respondents to plot the trust in the CC over party support. For the GFCC, I only used respondents of the control group, because the question for institutional trust was asked in the same wave as the experimental treatment.

A.8.1 Distribution of Institutional Trust in Germany and France over Party Support

Figure A.5 shows the distribution of institutional trust in the GFCC and the CC of respondents with different party affiliations. We can see clear differences between the two countries. In France, the CC only enjoys relatively high support amongst party members of the En marche!, and to some extent of the Republicans and the Socialist party members. Across all parties, the middle category (neither distrust nor trust) is most frequently selected answer. There is a substantial number of supporters of the Front National and, to some extent, of the Left Party who do not trust the Conseil at all.

In Germany, the picture is totally different. The majority of the respondents of all parties do have a high or very high trust in the GFCC. It is particularly interesting that even supporters of the AfD, a party which often attacks the government or other political institutions, do not hold especially negative views about the GFCC. This is, once more, an indicator that the GFCC is respected across the entire political spectrum and additional evidence for the broad public support the GFCC enjoys.

A.9 Robustness Tests and Diagnostic Checks

A.9.1 Using Another Policy Issue in the ENEF Survey

In the ENEF Wave 16, another policy issue was used to test the legitimacy-conferring capacity of the CC. This policy issue is about a retirement policy. The issue reads as follows:

D'après le Ministère des Solidarités et de la Santé, le régime général de retraite et le fonds de solidarité vieillesse présentaient, en 2016, un déficit de 2,7 milliards d'Euros. Imaginez le scénario suivant : le Parlement adopte une nouvelle loi sur les retraites. Pour avoir droit à une retraite à taux plein, la durée minimale de cotisations retraite nécessaire augmente de deux ans. Cette loi a pour objectif de pérenniser le régime des retraites en augmentant la durée de cotisations des actifs, compte tenu de l'augmentation de l'espérance de vie.

Respondents were then asked to rate this policy on a 5-point scale. The experimental manipulation was similar to the prior survey experiment: respondents were told that the CC either approves or disapproves the law, or that the Conseil d'Orientation des Retraites (a national expert body) approves the law or not. The dependent variable was recoded again into three categories (disagree, indifferent, agree). Table A.9 shows the results of an ordered probit regression using the experimental groups as the independent variable (included as dummies) and the control group as the reference category.

Table A.9 – Results of Ordinal Probit Regression, Retirement Policy ENEF

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	0.030
	(0.081)
CC Disapproves	0.114
	(0.082)
Conseil d'Orientation des Retraites Approves	0.018
	(0.082)
Conseil d'Orientation des Retraites Approves	0.021
	(0.084)
disagree | indifferent	0.135**
	(0.058)
indifferent | agree	0.575***
	(0.059)
N	2050

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment groups. The baseline category is the control group (no treatment). For this regression, the retirement policy of ENEF wave 16 was used.

A.9.2 Ordered Probit Regression Results Estimated Separately for GIP Wave 26 and Wave 27

Table A.10 – Results of Ordinal Probit Regression Using GIP Wave 26, November 2016

	DV: 3-Scale Rating of Likeability School Security Law
GFCC Approves	0.141^{*}
	(0.073)
GFCC Disapproves	-0.116
	(0.075)
Minister of Education Approves	-0.011
	(0.074)
Minister of Education Disapproves	-0.149^{**}
	(0.075)
disagree | indifferent	0.239^{***}
	(0.053)
indifferent | agree	0.476^{***}
	(0.053)
N	2748

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment). Only data from the GIP Wave 26 from November 2016 is used for estimation.

Table A.11 – Results of Ordinal Probit Regression Using GIP Wave 27, January 2017

	DV: 3-Scale Rating of Likeability School Security Law
GFCC Approves	0.116
	(0.074)
GFCC Disapproves	-0.180^{**}
	(0.075)
Data Security Official Disapproves	0.042
	(0.074)
Data Security Official Disapproves	-0.131^{*}
	(0.075)
disagree | indifferent	0.162^{***}
	(0.053)
indifferent | agree	0.385^{***}
	(0.053)
N	2693

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment). Only data from the GIP Wave 27 from January 2017 is used for estimation.

A.9.3 Baseline Results of Ordered Probit Using Original Five-point Scales

Table A.12 – Results of Ordinal Probit Regression GIP, Original 5-Point Scale

	DV: 5-point Scale Rating of Likeability School Security Law
Data Security Official Approves	0.087
	(0.056)
Data Security Official Disapproves	-0.126^{**}
	(0.057)
GFCC Approves	0.131^{***}
	(0.047)
GFCC Disapproves	-0.169^{***}
	(0.047)
Minister of Education Approves	-0.069
	(0.058)
Minister of Education Disapproves	-0.189^{***}
	(0.060)
1 | 2	-0.336^{***}
	(0.0348)
2 | 3	0.188^{***}
	(0.0346)
3 | 4	0.419^{***}
	(0.0349)
4 | 5	1.211^{***}
	(0.0379)
N	5208

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment). The original 5-point scale of the GIP survey was used.

Table A.13 – Results of Ordinal Probit Regression ENEF School Security Law, Original 5-point scale

	DV: 5-point Scale Rating of Likeability School Security Law
CC Approves	0.103
	(0.066)
CC Disapproves	−0.074
	(0.066)
Haut Conseil de l'Éducation Approves	0.145**
	(0.066)
Haut Conseil de l'Éducation Disapproves	−0.103
	(0.066)
1 | 2	−1.011***
	(0.052)
2 | 3	−0.328***
	(0.049)
3 | 4	0.039
	(0.049)
4 | 5	0.946***
	(0.051)
N	2566

***p < .01; **p < .05; *p < .1. All included variables are dummies for the treatment group. The baseline category is the control group (no treatment). The original 5-point scale of the ENEF survey was used.

A.9.4 Individual Heterogeneity - Knowledge about the Court

Table A.14 and A.15 show ordered probit regression results for an interaction of the experimental groups with an individual's knowledge. The GIP included two questions where respondents needed to identify the correct name of a judge currently sitting on the bench. Respondents who did not correctly identify any of the two individuals (Susanne Baer, Judge of the first Senate and Chief Justice Andreas Voßkuhle) comprise the group of not knowledgeable respondents (roughly 77% of all respondents in wave 26). The other respondents were those who were able to answer at least one of the two questions correctly.

In the German model, the knowledge dummy itself is statistically insignificant. Also, all the interaction effects are not significant. There is, therefore, no evidence that the treatment effects are conditional on a respondent's knowledge about the GFCC.

In France, respondents where directly asked about their knowledge about the CC:

Voici une liste d'institutions de la République. Pour chacune d'entre-elles, dites si...

- 1 Vous la connaissez de nom et vous êtes informé sur ses fonctions
- 2 Vous la connaissez de nom mais vous n'êtes pas informé sur ses fonctions
- 3 Vous ne la connaissez pas

46% of the respondents have heard about the CC and are knowledgeable about its functioning, whereas 42% have a least heard about it, and 5% do not know the CC (8% refused to answer). The respondents who have heard about the CC and know its functioning are coded as respondents with high knowledge (=1), whereas the rest is coded as not knowledgeable (= 0).

In the model for France (Table A.15), the knowledge dummy itself is statistically significant and the CC Disapproves endorsement turns out to be statistically significant, too. However, none of the interactions show a statistically significant effect. I use simulations to investigate whether knowledgeable respondents are different to not knowledgeable ones. For this, I compare the endorsement effect (the effect between the control group and the respective treatment group) for knowledgeable respondents and not knowledgeable ones. The corresponding first differences show that there is no statistically significant difference between these two groups.

Table A.14 – Results of Ordinal Probit Regression for Knowledge Interaction GIP

	DV: 3-Scale Rating of Likeability School Security Law
Data Security Official Approves	0.041
	(0.118)
Data Security Official Disapproves	−0.122
	(0.121)
GFCC Approves	0.168*
	(0.098)
GFCC Disapproves	−0.199**
	(0.099)
Conference of the Min. of Educ. Approves	−0.127
	(0.124)
Conference of the Min. of Educ. Disapproves	−0.197
	(0.128)
High Knowledge (=1)	−0.161
	(0.146)
Data Security Official Approves × High Knowledge	0.360
	(0.248)
Data Security Official Disapproves × High Knowledge	−0.179
	(0.258)
GFCC Approves × High Knowledge	0.144
	(0.202)
GFCC Disapproves × High Knowledge	−0.141
	(0.212)
Conference of the Min. of Educ. Approves × High Knowledge	0.023
	(0.257)
Conference of the Min. of Educ. Disapproves × High Knowledge	−0.311
	(0.256)
disagree | indifferent	0.271***
	(0.069)
indifferent | agree	0.647***
	(0.070)
N	5208

***p < .01; **p < .05; *p < .1. High Knowledge is a dummy variable indicating whether a respondent is knowledgeable of the GFCC (=1) or not (=0). Standard errors clustered by respondents.

Table A.15 – Results of Ordinal Probit Regression for Knowledge Interaction ENEF

	DV: 3-Scale Rating of Likeability School Security Law
CC Approves	0.062
	(0.110)
CC Disapproves	-0.217^{**}
	(0.105)
Haut Conseil d'Education Approves	0.079
	(0.107)
Haut Conseil d'Education Disapproves	-0.096
	(0.109)
High Knowledge (=1)	-0.214^{**}
	(0.109)
CC Approves × High Knowledge	0.093
	(0.153)
CC Disapproves × High Knowledge	0.182
	(0.154)
Haut Conseil d'Education Approves × High Knowledge	0.178
	(0.153)
Haut Conseil d'Education Disapproves × High Knowledge	0.008
	(0.155)
disagree | indifferent	-0.409^{***}
	(0.076)
indifferent | agree	-0.063
	(0.076)
N	2401

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$. High Knowledge is a dummy variable indicating whether a respondent is knowledgeable of the CC (=1) or not (=0).

APPENDIX B

Chapter 3

B.1 Appendix with supplementary information for the expanded dictionary

German seed words from the German LIWC dictionary (Wolf et al., 2008) in their basic form (inflicted versions to infinitive):

- “angeblich”, “beinahe”, “duerfte”, “eigentlich”, “erscheinen”, “etwa”, “eventuell”, “fast”, “gelegentlich”, “gezoegert”, “gezweifelt”, “hoffen”, “hoffentlich”, “Hoffnung”, “hoffnungslos”, “hoffnungsvoll”, “irgendein”, “irgendetwas”, “irgendjemand”, “irgendwann”, “irgendwer”, “irgendwie”, “irgendwo”, “jederzeit”, “jemand”, “koennte”, “labil”, “manche”, “moeglich”, “mutmaßen”, “mutmaßlich”, “nahezu”, “probieren”, “provisorisch”, “schaetzen”, “scheinen”, “sozusagen”, “unbestimmt”, “uneins”, “ungefaehr”, “ungewiss”, “unklar”, “unsicher”, “vage”, “vermeintlich”, “vermutlich”, “vielleicht”, “vielleicht”, “vorsichtig”, “wahrscheinlich”, “zeitweise”, “ziemlich”, “zoegerlich”, “zoegern”, “Zufaelle”, “zufaellig”, “Zufall”

Manually selected similarity candidates based on word embeddings (terms with a “_” indicate a bi-gram):

- “unbenommen” , “Differenzierung”, “Typisierung” , “annähernd”, “angemessener” , “gegebenenfalls”, “freistehen”, “ausreichend”, “tragfähiger”, “genügend”, “zureichend” , “Ermessen” , “Disposition”, “Generalisierung”, “Typisierung” , “auslegungsfähig” , “Gestaltung”, “Ausgestaltung”, “generalisierende” , “Möglichkeiten”

, “unterschiedlich”, “verschieden”, “differenzieren” , “Weisungsrechts” , “Einschätzung”, “bedarfsgerechten,” “zweifelhaft” , “überlassen” , “Typisierungsspielraum”, “Gestaltungsraum”, “Gestaltungsspielraum” , “Spielraum” , “Gestaltungsfreiraum”, “Gestaltungsbereich” , “einen_weiten” , “Gestaltungsfreiheit” , “Typisierungsspielraum”, “Gestaltungsspielraums”, “Entscheidungsspielraum”, “besonders_weiten”, “eingeräumten_weiten”, “Gestaltungsraums” , “Ermessensspielraum” , “Prognosespielraums”, “Einschätzungs_Wertungs”, “weiten Gestaltungsraum”, “politischen_Ermessens” “Einschätzungs” , “Einschätzungsspielraum”

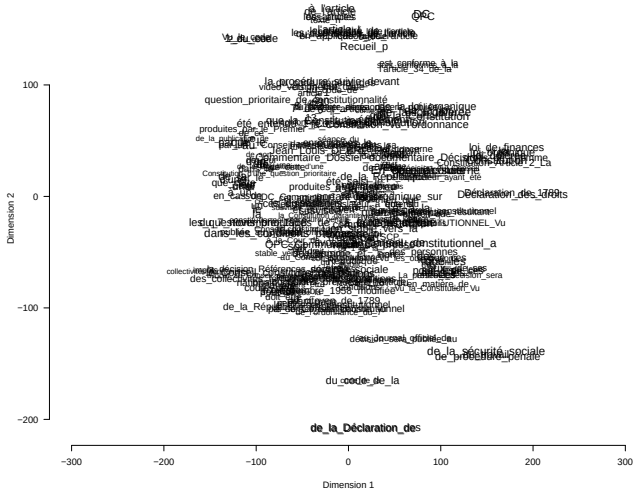
French seed words from the French LIWC dictionary (Piolat et al., 2011) in their basic form (inflicted versions to infinitive):

- “apparaître”, “apparemment”, “approximativement”,”assumer”, “brouiller”, “chance”, “déconcerter”, “dépendre”, “déranger”, “désorientant”, “désorienter”, “embarrasser”, “embrouiller”, “espérer”, “estimant”, “estimer”, “éventuellement”, “éventuels”, “éventuelles”, “figurer”, “guère”, “hésiter”, “hypothétique”, “imaginer”, “à l’occasion”, “occasions”, “paraît”, “parfois”, “parier”, “partiellement”, “peut-être”, “pouvoir”, “pratiquement”, “presque”, “présumer”, “probablement”, “projeter”, “quelquefois”, “questionner”, “reconsidérer”, “remarquer”, “repartir”, “reprendre”, “rétorquer”, “risquer”, “soupçonner”, “suggérer”, “supposer”

Manually selected similarity candidates based on word embeddings (words with a “_” indicate a bi-gram/tri-gram):

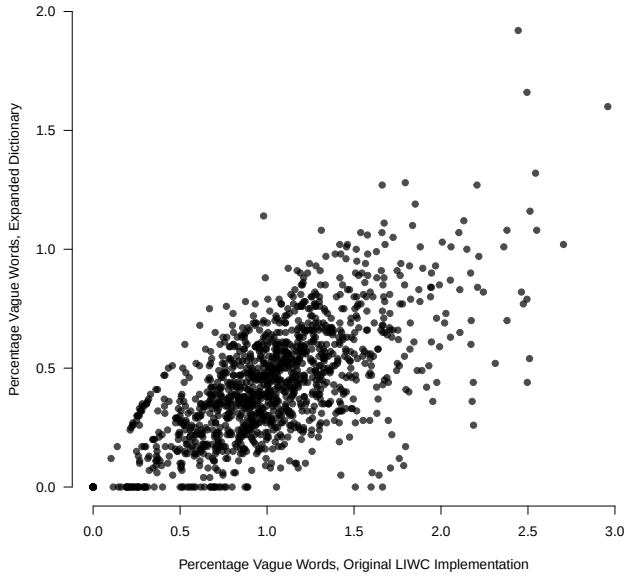
- “rester_à_l’appréciation”, “mesures_nécessaires”,”mesure_le_législateur”, “s’avèrer”, “cas_échéant_d’autres”, “convenable”, “appropriée”, “convenir”, “traitement_approprié”, “alternatives”, “imprécis”, “inexact”, “détachement”, “fair_la_part”, “s’il_est_vrai_que”, “latitude”, “marge_d’appréciation”

Figure B.1 – Two-dimensional reduction of the 300 most frequently occurring terms, France



Note: This figure shows a mapping of the 300 most frequently appearing terms in the word embeddings trained on the French constitutional court decision corpus. The dimensions are not identified. We can see that similar words or terms are clustered next to each other. QPC, DC and other names for decisions types are, for instance, grouped next to each other in the upper right corner. The same accounts for different references to law types (loi de finance, loi déférée, loi organique) that are clustered (right side). Because of the less amount of training data in the French application, the map is less useful so see word clusters than the German map in the main text.

Figure B.2 – Scatterplot of Proportion of Vague Words, LIWC versus Expanded Dictionary, CC Application



Note: This figure shows the proportion of vague words identified by the original French LIWC dictionary on the X-axis and the French expanded dictionary on the Y-axis when scoring the same decision texts.

B.2 Appendix with supplementary information for NLP Classifier in the GFCC application

B.2.1 Annotation procedure and coding instructions

The annotation procedure was conducted in two stages. In the first stage, the pre-test stage, the annotators received detailed project instructions, including a brief explanation of the project's purpose and context, and some information about the structure of court decisions. One of the annotator has a minor in public law. The other two annotators have had at least two public law classes in their undergraduates, and received detailed coding instructions and examples. The instructions also included detail information about the concept of judicial policy implementation vagueness. The annotators then coded a random sample of over 300 sentences. Afterwards, a meeting was held to discuss potential issues. In the second phase, the annotators had to label the remaining 3,500 sentences. The following page show the coding instructions that the annotators received.

Leitfaden für die Annotation von Vagheit in Bundesverfassungsgerichtsentscheidungen

Sebastian Sternberg

20. Februar 2019

Kontext des Projekts:

Das Bundesverfassungsgericht ist das Verfassungsgericht der Bundesrepublik Deutschland und unter anderem dafür verantwortlich, die Verfassungsmäßigkeit von verabschiedeten Gesetzen zu überprüfen. Dabei kann das Gericht nicht nur ein Gesetz als unvereinbar mit dem Grundgesetz und dafür für nichtig erklären, sondern auch konkrete Vorschläge machen, wie ein Gesetz ausgelegt oder geändert werden muss, damit es im Einklang mit der Verfassung ist. Das Gericht legt seine Entscheidung immer in schriftlicher Form in den sogenannten Bundesverfassungsgerichtsentscheidungen vor.

Ziel des Projekts

Ziel dieses Projektes ist es herauszufinden, unter welchen Umständen sich das Gericht dazu entschließt, dem Gesetzgeber genau vorzuschreiben wie ein neues Gesetz umgesetzt werden soll und wann es davon absieht. Der Hintergrund davon ist, dass bestimmte Annahmen getroffen werden können, wann das Gericht strategisch darauf verzichtet, dem Gesetzgeber etwas vorzuschreiben (eine *vage* Entscheidung zu schreiben) und wann es, ebenfalls strategisch, präzisiert wie ein Gesetz modifiziert werden soll (eine *nicht vage* Entscheidung).

Um diese Fragestellung beantworten zu können ist es notwendig zu messen, wann eine Entscheidung vage oder nicht vage ist. Hierfür soll mit Hilfe dieses Projekts eine computerbasierte automatisierte Identifikation von vagen Textstellen erfolgen, die es erlaubt, Sätze mit vagen Vorschriften zur Umsetzung von Gesetzen innerhalb einer großen Anzahl von Texten automatisiert (d.h. ohne menschliche Steuerung) zu erkennen. Dies erfolgt mithilfe eines Klassifizierungsalgorithmus, welcher anhand von Beispielsätzen lernen kann, welche Satzeigenschaften einen vagen oder nicht vagen Satz repräsentieren.

Definition und Motivation des Konzept: Vagheit in juristischen Texten

Vagheit ist konzeptionnel nicht äquivalent zu *Ambiguität*. Ein sprachlicher Ausdruck wird als ambiguitiv (mehrdeutig) angesehen, wenn er zwei oder mehr klar abgegrenzte Bedeutungen hat. Ein Standardbeispiel für ein doppeldeutiges Wort ist "Bank", da es sowohl für eine Sitzgelegenheit aus Holz, Stein oder Ähnlichem stehen kann oder für ein Unternehmen, das Geld- und Kreditgeschäfte betreibt. Im Gegensatz dazu wird ein Ausdruck als vage bezeichnet, wenn die sprachlichen Grenzen des Begriffs nicht klar separiert sind. Ein Standardbeispiel für "*Vagheit*" ist der Begriff "groß". Es ist nicht eindeutig definiert, was genaue eine "große" Person ausmacht. Jemand der anhand der Durchschnittsgröße als "groß" bezeichnet wird muss nicht als groß in-

nerhalb eines Basketballteams bezeichnet werden.

Weil das Konzept der Vagheit bereits an sich vage ist und stark auf den Satzkontex ankommt, ist es notwendig eine größere Anzahl an Beispielssätze manuell als vage bzw. nicht vage zu kennzeichnen. Dazu ist es notwendig das menschliche AnnotatorInnen entscheiden, welche Sätze vage oder nicht vage sind, damit später der Computer bzw. der Algorithmus lernen kann, welche Eigenschaften eines Satzes für einen vagen bzw. nicht vagen Satz stehen. Die Erfahrung aus anderen Projekten hat gezeigt, dass vage Sätze im Durchschnitt wesentlich seltener vorkommen als nicht vage Sätze.

Das Konzept der Vagheit sollte beim Annotieren immer innerhalb des oben erklärten juristischen Kontextes verstanden werden. Dies bedeutet, dass Vagheit im Sinne von der Vagheit von Vorgaben des Gerichts an den Gesetzgeber für die Umsetzung von Entscheidungen verstanden werden soll. Weitere Erläuterungen und Beispiele finden sich im weiteren Verlauf des Textes.

Erläuterung der Textform:

Die Entscheidungstexte werden von den Richtern selbst geschrieben und unterschrieben. Die Texte sind typischerweise sowohl an die juristische Community als auch die Medien und die Öffentlichkeit addressiert.

Typischerweise wird in den Entscheidungen zuerst der Kontext dargelegt, d.h. um was es in dem zu beurteilenden Gesetz inhaltlich geht und welcher Grundrechtsverstoß vom Antragsteller anhängig gemacht wird. Danach werden, falls vorhanden und nötig, Expertenmeinungen sowie die Meinungen von Kläger und Beklagten diskutiert. Danach beginnt das Gericht mit der eigentlichen Prüfung des Gesetzes, in dem es beispielsweise prüft ob die vom Antragssteller genannten Verstöße einschlägig sind oder aber andere Grundrechte verletzt wurden. Im letzten Abschnitt des Textes wird sodann die Entscheidung des Gerichts bekanntgegeben. Hier kann das Gericht auch, falls es dies als nötig befindet, konkrete Anweisung für die Umsetzung bzw. Änderung eines neuen Gesetzes geben. In den Entscheidungen selbst kann es um eine Vielzahl von Themen gehen, zum Beispiel Schwangerschaftsabbruch, Beamtenrente, Numerus Clausus für Medizinstudenten oder die Pleichstellung von gleichgeschlechtlichen PartnernInnen. Bundesverfassungsgerichtsentscheidungen und deren Sätze sind im Durchschnitt eher länger und sprachlich auf gehobenem Niveau, d.h. benutzen viele Fremdwörter oder auch juristische Fachbegriffe.

Annotations-Aufgabe und Hinweise

Konkret besteht die Annotations-Aufgabe darin, 1500 zufällig aus Gerichtsentscheidungen ausgewählte Sätze als entweder vage oder nicht vage zu klassifiziert. Dabei wird jeweils ein Satz gezeigt, und die Annotatorienden müssen entscheiden, ob dieser Satz vage ist oder nicht. Mithilfe dieser sogenannten Trainingsdaten kann ein Algorithmus dann lernen, welche Satzeigenschaften für einen vagen bzw. nicht vagen Satz stehen. Hierbei ist es wichtig zu betonen, dass es ausschließlich auf die Beurteilung des/der Annotierenden ankommt. Auch gibt es keine falsche oder richtige Antwort, sondern es zählt ausschließlich die wahrgenommene Vagheit der Sätze. Es ist nicht möglich, ein und denselben Satz als vage bzw. nicht vage zu annotieren. An dieser Stelle soll nach

einmal betont werden, dass vage Sätze im Durchschnitt wesentlich seltener vorkommen als nicht vage Sätze.

Da einzig und allein die Qualität der annotierten Sätze darüber entscheidet, ob der Algorithmus in der Lage ist, das Konzept der Vagheit in juristischen Texten zu erlernen, sollten AnnotatorInnen nie länger als eine Stunde am Stück ohne Pause annotieren. Ansonsten kann es zum sogenannten "fatigue effect"kommen, welcher die zunehmende Unkonzentriertheit und Veringerung der Zurechnungsfähigkeit über die Zeit einer zu bearbeitenden Aufgabe beschreibt, mit der die Qualität der Arbeit stark sinkt.

Die Annotation wird über eine vorprogrammierte Online-App durchgeführt. Dabei werden nacheinander Sätze eingeblendet, welche anschließend annotiert werden sollen. Um unnötiges Klicken mit der Maus zu vermeiden, können Tastaturtasten frei belegt werden, sodass ausschließlich mit der Tastatur annotiert werden kann. Wenn ein Satz als vage annotiert wurde gibt es darüber Hinaus noch die Möglichkeit in einem Textfeld einzugeben, welches Wort oder welcher Ausdruck im Satz die Entscheidung zur Klassifikation als vage beeinflusst hat (sogenannte trigger terms).

Beispielsätze

Im Folgenden werden einige Beispielsätze für die "vage"bzw. "nicht vage"Kategorie aufgeführt. Vage Sätze können beispielsweise so aussehen:

- Dem Gesetzgeber kommt bei der Erfüllung dieser Schutzpflicht ein weiter Einschätzungs-, Wertungs- und Gestaltungsfreiraum zu, der auch Raum für die Berücksichtigung konkurrierender öffentlicher und privater Interessen läßt.

Hierbei betont das Gericht den weiten Entscheidungsspielraum des Gesetzgebers, und betont noch einmal den "Raum"für etwaige Abwägungen. Der folgende Satz ist schon weniger klar:

- Es ist dem Gesetzgeber aber unbenommen, die Wirkung der vorliegenden Entscheidung auch auf bereits bestandskräftige Bescheide zu erstrecken; von Verfassungs wegen verpflichtet ist er hierzu nicht.

Hierbei zeigt das Wort “unbenommen”, dass der Gesetzgeber eine gewisse Freiheit hat, wie er die vorliegende Entscheidung umsetzt. Dies wird durch den Nachtrag” von Verfassungs wegen verpflichtet” noch einmal betont.

Nicht vage Sätze können viele Formen annehmen. Konkret auf die Umsetzung von Entscheidungen bezogen kann dies so aussehen:

- Der Gesetzgeber ist verpflichtet, bis zum 30. Juni 2001 eine verfassungsgemäße Regelung zu treffen.

Hier verpflichtet das Gericht den Gesetzgeber, eine Regelung bis zu einem festgesetzten Datum zu treffen. Da die Sätze zufällig aus den Gerichtsentscheidungen ausgewählt wurden, können es auch Sätze sein, welche nicht konkret auf Gesetze und deren Umsetzung eingehen, sondern aus anderen Stellen im Text stammen.

- In der Regel ermaechtigten diese Gesetze die Aerztekammern, die Berufspflichten der Aerzte in einer Berufsordnung zu regeln.

Insgesamt werden solche Sätze die Mehrheit der zufällig ausgewählten Sätze darstellen. Darüber hinaus gibt es Fälle, welche weniger klar und deshalb schwerer zu annotieren sind:

- Einen gewissen Spielraum verschafft dem Gesetzgeber insoweit nur noch die rechnerische Einbeziehung der Ertraege, die er als abziehbare Aufwendungen und sonstige Entlastungen unbesteuert lassen will.

Hier wird zwar der “gewisse Spielraum” durch den Ausdruck “nur noch” eingeschränkt, jedoch sollte dieser Satz immer noch als vage annotiert werden. Folgender Satz ist noch schwieriger:

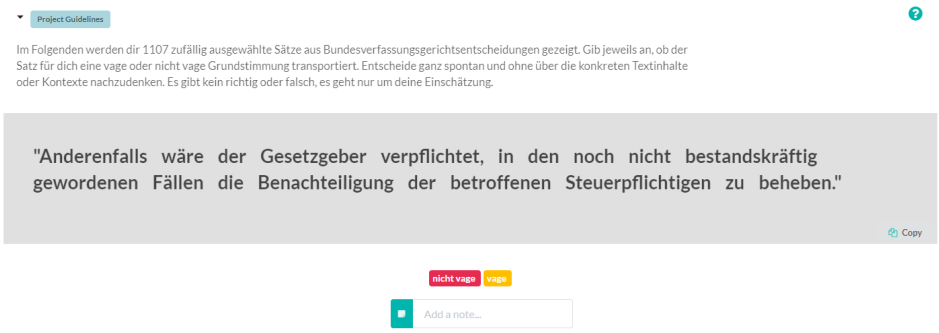
- Damit ist dem Gesetzgeber allerdings nicht jede Differenzierung verwehrt.

Auch hier sollte eher als vage annotiert werden, da dem Gesetzgeber zugestanden wird, ein gewisses Maß an Differenzierung anlegen zu dürfen.

B.2.2 Screen-shot of the software used for the annotation task

Figure B.3 shows a screenshot (in German) of the software (an onlineplatform called *dataturks* (`https://dataturks.com/`)) used for the annotation task with an example sentence. For each of the randomly sampled sentences from the GFCC decisions, the three annotators had to classify each sentence into vague or not vague. Additionally, they could add a note when they thought a certain trigger word has shaped their classification decision.

Figure B.3 – Screenshot of Software used for the Annotation Task



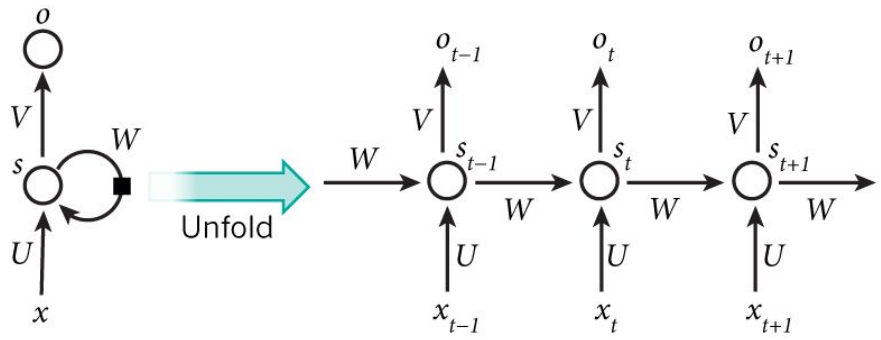
B.3 Brief overview over typical deep learning architectures used in this study

The typical RNN architecture

Figure B.4 shows an "unfolded" RNN. Unfolded here refers to writing out the network for the complete sequence. A sentence consisting of three words, for instance, would be unfolded into a three-layer network, with one layer for each word. Stated more generally, a RNN can map an input sequence with an arbitrary number of elements x_t at time t into an output sequence with elements O_t, with each O_t depending on all the previous $x_{t'}$ for $t' < t$ (Lecun, Bengio and Hinton, 2015, 442). Unlike traditional feed-forward networks, which uses different parameters at each layer, a RNN shares the same parameters (weight matrices U, V, W) across all steps. Together U and W define how to calculate the new state s_t of the network given the previous state s_{t-1} and the input x_t (together with the function f). V defines how to map the hidden state back into the outcome space.

In Figure B.4, x_t is the input at time step t, for instance a one-hot encoded vector corresponding to a word in a sentence. s_t is the hidden state at time step t, and is the "memory" of the network. s_t is calculated based on the previous hidden state and the input at the current step as follows:

Figure B.4 – An Unfolded Recurrent Neural Network

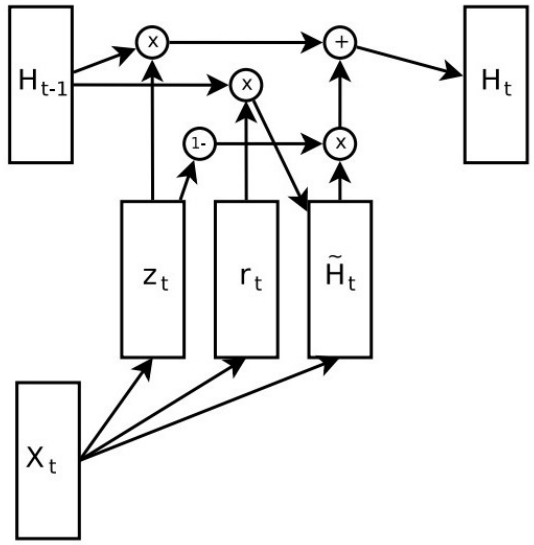


Note: Example architecture of a simple RNN, unfolded over three iterations. Figure taken from Lecun, Bengio and Hinton (2015, 442).

$$s_t = f(Ux_t + Ws_{t-1}) \tag{B.1}$$

where the function f is typically a non-linearity such as *tanh* or *ReLU*, and U and W are weight matrices. o_t is the output at step t. Is the task for instance predicting the next word in a sentence given a sequence of previous words, o_t would be a vector of probabilities over all words of the vocabulary given by $softmax(Vs_t)$.

The typical GRU architecture

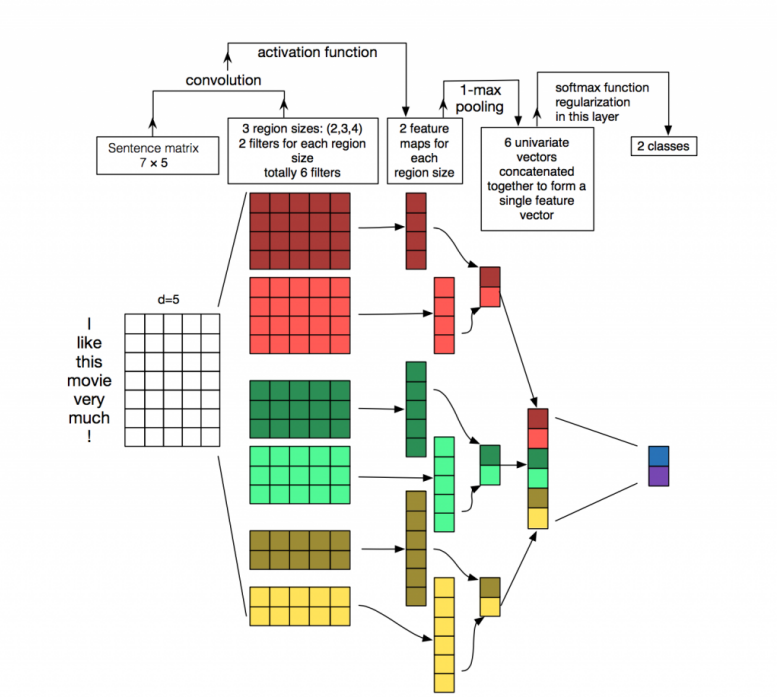


$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1}) \\
\tilde{h}_t &= tanh(W_h x_t + U_h(r_t \odot h_{t-1}) \\
h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\end{aligned}
\tag{B.2}
$$

Note: Gated Recurrent Unit network topology, with a hidden unit (state) that can adaptively remember and forget. In the above equation, for every time-step, x_t is the input vector, z_t is the update gate vector, r_t is the reset gate vector, h_t is the output vector, and $\tilde{h}_t$ is the new hidden state as the combination of new input x_t with the past hidden state h_{t-1}. W and U are parameter matrices.
The hidden state h_t is calculated using the previous hidden input h_{t-1} and the new hidden state generated $\tilde{h}_t$ considering the update gate. The reset signal r_t is responsible for determining how important the previous hidden state h_{t-1}. The reset gate controls the information flow from the hidden state and can erase the past hidden state if it finds that h_{t-1} is irrelevant. The update signal z_t is responsible for determining how much of h_{t-1} should be carried forward to the next state. For instance, if $z_t \approx 1$, then h_{t-1} is almost entirely copied out to h_t. Conversely, if $z_t \approx 0$, then mostly the new memory $\tilde{h}_t$ is forwarded to the next hidden state. Figure and its description including the notation are based on Jozefowicz, Zaremba and Sutskever (2015, 2344). Note that $\odot$ is the element-wise product. For more information on calculation and derivation see the original paper of Cho et al. (2014).

The typical CNN architecture

Figure B.5 – Example CNN architecture for sentence classification



Note: This figure shows an example Convolutional Neural Network (CNN) architecture for sentence classification. The toy sentence contains six word tokens, which are usually encoded in a word embedding space (here the dimension of this space is of arbitrary size $d = 5$). Three filter region sizes are depicted with again arbitrary size (2,3 and 4), each of which has 2 filters, so that overall 6 different filters are applied. Every filter performs convolution on the sentence matrix and generates (variable-length) feature maps. Another, more layman way of thinking about this is having six people who independently have to classify a sentence. Two persons are allowed to looking at four consecutive words at a time (second column in the graph, where the convolution takes place). Two persons are allowed to look at three consecutive words and two are allowed to only look at two consecutive words at a time. It is important to note that even though the same two people look at the same four words at a time, they will most likely build different intuitions. In neural network terms this is due to the different random initialization of the feature maps.
Given the filter or region size 2, this means a 2-word filter is applied to the sentence matrix by sliding over the first two rows the sentence matrix (thus the 2-word filter matrix is of size 2×5), then going to the next row and so on. Every time, a feature map is created by calculating the sum of the element-wise product for all its 2×5 elements. Then 1-max pooling is performed over each feature map, which simply means the largest number from each feature map is recorded. This results in an univariate feature vector that is based on all six maps. These 6 features are concatenated such that they form the final vector for the last layer. The final *softmax* layer then converts this feature vector into probabilities for classification. In this example, it is a binary classification. Figure and description are based on (Zhang and Wallace, 2015, 4).

B.3.1 Definition of performance measures

Generally, the results of a binary classifier (and any other classifier) can be summarized by a confusion matrix. In the case of binary classification this is a 2×2 table of the four possible classification outcomes of a model. The used can all be explained with the help of confusion matrices. To get class predictions from predicted probabilities of belonging to the positive class, one has to set a threshold for positive prediction. Usually, the default value of this threshold for positive prediction is 0.5. However, any other value between 0 and 1 could be a sensible threshold for positive prediction.

Confusion Matrix

		Observed	
		Positive	*Negative*
Predicted	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- **Accuracy**: $\frac{TP+TN}{TP+FP+TN+FN}$
- **Precision**: Precision is defined as : Precision $= \frac{TP}{TP+FP}$, that is the ratio of correctly classified positives and all predicted positives.
- **Recall**: Recall (also called True Positive Rate (TPR)), is defined as Recall $= \frac{TP}{TP+FN}$. It measures the fraction of positive examples that are correctly labeled.
- F_1 **score**: The F_1 score is the harmonic mean of precision and recall, and defined as $F_1 = \frac{2 \cdot TP}{2 \cdot TP+FN+FP}$.
- **Receiver operating characteristic area under the curve:** Sensitivity (recall) plotted against 1- specificity ($\frac{TN}{TN+FP}$) at various threshold settings.
- **Kappa** = $\frac{p_o - p_e}{1-p_e}$, where p_o is the observed agreement (analog to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

B.3.2 Confusion matrices of the different classifiers

Table 1: Confusion Matrix SVM

	pred:not vague	pred:vague
true:not vague	668	4
true:vague	32	13

Table 2: Confusion Matrix Random Forest

	pred:not vague	pred:vague
true:not vague	669	3
true:vague	39	6

Table 3: Confusion Matrix Naive Bayes

	pred:not vague	pred:vague
true:not vague	648	24
true:vague	34	11

Table 4: Confusion Matrix Logistic Regression

	pred:not vague	pred:vague
true:not vague	648	24
true:vague	21	24

Table 5: Confusion Matrix Expanded Dictionary

	pred:not vague	pred:vague
true:not vague	561	111
true: vague	18	27

Table 6: Confusion Matrix XGBoost

	pred:not vague	pred:vague
true:not vague	565	107
true:vague	12	33

Table 7: Confusion Matrix GRU

	pred:not vague	pred:vague
true:not vague	667	33
true:vague	5	12

Table 8: Confusion Matrix CNN

	pred:not vague	pred:vague
true:not vague	672	33
true:vague	0	12

B.3.3 Summary statistics of original and translated court decisions

Table B.1 – Summary Statistics for original and translated decisions of the CC

Statistic	N	Mean	St. Dev.	Min	Max
Word Count German	20	6,445	4,382	455	16,461
Vagueness Score German	20	0.2	0.1	0.0	0.5
Word Count French	20	7,206	4,878	1,154	19,394
vagueness Score French	20	0.4	0.2	0.1	0.8

Table B.2 – Summary Statistics for original and translated decisions of the GFCC

Statistic	N	Mean	St. Dev.	Min	Max
Word Count French	20	370.2	224.7	38	837
Vaguenss Score French	20	0.6	0.5	0.0	1.4
Word Count German	20	307.5	180.9	35	692
Vagueness Score German	20	0.4	0.6	0.0	1.7

APPENDIX C

Chapter 4

C.1 Summary statistics of the data used in the analyses

Table C.1 and Table C.2 show summary statistics for all variables used in the analysis. Note that the small numbers for the percentage of vague sentences are not erroneous, but due to the fact that the appearance of vague *sentences* is more rare than the occurrence of vague *words*. Nevertheless, as I have demonstrated in this chapter, even a single vague sentence can be enough to express judicial policy implementation vagueness.

Table C.1 – Descriptive Statistics of Variables Used in the GFCC Analysis

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Percentage vague sentences	372	0.002	0.01	0	0	0	0	0.046
Percentage vague words	372	0.38	0.22	0.00	0.24	0.35	0.48	1.64
Ideological Distance (CMP)	372	13.98	10.07	0.53	6.75	9.81	25.48	32.99
Ideological Distance (MCSS)	372	1.74	0.60	0.01	1.65	1.91	2.17	2.48
Dummy Simple Case	372	0.35	0.48	0	0	0	1	1
Dummy Second Senate	372	0.38	0.49	0	0	0	1	1
Dummy Oral Hearing	372	0.15	0.36	0	0	0	0	1
Dummy Risk Non-Compliance	372	0.87	0.34	0	1	1	1	1
Length Casefacts	372	67.35	44.59	0	39	56	83	200

Table C.2 – Descriptive Statistics of Variables Used in the CC Analysis

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Percentage vague words	558	0.42	0.23	0.00	0.28	0.40	0.57	1.66
Ideological Distance (CMP)	558	15.06	12.37	0.00	5.00	9.47	24.51	41.45
Number of legal issues	558	5.20	3.70	2	2	4	7	28
Dummy Risk Non-Compliance	558	0.32	0.47	0	0	0	1	1

C.2 Results of Fractional Logistic Regressions Using the Robust Variance Estimator Proposed by Papke and Wooldrige (1996)

Table C.3 and Table C.4 show the fractional logit regressions of the main analyses. This means that the coefficients are directly estimated using the QMLE outlined in section 4.4.3 (the fractional logit regression), and the standard errors are robust standard errors estimated by the well-known sandwich estimator.

Table C.3 – Results of Fractional Logit Regression, GFCC Analysis

	DV: Percentage of Vague Sentences	
	Model 1	**Model 2**
Ideological Distance	-0.026^{*}	-0.092^{**}
	(0.014)	(0.039)
Case Complexity (= 1)	0.592^{*}	0.566^{*}
	(0.325)	(0.321)
Second Senat (= 1)	-0.553^{*}	-0.554^{*}
	(0.323)	(0.322)
Oral Hearing	-0.418	-0.397
	(0.313)	(0.316)
Risk Non-Compliance (= 1)	0.429	-0.280
	(0.905)	(0.773)
Ideological Distance $\times$ Risk Non-Compliance		0.069^{*}
		(0.042)
Constant	-5.923^{***}	-5.263^{***}
	(1.005)	(0.802)
N	372	372

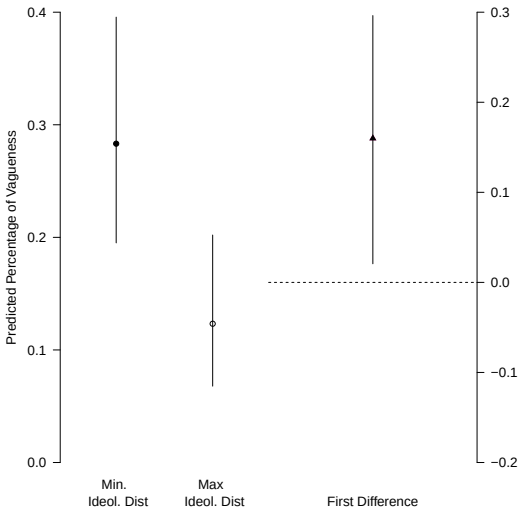
$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$
Robust Standard Errors in Parentheses.

Table C.4 – Results of Fractional Logit Regression, CC Analysis

	DV: Percentage of Vague Words	
	Model 1	**Model 2**
Ideological Distance	0.010^{***}	0.009^{***}
	(0.002)	(0.002)
Case Complexity (Number of Legal Issues)	0.025^{***}	0.025^{***}
	(0.006)	(0.006)
Risk Non-Compliance (= 1)	-0.035	-0.064
	(0.055)	(0.071)
Ideological Distance $\times$ Risk Non-Compliance		0.004
		(0.005)
Constant	-5.705^{***}	-5.695^{***}
	(0.056)	(0.059)
N	558	558

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$
Robust Standard Errors are used.

Figure C.1 – Distribution of First Difference between Minimum and Maximum Ideological Distance

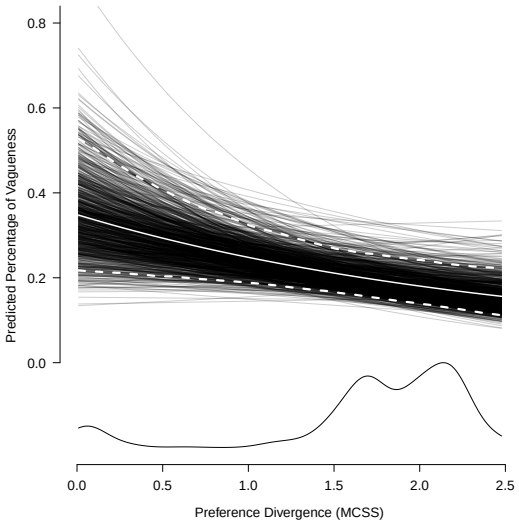


Note: This figure shows the distribution of the first difference between minimum and maximum ideological distance of Figure 4.2, including 90% confidence intervals (dashed lines) and the corresponding mean (thick line). The first difference is calculated using the observed value approach and Model 1 of Table C.3. We see that the first difference is statistically significant different from zero at the 90% level.

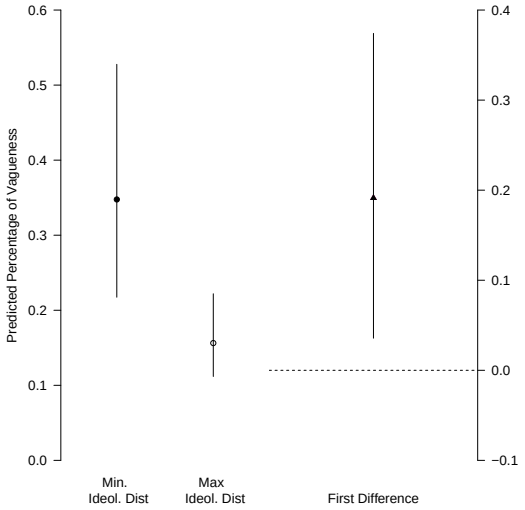
C.3 First Difference Minimum Maximum Ideological Distance, Observed value Approach

Figure C.1 explicitly tests the first difference between minimum and maximum ideological preference for the German analysis. For the simulation, the ideological distance variable was set to the minimum and the maximum. The observed value approach is used. Because the preference divergence distribution is bimodal, I also tested the first difference between the 10th and 90th quantile. All these first differences are statistically significant on the 90% level.

Figure C.2 – Effect of Ideological Distance on Decision Vagueness, GFCC



(a) Ideological Distance on Decision Vagueness



(b) First Difference Minimum Maximum Ideological Distance using MCSS

C.4 Robustness Checks

Robustness Checks GFCC

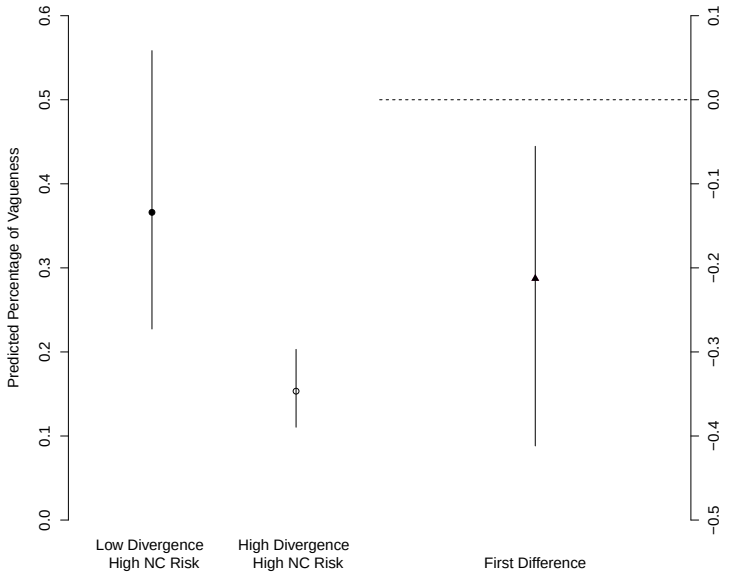
Using alternative ideology score (MCSS)

For the main analysis, I used the scores from the Comparative Manifesto Project project to calculate the ideological position of court and legislator. These scores are increasingly criticized with regard to their spatial and temporal comparability (Lowe et al., 2011; König, Marbach and Osnabrügge, 2013). In order to check whether my findings are robust to the measurement of ideological distance, I replicate my analyses but use the Manifesto Common Space Scores (MCSS) of König, Marbach and Osnabrügge (2013) instead of the original CMP scores. My findings remain robust to the usage of this different ideology measure. Like in the main analysis, I used the observed value approach and the QMLE of the fractional logit.

Figure C.2 shows that my findings are robust to the usage of the MCSS scores instead the CMP scores. The higher the preference divergence between court and legislator, the less vague the decisions of the GFCC become. Also, I tested the first difference between minimum and maximum ideological distance to evaluate whether the overall effect is statistically significant. As we can see on the right side of the Figure, the first difference is significant on the 90% level.

Figure C.3 shows the conditional effect of preference divergence and the perceived risk of noncompliance on decision vagueness for the GFCC. Again, using the MCSS scores my findings remain robust. Decision vagueness decrease when the German

Figure C.3 – Conditional Effect of Preference Divergence and Non-Compliance Risk on Decision Vagueness, GFCC



judges face the risk of noncompliance. This is because they know that due to their robust levels of public support, they can increase the pressure on the government. The corresponding first difference is statistically significant on the 90% level.

Using alternative dependent variable

Table C.5 shows the fractional logit regression results for the GFCC, using the alternative judicial policy implementation vagueness score of the German expanded dictionary as dependent variable. The results show that although all coefficients exhibit the correct directions, none of them are statistical significant. One reason for this could be that the dictionary produces a substantial amount of false positives when testing on the out-of-sample prediction in Chapter 3.4.4.

Using random sub-sampling validation

I apply repeated random sub-sampling validation to account for unobserved heterogeneity. I re-run my analyses twenty times using only a two-third subset of the data, collect the results and combine the estimates as outlined in King et al. (2001). I find that the direction and size of the estimates are robust across different subsets, but that the combined estimates are not statistically significant anymore, indicated by the large standard errors. However, this is a result of the rather small N of the German data ($N = 372$ decisions), so that the larger standard errors are a result of sampling variability, and therefore to be expected.

Table C.5 – Results of Fractional Logit Regression, Percentage of Vague Words as Dependent Variable, GFCC Analysis

	DV: Percentage Vague Words	
	Model 1	**Model 2**
Ideological Distance	−0.002	0.003
	(0.003)	(0.014)
Complex Case (= 1)	0.086	0.085
	(0.064)	(0.064)
Second Senate (= 1)	−0.095	−0.095
	(0.065)	(0.065)
Oral Hearing (=1)	−0.073	−0.077
	(0.077)	(0.077)
Risk Noncompliance (=1)	0.020	0.103
	(0.123)	(0.227)
Ideological Distance × Risk Non-Compliance		−0.006
		(0.014)
Constant	−5.555***	−5.628***
	(0.125)	(0.226)
N	372	372

***p < .01; **p < .05; *p < .1
Robust Standard Errors are used.

Table C.6 – Combined estimates from repeated random sub-sampling validation, GFCC

	DV: Percentage Vague Sentences	
	Model 1	**Model 2**
Ideological Distance	−0.028	−0.257
	(0.018)	(0.655)
Complex Case	−0.724	−0.665
	(0.529)	(0.528)
Second Senate (=1)	−0.653	−0.668
	(0.547)	(0.547)
Oral Hearing (=1)	−0.502	−0.453
	(0.495)	(0.500)
Risk Non-Compliance (= 1)	0.101	−0.721
	(0.977)	(1.235)
Ideological Distance X Risk Non-Compliance		0.232
		(0.654)
Constant	−5.503***	−4.748***
	(1.094)	(1.219)

***p < .01; **p < .05; *p < .1

Robustness Checks Conseil Constitutionnel Analysis

Alternative operationalization of judicial uncertainty using CAP coding

First, I re-run the main analyses but this time I use a dummy variable indicating whether a case is complex or not (using the same Comparative Agenda Project coding scheme than in the German analysis) instead of the count of the numbers of legal doctrines examined as a measure for judicial uncertainty. The results remain unchanged when using this alternative measure.

Table C.7 – Results of Fractional Logit Regression Using Case Complexity as Judicial Uncertainty Measure, CC

	DV: Percentage Vague Words
	Model 1
Ideological Distance	0.009***
	(0.002)
Complex Case (= 1)	0.145***
	(0.049)
Government Brief	0.057
	(0.052)
Intercept	−5.632***
	(0.052)
N	494

***p < .01; **p < .05; *p < .1
Robust standard errors in parentheses.

Using random sub-sampling validation

Second, I apply the same repeated random sub-sampling validation to account for unobserved heterogeneity than in the robustness analysis of the German data. I re-run my analyses twenty times using only a two-third subset of the data, collect the results and combine the estimates as outlined in King et al. (2001). My results remain unchanged: coefficient and the combined standard errors are similar to the ones the main analysis using the full data set.

Using alternative ideology score (MCSS)

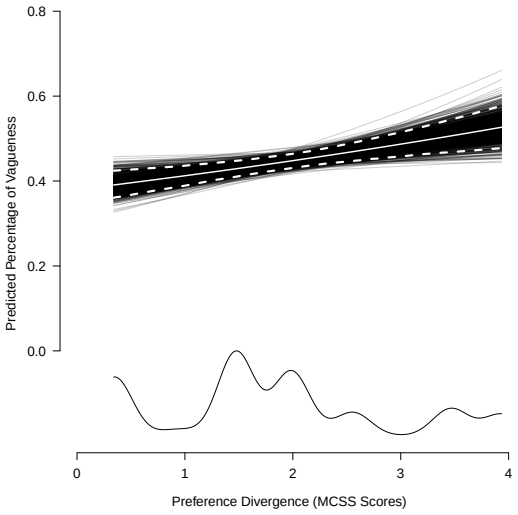
I replicated the main analysis but this time using the MCSS scores instead of the CMP scores. In Figure C.4 shows the corresponding simulations, which were obtained in the same way as in the main analysis (observed value approach using the QMLE of the fractional logit). We observe the same effect than in the main analysis: the higher the distance between Conseil and legislator, the higher the expected decision vagueness. This is again in contrast to the formal model's prediction. Because I obtain the same

Table C.8 – Combined estimates from repeated random sub-sampling validation, CC

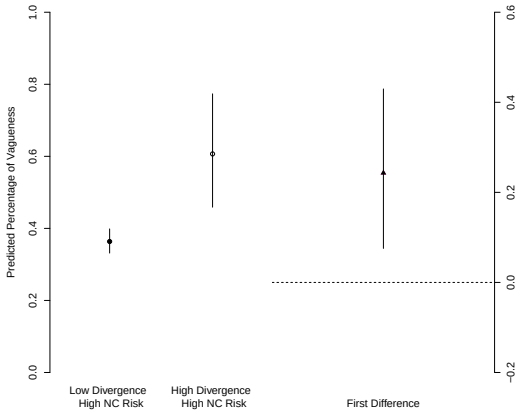
	DV: Percentage Vague Words	
	Model 1	**Model 2**
Ideological Distance	0.010***	0.009**
	(0.003)	(0.004)
Case Complexity (Number of legal issues)	0.023**	0.023**
	(0.010)	(0.010)
Risk Non-Compliance (= 1)	−0.022	−0.070
	(0.085)	(0.110)
Ideological Distance X Risk Non-Compliance		0.005
		(0.006)
Constant	−5.699***	−5.686***
	(0.088)	(0.092)

***p < .01; **p < .05; *p < .1

Figure C.4 – Effect of Ideological Distance (MCSS) on Decision Vagueness, CC



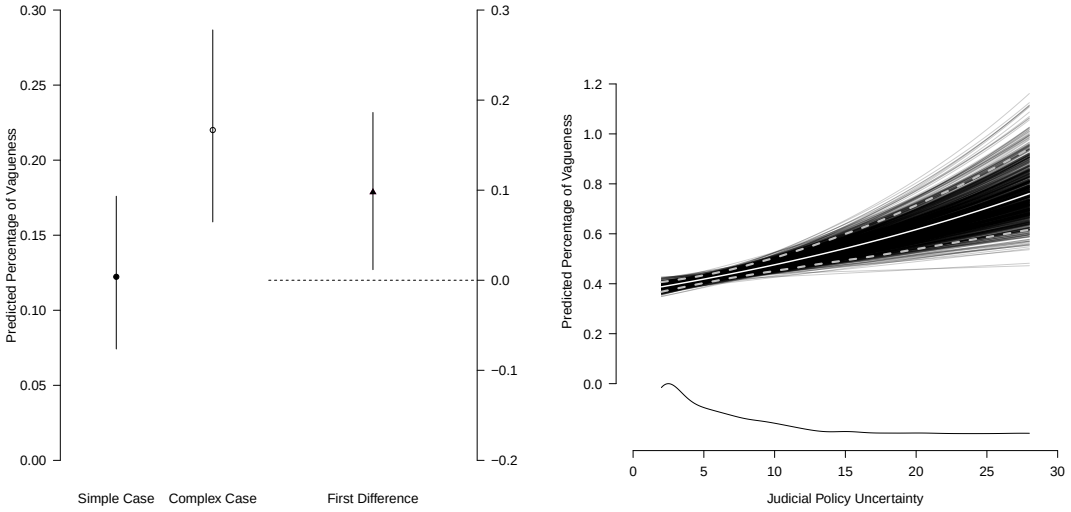
(a) Ideological Distance on Decision Vagueness



(b) First Difference Minimum Maximum Ideological Distance using MCSS

result using the MCSS scores and the CMP scores, I conclude that this contradictory finding is not due to measurement error of the independent variable.

Figure C.5 – Effect of Judicial Uncertainty on Decision Vagueness, GFCC and CC



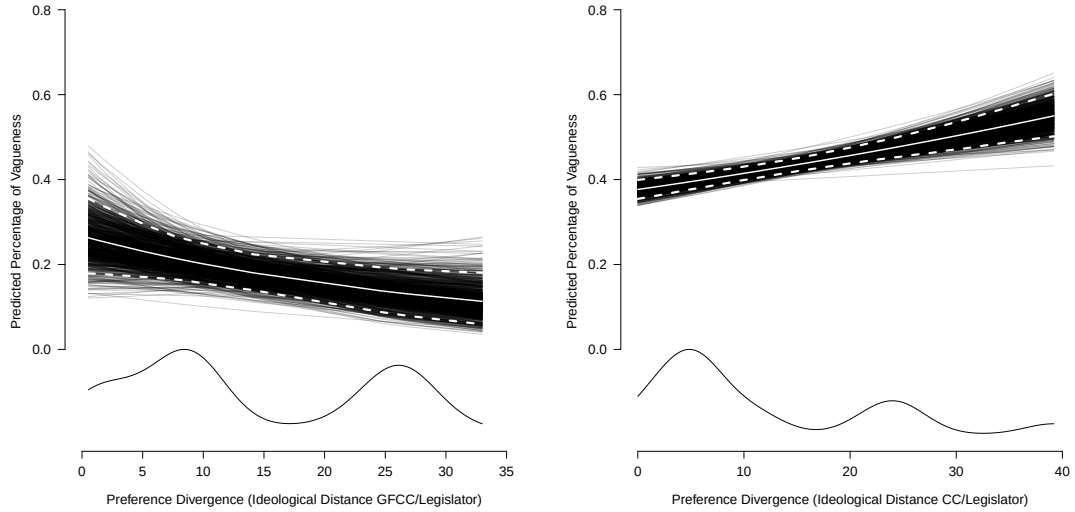
(a) GFCC analysis *(b) CC analysis*

Note: Left Side: Expected percentage of decision vagueness for simple and complex cases, with the corresponding first difference using simulations. For the simple and complex case scenario, the *complex case* dummy was set to zero and one, respectively. The points represent the point estimates and the bars represent 90% confidence intervals.
Right side: Expected percentage of decision vagueness over a range of judicial uncertainty, including 90% confidence intervals. Judicial uncertainty is measured by the number of legal doctrines considered in a decision.

C.5 Results of the Fractional Logistic Regressions Based on Bootstrapping

In this section, I replicate the main analysis (see Chapter 4.4.3) but instead of the robust variance estimator proposed by Papke and Wooldridge (1996), I use bootstrapping to obtain coefficients and standard errors. In bootstrapping, the sampling distribution of a parameter is approximated by repeatedly taking n samples *with replacement* from the original data. In my analysis, I use $n = 1,000$ where the size of each bootstrap sample is identical to the original data. The bootstrapped samples are then used to calculate means and standard errors. For the simulation, I use the observed value approach outlined in Chapter 4.4.3 and the same model specifications. The sampling distributions obtained via the bootstraps are directly used for the simulations. The following figures show that I obtain results similar to the main analysis. Overall, the uncertainty surrounding the estimates is a little smaller than in the main analysis. For instance in the simulation of the effect of ideological distance on the decision vagueness in Figure C.6, the lines representing one draw of the simulation in the "spaghetti" plot are a little closer to the estimated mean than in the main analysis. This demonstrates that even if no assumptions about the variance specification are made, the results remain robust.

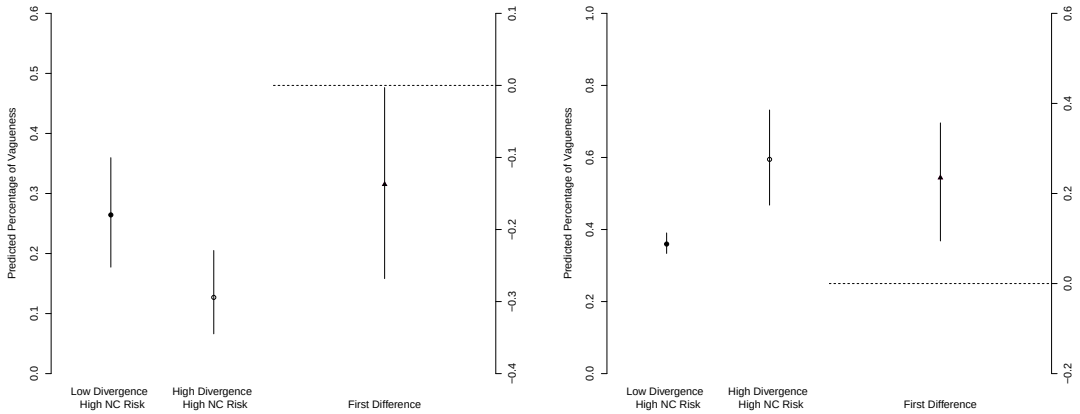
Figure C.6 – Effect of Preference Divergence on Decision Vagueness, GFCC and CC



(a) GFCC analysis *(b) CC analysis*

Note: Left Side: Expected percentage of decision vagueness over the range of preference divergence, measured by the ideological distance between court and legislator. 90% confidence intervals are used.
Right side: Expected percentage of decision vagueness over the range of preference divergence, measured by the ideological distance between court and legislator. 90% confidence intervals are used.

Figure C.7 – Conditional Effect of Preference Divergence and Non-Compliance Risk on Decision Vagueness, GFCC and CC



(a) GFCC analysis *(b) CC analysis*

Note: This figure shows the conditional effect of preference divergence and the risk of non-compliance (NC) on decision vagueness. The black points represent the expected decision vagueness for a case with a high perceived risk of non-compliance and low preference divergence. The white points represent the expected decision vagueness for a case with a low perceived risk of non-compliance and high preference divergence. The first difference between these two scenarios is displayed on the right of each graph. The bars represent 90% confidence intervals. For the scenarios, the risk of non-compliance variable is set to 1 and ideological distance is set to the minimum (low divergence) and maximum (high divergence).

APPENDIX D

Chapter 5

D.1 Outline of the Random Forest Algorithm

Algorithm outline of random forest, directly adopted from (Hastie, Tibshirani and Friedman, 2009, 558):

1. For b = 1 to B:

   (a) Draw a bootstrap sample **Z*** of size N from the training data.

   (b) Grow a random forest tree T_b to the bootstrap data, by repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables.

      ii. Pick the best variable/split-point among the m.

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

The final prediction of a new data point x is then in the classification case:

$$\hat{C}_{rf}^B(x) = majorityvote\{\hat{C}_b(x)\}_1^B$$

where $\hat{C}_b(x)$ is the class prediction of the bth random forest tree.

D.2 Comparison of predictive performance of different classifiers

Figure D.1 – Performance of different algorithms on the Constitutional Complaints Data, Combined Model

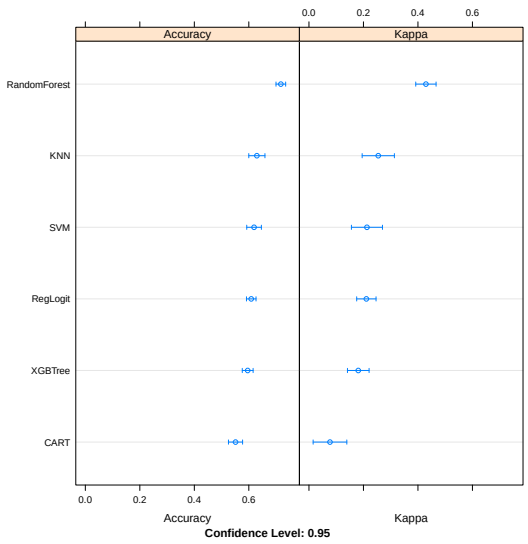


Figure D.1 shows the predictive performance of multiple machine learning algorithms using 10-fold cross-validation (without hyper-parameter tuning) and the constitutional complaints data set. Classification and Regression Trees (CART), extremely boosted trees (XGBTree), regularized regression, support vector machines (SVM), *k*-nearest neighbors and random forests. Accuracy and Kappa are reported. Confidence intervals are just for visualization purposes and are calculated using the standard error of the respective mean (across the 10-folds).

D.3 Additional Model Performance Metrics

Table D.1 – Model Evaluation Based on Aggregated Cross-Validation Scores, Additional Performance Metrics

	Accuracy			Kappa		ROC AUC		PR AUC	
	Legal	Combined	Baseline	Legal	Combined	Legal	Combined	Legal	Combined
Constitutional Complaints	60.14	**68.93**	53.47	0.20	**0.37**	0.66	**0.76**	0.70	**0.76**
Concrete Review	68.42	**80.18**	67.02	0.08	**0.50**	0.66	**0.83**	**0.85**	0.83
Abstract Review/Organstreit	63.54	**73.04**	60.26	0.19	**0.41**	0.68	**0.77**	0.73	**0.77**

Note: Model performances of the legal model and the combined model based on the aggregated 10-fold cross-validation scores. The random forests were build with a fixed m. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

Table D.2 – Model Evaluation Based on Out-of-Sample Prediction, Additional Performance Metrics

	Accuracy			Kappa		ROC AUC		PR AUC	
	Legal	Combined	Baseline	Legal	Combined	Legal	Combined	Legal	Combined
Constitutional Complaint	66.67	**74.49**	52.67	0.33	**0.49**	0.73	**0.83**	0.75	**0.83**
Concrete Reviews	75.26	**81.05**	65.79	0.41	**0.57**	0.75	**0.86**	0.65	**0.82**
Abstract Reviews/Organstreit	60.38	**77.36**	58.49	0.17	**0.52**	0.66	**0.79**	0.67	**0.82**

Note: Model performances of the legal model and the combined model based on out-of-sample prediction. The legal model only uses legal context variables, while the combined models used both legal and political context variables. The baseline category for accuracy is a naive classifier who always votes the majority category of the training set. The best performances are highlighted in bold.

D.4 Confusion Matrices of the different classifiers

Table 1: Constit. Complaints, Legal Model

Predicted/Reference	against	in favor
against	147	79
in favor	83	177

Table 2: Concrete Reviews, Legal Model

Predicted/Reference	against	in favor
against	110	32
in favor	15	33

Table 3: Abstract Reviews/Organstreit Proceedings, Legal Model

Predicted/Reference	against	in favor
against	23	12
in favor	8	10

Table 4: Constit. Complaints, Combined Model

Predicted/Reference	against	in favor
against	169	52
in favor	61	204

Table 5: Concrete Reviews, Combined Model

Predicted/Reference	against	in favor
against	111	22
in favor	14	43

Table 6: Abstract Reviews/Organstreit Proceedings, Combined Model

Predicted/Reference	against	in favor
against	27	8
in favor	4	14

D.5 Model Evaluation of Legal, Combined and Random Model based on Out-of-Sample Prediction

Table D.3 – Model Evaluation of Legal, Combined and Random Model based on out-of-sample prediction

	Accuracy			Kappa		
	Legal	Combined	Random	Legal	Combined	Random
Constitutional Complaints	64.81	**76.34**	63.58	0.29	**0.53**	0.27
Concrete Review	75.26	**81.05**	73.68	0.41	**0.57**	0.36
Abstract Reviews/Organstreit	69.93	**73.58**	72.73	0.39	**0.44**	0.42

Note: Model performances of the legal, combined and random model based on out-of-sample prediction. The best performances are highlighted in bold.

Table D.3 reports the model performance of the legal, the combined and the random model using the same out-of-sample data set than used in the main analysis. Again, the combined model performs best across all metrics. We can also see that again, although less stark than in the main analysis, the addition of noise features to the model improves the predictive performance compared to the legal model for abstract reviews and Organstreit proceedings.

D.6 Model evaluation based on out-of-sample prediction using the time dimension for splitting

Table D.4 – Model evaluation based on out-of-sample prediction using the time dimension for splitting

	Accuracy		Kappa		ROC AUC		PR AUC	
	Legal	Combined	Legal	Combined	Legal	Combined	Legal	Combined
BvR	59.33	**59.33**	**0.18**	0.13	**0.65**	0.62	0.74	**0.76**
BvL	75.26	**81.58**	0.41	**0.58**	0.74	**0.86**	0.64	**0.82**
BvE/BvF	51.51	**54.55**	-0.03	**0.06**	0.50	**0.41**	**0.50**	0.26

Table D.4 reports the performance measures for the legal and combined model using out-of-sample prediction. The test data was created by splitting the data set on each proceeding such that all observation after 2005 were assigned to the test set and all observations after where assigned to the training set. Note that this, however, results in unequal train/test splits, such that not all test sets contain the same percentage of observations. Again, the combined model achieves the best classification performance across most of the performance metrics.