Regulatory Compliance with Limited Enforceability: Evidence from Privacy Policies*

Bernhard Ganglmair† Julia Krämer‡ Jacopo Gambato§

July 8, 2024

Abstract

We study how asymmetric enforceability of regulatory rules affects firms' compliance using a simple inspection model with two target outcomes and a large sample of privacy policies from German firms. We empirically exploit the introduction of stringent transparency rules through the EU General Data Protection Regulation (GDPR) of 2018, compelling firms to disclose, in accessible language, details of how they collect, process, and use data. The specifics of disclosure are objective (and easy to verify), whereas readability is subjective and difficult to enforce. We use text-as-data techniques to construct measures of disclosure and readability and show that firms increased the disclosure volume, but the readability of their privacy policies did not improve. In line with our theoretical predictions, firms anticipating heightened regulatory scrutiny and those facing higher-budget data protection authorities demonstrated a stronger response in readability compliance without sizeable effects on disclosure.

Keywords: data protection, GDPR, information disclosure, privacy policies, regulation, text-as-data, transparency

JEL Codes: D22; K20; L51.

*The order of authors is randomized using the AEA's Author Randomization Tool. We thank Heski Bar-Isaac, Kirsten Bock, Guido Friebel, Alberto Galasso, Yangguang Huang, Wolfgang Kerber, Jan Krämer, Tesary Lin, Ryan Steed, Hannes Ullrich, Kathrine von Graevenitz, Martin Watzinger, and conference and seminar participants at HEC Paris, Santa Clara University, University of Antwerpen, University of Toronto (Rotman), AEA, EALE, EPCS, the European Commission's Annual Research Conference (2023), the Annual Meeting of German Economists Abroad, IIOC, MWZ Text-As-Data Workshop, NBER Privacy Conference, Royal Economic Society, and SIOE for useful comments and suggestions. We thank Daniel Erdsiek and Sandra Gottschalk for guiding us through the data sources at ZEW and Farshad Ravasan for sharing the data on enforcement activities at the UK Information Commissioner's Office. We also thank Jianming Cui, Natalia Garcia Soto, Pujit Golchha, Ana Rantes Lozano, and Lion Szlagowski for excellent research assistance. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Projects B02 and B04) is gratefully acknowledged. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

†Corresponding author: University of Mannheim and ZEW Mannheim, b.ganglmair@gmail.com

‡Erasmus University Rotterdam, j.k.kramer@law.eur.nl

§University of Mannheim and ZEW Mannheim, ja.gambato@gmail.com

1 Introduction

In 2018, the General Data Protection Regulation went into effect; it transformed the digital landscape in Europe and beyond, often to the detriment of firms but with some privacy improvement on the consumer side.[1] A central contribution of the GDPR was its *transparency principle* that compels firms to disclose information about the nature of their data collection, processing, and use (Art. 13–14 GDPR) in accessible and readable language (Art. 12(1) GDPR). However well-intended, the rules come with major enforceability concerns. While the disclosure requirement is based on an objective list of items to be disclosed, the readability requirement is vague and subjective; data protection authorities are left to interpret these rules as they lack enforcement experience and established precedents or best practices. Because enforcement is costly, data protection authorities will prioritize and eventually give more weight to the disclosure requirement in their enforcement activities. However, if firms anticipate limited enforcement in one dimension, compliance will suffer. This paper asks how the asymmetric (and limited) enforceability of the GDPR's transparency principle affects firms' compliance decisions. We also explore in greater detail the role of regulatory scrutiny (when firms anticipate they are a regulator's primary target) and regulatory capacity (e.g., Stern, 2000; Armstrong and Sappington, 2006) in the strategic interaction of enforcement and compliance.

We first propose a theoretical model to address these questions. In our framework, a firm can choose costly compliance with some requirements when drafting a privacy policy, and a regulator can audit the privacy policy to confirm compliance. Our framework is closest to that of Heyes (1994), which models the thoroughness of inspection of a single requirement as an endogenous choice. We deviate from this approach by taking the probabilities of success of an audit in detecting non-compliance with multiple requirements as given (Macho-Stadler and Perez-Castrillo, 2006), and instead focus on the firm's and regulator's choices of which of the requirements to comply with and audit. We assume asymmetric enforcement success, with a higher probability of detecting non-compliance for the disclosure requirement and a lower probability for the readability requirement. We derive equilibrium outcomes for constrained regulators (that can enforce only one of the requirements) and unconstrained regulators (that can enforce one or both requirements).

This model of asymmetric enforceability predicts better disclosure compliance than readability compliance as a firm's response to the GDPR (Prediction 1). Moreover, when a firm expects to be a more prominent target for the regulator and thus anticipates stricter

[1] See Johnson (forthcoming) for a comprehensive review of the economics literature studying the effect of the GDPR.

enforcement (*regulatory scrutiny*), it will showcase better readability compliance than disclosure compliance (Prediction 2). Intuitively, such a firm already exhibits extensive disclosure, and higher regulatory scrutiny induces more effort in the previously understated readability dimension. Last, relaxing the regulator's budget constraint (and allowing for broader enforcement through increased *regulatory capacity*) again triggers a catching-up effect. A firm facing an unconstrained regulator will comply more with the readability requirement than a firm subject to a constrained regulator (Prediction 3). A firm with high compliance costs will reduce its disclosure compliance in response.

To test the predictions from our model, we construct a quarterly (unbalanced) panel of privacy policies posted by German firms between 2014 and 2021. The final dataset contains more than 585,000 policies posted by more than 75,000 firms. For Prediction 1, we conduct a simple before-and-after analysis. For Prediction 2, we use the data protection enforcement history of the UK Information Commissioner's Office (Koutroumpis et al., 2022), the prevalence of the use of personal data in the German information economy (ZEW, 2017), and market concentration (4-digit industry Herfindahl-Hirschman Index using firm-level information from the Mannheim Enterprise Panel (Bersch et al., 2014)) as proxies for regulatory scrutiny. For Prediction 3, we leverage Germany's decentralized enforcement of the GDPR by 16 state data protection authorities regulating firms in their respective states. We exploit the variation of the authorities' budget across states and over time, assuming that higher-budget data protection authorities are less likely to be budget-constrained.[2] We collect budget information for the 16 German state data protection authorities and use state-level firm population numbers (Bersch et al., 2014) to calculate per-firm budget variables.

For our outcome variables of a firm's compliance with the transparency requirements, we construct metrics for disclosure and readability. We estimate LDA topic models (Blei et al., 2003) and count a policy's *distinct topics* to capture the breadth of a policy's content. We also use these results to identify paragraphs mentioning terms indicative of the disclosure of information required by Art. 13–14 GDPR (i.e., disclosing paragraphs). To construct our primary disclosure measure (*topic-weighted word count*), we use the relative distribution of topics of these disclosing paragraphs as weights of a policy's paragraphs' word count.

For readability measurements, we borrow from the toolkit of linguists, who have constructed many indices and scores to measure the readability of texts. We use two scores. First, as best-practice approach, we construct the German version of the Flesch Reading Ease (German FRE) score (Flesch, 1948; Amstad, 1978) that has been used in the U.S. to

[2]Enforcement is costly for regulators, and data protection authorities (in Germany and across Europe) vary in financial resources (the result of political decisions by the respective legislatures). Such variations are likely contributing to differences in authorities' strictness (e.g., survey evidence suggests the strictest regulators are found in Germany and Sweden (see Johnson, forthcoming)).

regulate the readability of insurance contracts. Second, we take a data-driven approach. We compile a set of roughly 4,000 human-coded comparisons of the readability of short snippets of text from our sample of privacy policies. Using the methodology laid out in Benoit et al. (2019), we then identify the *läsbarhetsindex* (LIW) (Björnson, 1968) as the readability index that best explains our comparison data.

We find that, in response to the GDPR, firms increase the amount of disclosure in their privacy policies by 50% (topics) to 80% (topic-weighted words). This is strong evidence for disclosure compliance, whereas the results for readability compliance are weak and mixed (perfectly in line with Prediction 1 in our theoretical framework). The GDPR response for the German FRE score indicates a decline in readability and, for the LIW, an increase in readability. Both are an order of magnitude smaller than the effects on disclosure. We further explore the impact of the GDPR on disclosure and readability as a function of a firm's exposure to the GDPR (or the treatment intensity). We find that firms with low pre-GDPR compliance (that were therefore more exposed to the GDPR's requirements) exhibited more robust increases in disclosure and readability.

Our model predicts a stronger response in readability than in disclosure for firms that expect to see more attention from regulators (Prediction 2). Our empirical results are in line with this prediction. Firms in industries with prior enforcement history improve the readability more (or lower the readability less) than firms in industries without. At the same time, the number of topics increases more, but the volume of disclosure increases less (both at much smaller relative magnitudes than the effects on readability). We observe the same patterns for readability in industries with a higher prevalence of the use of personal data and higher market concentration (both are presumably primary regulation targets). Last, we find a positive effect on the firms' disclosure response to the GDPR for firms in industries with a higher prevalence of personal data. This result is likely because when the use of personal data is important, firms will have additional content to disclose in their policies.

Our regulator budget results are in line with Prediction 3. We find that, as predicted, firms in higher-budget states do not exhibit different levels of disclosure compliance. If anything, these firms' disclosure compliance declines (a result we obtain for firms with high compliance costs). The results on readability compliance are strongest for the German FRE. Firms in higher-budget states see a smaller decline in readability than firms with more constrained data protection authorities. We see similar, albeit statistically weaker, patterns in support of Prediction 3 for the LIW.

Related Literature Our study contributes to various strands of the literature in economics. The nature and characteristics of regulatory environments have been the subject

of a line of studies in the regulation economics literature. Systematic limitations to regulation generally relate to information asymmetries between the regulator and the regulated industry (Laffont, 1994) and limited regulatory resources (Stern, 2000; Armstrong and Sappington, 2006).[3] We focus our attention on a different kind of impediment to regulation, namely the limited enforceability of uncertain or vague requirements. Uncertainty of regulatory requirements can interact with a regulator's budget constraint, mainly when regulation is multi-dimensional (with several requirements that must be met), and limited resources reduce the ability of a regulator to enforce them all. Our empirical results show that the effect of relaxing a regulator's budget constraint is stronger for the requirement that comes with a higher level of vagueness (and lower verifiability).

Our theoretical framework builds on the game-theoretical literature on audits and tax avoidance (Greenberg, 1984; Fellingham and Newman, 1985; Graetz et al., 1986), which builds on seminal work by Dresher (1962) who first formulated *inspection games*. We include a novel dimension to the strategy of the regulator: We focus on the optimal regulator strategy with regard to what she should audit when agents are compelled to comply with multiple requirements. Unlike earlier work, we model imperfect regulation (Heyes, 1994; Bardsley, 1996; Macho-Stadler and Perez-Castrillo, 2006)[4] and assume exogenous success probabilities (as in Macho-Stadler and Perez-Castrillo, 2006). We add to this literature by studying the enforcement and compliance of multiple requirements that must be audited separately with different (and independent) success probabilities.

A growing number of studies examine the effects of the GDPR on firm behavior and performance. Examples are Yuan and Li (2019) (a sharp decline in financial performance for hospitals that attach importance to digital health services), Goldberg et al. (2024) (a drop in page views and revenue for online firms), or Johnson et al. (2023) and Peukert et al. (2022) (examining the effects of the GDPR on firms' use of and interaction with web technology vendors). Koski and Valmari (2020) find that small and medium-sized enterprises in data-intensive industries are affected the most by the GDPR, arguing that economies of scale may result in different economic effects of the GDPR when adhering to its provisions. We add to this literature by providing a nuanced picture of the effectiveness of the GDPR and highlighting that compliance with the new regulation is not a given but rather the result of firms' anticipation of strategic enforcement decisions by constrained regulators.

The law and economics literature studying privacy policies has seen a sharp increase in attention with the introduction of the GDPR. While earlier work uses small samples of

[3]These limitations are typically (but not exclusively) studied in the context of developing countries where the premises for perfect enforcement are not generally met (Stern, 2000; Laffont, 2005).

[4]We also assume audits (or regulation inspections) are error-free, which means, they do not produce false negatives by mistaking compliance for non-compliance.

privacy policies (e.g., Jensen and Potts, 2004; Milne et al., 2006), more recent studies compile large datasets of privacy policies for thousands of firms (Frankenreiter, 2022; Amos et al., 2021; Wagner, 2023). Both small and large-scale studies have found a downward trend in the readability of privacy policies (e.g., Milne et al., 2006; Amos et al., 2021), with some recent results also hinting at no-changes (Linden et al., 2020) or slight improvements (Becher and Benoliel, 2021) post-GDPR. Moreover, studies show that post-GDPR, privacy policies are significantly longer and show greater detail (Degeling et al., 2019; Linden et al., 2020). We can match privacy policies to firm and industry-level data and thus paint a more detailed picture of trends in disclosure and readability. Additionally, our approach provides evidence for limited enforceability (and a resulting lack of enforcement) as a potential explanation for the failure of the GDPR to provide more readable and transparent privacy policies (e.g., European Commission, 2019).

Our results also relate to the literature on contractual terms of use, "fine print," and boilerplate (or standardized) contract language. Bakos et al. (2014) show overwhelming evidence supporting the notion that users rarely even skim through the fine print of contracts and terms of use online. Given this lack of attention by consumers, it is sensible to ask whether firms display more or less predatory contractual terms in response to the clients' disregard for the content of contracts. Marotta-Wurgler (2007) studies software end-user license agreements and shows a striking heterogeneity and a negative correlation between firm revenue and pro-consumer bias in these contracts' terms. Drawing a parallel between readability and the author's definition of "friendliness" of contract terms, our findings are well in line with hers. In a separate article (Marotta-Wurgler, 2008), however, the author finds no correlation between bias in contract terms and firm-relevant market concentration measures. In contrast, we highlight a positive relationship between the two: firms active in more concentrated markets tend to draft more readable (i.e., user-friendly) policies.

Last, our methodological approach relates our study to a growing literature that uses text-as-data methods.[5] A central method in our paper is the estimation of topic models. These models have been used on a number of different types of document corpora,[6] and we add privacy policies to this ever-growing list.

The remainder of this paper is structured as follows. In Section 2, we introduce our theoretical framework and derive predictions for the empirical analysis. In Section 3, we

[5]For a comprehensive survey of this growing literature, see Loughran and McDonald (2016) (in finance and accounting) or Gentzkow et al. (2019) and Ash and Hansen (2023) (in the social sciences).

[6]For example, emails (McCallum et al., 2007), scientific abstracts and articles (Blei et al., 2003; Griffiths and Steyvers, 2004), newspaper archives (Larsen and Thorsrud, 2019), U.S. Supreme Court decisions (Livermore et al., 2017), patents (Ruckman and McCarthy, 2017), loan agreements (Ganglmair and Wardlaw, 2017), or analyst reports (Bellstam et al., 2021).

describe the construction of our estimation sample and introduce our text-based measures of disclosure and readability. In Section 4, we document how firms have responded to the introduction of the GDPR using simple before-and-after analyses. In Section 5, we explore the role of regulatory exposure (through a treatment-intensity design), scrutiny, and capacity. We conclude in Section 6. All proofs are relegated to the Appendix.

2 A Model of Compliance

We propose a simple model to capture a firm's decision to comply with the requirements of the GDPR. In our inspection game, a firm can choose costly compliance with some requirements when drafting a privacy policy, and a regulator can audit the privacy policy to confirm compliance. We focus on the role that the expected level of received scrutiny plays in the way firms draft their policies. Furthermore, we highlight the role of a regulator's budget constraint. Through the model, we make several predictions that we apply to the data.

2.1 Framework

A firm (she) is tasked with drafting a privacy policy, subject to two requirements. First, the policy must provide the right type and amount of information ("disclosure"); second, the policy must be accessible to consumers ("readability"). Compliance with these requirements is costly for the firm. A regulator (he) is tasked with enforcing the disclosure and readability requirements. He audits policies to assess their compliance. The regulator can choose the intensity of this audit and inspect the policy for either, neither, or both requirements. These audits are imperfect: The regulator learns, with positive probability, whether the policy complies with either requirement. If he finds non-compliance, he challenges the policy, resulting in a penalty for the firm.

We model the interaction between the firm and the regulator as a simultaneous-move game in which the firm chooses how many and which requirements to comply with, and the regulator chooses the intensity of the audit (that is, which requirements, if any, to inspect). The firm's stand-alone value from an unchallenged policy is $v > 0$. Compliance with each requirement $j \in \{d, r\}$ (for *disclosure* and *readability*) comes at a fixed cost k per requirement. Formally, the firm selects $(d, r) \in \{0, 1\} \times \{0, 1\}$ and generates an unchallenged value $v - kd - kr$. If the regulator audits the policy and finds non-compliance, the firm's payoffs are zero.[7] We assume that compliance never leads to negative payoffs for the firm:

[7] A challenged policy always generates zero utility for the firm, regardless of its choice of (d, r). This implies that the fee paid by the firm for non-compliance is different if she does not comply with either or both requirements. The assumption allows for immediate comparison of all possible outcomes. A flat fee

Assumption 1. $0 < k < \frac{v}{2}$

For an audit, the regulator chooses to inspect either, neither, or both requirements. When he finds non-compliance in either requirement, he challenges the policy. Both audit and challenge are without cost. We consider two types of regulators: *unconstrained* and *constrained*. An unconstrained regulator has sufficient resources to inspect both requirements. A constrained regulator has limited resources and can inspect at most one requirement.[8]

Non-compliant policies generate a social loss of $-\gamma < 0$.[9] The regulator's objective is to minimize this social loss, and he audits policies to detect non-compliance, subject to his constraint type. His payoffs from an unchallenged non-compliant policy are $-\gamma$; the payoffs from a challenged or compliant policy are zero.

Audits are imperfect, and the inspection of a policy for requirement j leads to the discovery of its state (either $d, r = 1$ or $d, r = 0$) with probability π_j. We assume, per our earlier discussion, that inspecting disclosure d has a higher chance of discovering the true state of the policy: disclosure of specific information items is objective, whereas readability is subjective. We further introduce a lower bound for the regulator's success probabilities:[10]

Assumption 2. $\frac{1}{2} < \pi_r < \pi_d < 1$

The strategies of the players are as follows: The firm chooses between full non-compliance $(d = 0, r = 0)$, non-compliance in readability $(d = 1, r = 0)$, non-compliance in disclosure $(d = 0, r = 1)$, and full compliance $(d = 1, r = 1)$. To ease notation, we refer to the strategic decision $j = 1$ as j, and $j = 0$ as 0, for $j \in \{d, r\}$.

An unconstrained regulator chooses between no inspection (a_0), inspection of the disclosure requirement (a_d), inspection of the readability requirement (a_r), and full inspection (of both requirements) $(a_{d,r})$. A constrained regulator has only single inspection choices and cannot choose full inspection. Table 1 summarizes the players' strategies and corresponding outcomes. The first value in each cell represents the firm's payoffs; the second value represents the regulator's payoffs.

would not affect the compliance incentives beyond a numerical difference. We model the payoffs this way to highlight the interaction between the firm's and regulator's choices rather than produce direct numerical estimates.

[8]This regulator-type distinction is a reduced-form characterization of a regulator's budget constraint. Suppose inspecting a given requirement comes at a cost, say, c. Then, an unconstrained regulator has sufficient resources to incur costs of $2c$, whereas a constrained regulator can afford only audit costs of c.

[9]We assume a non-compliant policy generates the same social loss $-\gamma$ for any form of non-compliance. Thus, the regulator does not value one requirement more than the other and is ambivalent toward either requirement. This assumption is for simplicity and not an assessment of the social loss from lack of disclosure relative to lack of readability.

[10]This second assumption ensures that, for all feasible values k (Assumption 1), the firm does not always strictly prefer not to comply with either requirement.

Table 1: Normal-Form Representation of the Compliance-Enforcement Game

Firm's strategy	Regulator's strategy: a_0	a_d	a_r	$a_{d,r}$
$(0,0)$	$\begin{pmatrix} v \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)v \\ -(1-\pi_d)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)v \\ -(1-\pi_r)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(1-\pi_r)v \\ -(1-\pi_d)(1-\pi_r)\gamma \end{pmatrix}$
$(d,0)$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)(v-k) \\ -(1-\pi_r)\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_r)(v-k) \\ -(1-\pi_r)\gamma \end{pmatrix}$
$(0,r)$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(v-k) \\ -(1-\pi_d)\gamma \end{pmatrix}$	$\begin{pmatrix} v-k \\ -\gamma \end{pmatrix}$	$\begin{pmatrix} (1-\pi_d)(v-k) \\ -(1-\pi_d)\gamma \end{pmatrix}$
(d,r)	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$	$\begin{pmatrix} v-2k \\ 0 \end{pmatrix}$

2.2 Equilibrium

We first derive the Nash equilibrium for the unconstrained regulator (Proposition 1) and then proceed to the constrained regulator (Proposition 2). Last, we compare regulatory environments with unconstrained relative to constrained regulators (Proposition 3).

2.2.1 Unconstrained Regulator

In Table 1, we can see that if the regulator is unconstrained, he chooses to inspect both requirements since $a_{d,r}$ is a dominant strategy. Given this dominant choice by the regulator, the firm always prefers $(d,0)$ to $(0,r)$. If she decides to comply with only one requirement, she optimally chooses to comply with what is easier to detect (e.g., disclosure). Moreover, by the lower bound of the regulator's success probability in Assumption 2, the firm either chooses full compliance (d,r) or non-compliance in readability $(d,0)$.

Proposition 1 (Unconstrained Regulator)**.** *Suppose the regulator is unconstrained and can inspect both requirements. Let $k^u = \frac{\pi_r}{1+\pi_r}v$. For low compliance costs with $k < k^u$, the equilibrium is $(a_{d,r},(d,r))$. For high compliance costs with $k \geq k^u$, the equilibrium is $(a_{d,r},(d,0))$. In both cases, the regulator inspects both requirements, and the firm always complies with disclosure. The firm also complies with readability when enforcement costs are low.*

The extent to which the firm complies with *both* d and r depends on the compliance cost k. Because $\pi_d > \pi_r$, it is always better for the firm to comply with the disclosure requirement than with the readability requirement. It is also strictly better than not complying at all under the assumption that compliance never leads to negative payoffs for the

firm. Furthermore, if k is small enough, full compliance is cheap, and the benefits (by avoiding a challenge by the regulator) outweigh the costs. Conversely, if compliance costs k are high, the firm prefers not to comply with r, hoping that her non-compliance goes undetected. The threshold value at which the firm is indifferent between these options, k^u, is increasing in π_r. All else equal, an increase in π_r implies that the firm strictly prefers to comply with both requirements for more values of k. At the limit, $\pi_r = 1$ leads to $k^u = \frac{v}{2}$, and the firm always complies with both requirements when audits are perfect ($k < \frac{v}{2}$ by Assumption 1).

2.2.2 Constrained Regulator

For a constrained regulator, a full audit with $a_{d,r}$ is not feasible. However, given cost-free audits, an audit with *some* inspection dominates no audit. As a consequence, the regulator selects either a_d or a_r, possibly using a mixed strategy.

It is straightforward to see that a pure-strategy equilibrium does not exist. Suppose the regulator selects to inspect disclosure, a_d, with probability one. The firm's best response is to choose $(d, 0)$, that is, comply with respect to disclosure and ignore the readability requirement. The regulator is then unable to challenge the non-complying firm and would want to deviate, choosing a_r instead to be able to challenge the policy (that is not readability compliant). And so forth. In equilibrium, the regulator will always play a mixed strategy, choosing a_d and a_r with strictly positive probabilities.

The firm does not want to comply with both requirements if she can avoid it. To find the respective equilibria, we proceed as follows: First, we obtain the firm's mixed strategies with probabilities of playing $(d, 0)$ (denoted by p_d), $(0, r)$ (denoted by p_r), and $(0, 0)$ (probability $1 - p_d - p_r$). Second, we derive the regulator's mixed strategy that makes the firm indifferent between playing two of these strategies and for which parameters the firm is better off not deviating from the resulting mix.

In any mixed-strategy equilibrium, each player randomizes over some actions to make the other player indifferent between their selected strategies. We first find that the probabilities p_d and p_r, which make the regulator indifferent between a_d and a_r. These probabilities are:

$$p_r \in \left[0, \frac{\pi_r}{\pi_d + \pi_r}\right]; \tag{1}$$

$$p_d = \frac{\pi_d - (1 - p_r)\,\pi_r}{\pi_d}; \tag{2}$$

$$1 - p_d - p_r = (1 - p_r)\,\frac{\pi_r}{\pi_d} - p_r. \tag{3}$$

Note that p_d is always positive so that the firm satisfies the disclosure requirement with

strictly positive probability. With the complementary probability, the firm plays either $(0,0)$ (satisfying the readability requirement with zero probability), $(0,r)$ (always satisfying one or the other requirement), or both, if she is indifferent between $(0,0)$ and $(0,r)$.

The regulator has only two non-dominated strategies, a_d and a_r. Because the firm always plays $(d,0)$ with positive probability, we consider next the mixed strategy the regulator adopts to render the firm indifferent between $(d,0)$ and either $(0,r)$ or $(0,0)$. We use $p^r_{a_d}$ and $p^0_{a_d}$ to denote the probabilities of playing (a_d) that make the firm indifferent between $(d,0)$ and $(0,r)$ or $(0,0)$, respectively:

$$p^r_{a_d} = \frac{\pi_r}{\pi_d + \pi_r}; \tag{4}$$

$$p^0_{a_d} = \frac{(1-\pi_r)\,k}{\pi_d v - \pi_r k}. \tag{5}$$

With these probabilities, we characterize all possible equilibria. First, given the regulator's strategy $p^0_{a_d}$ or $p^r_{a_d}$, we find the expected payoffs of the firm playing $(d,0)$ and $(0,r)$ or $(0,0)$, respectively. Then, it suffices to identify the parametric values such that the other strategies are dominated, given the regulator's strategy.

Proposition 2 (Constrained Regulator)**.** *Suppose the regulator is constrained and can inspect a policy for only one requirement. Let*

$$\underline{k} := \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v < \frac{\pi_r \pi_d}{\pi_r + \pi_d - \pi_r \pi_d} v =: \overline{k}.$$

Then the following equilibria exist:

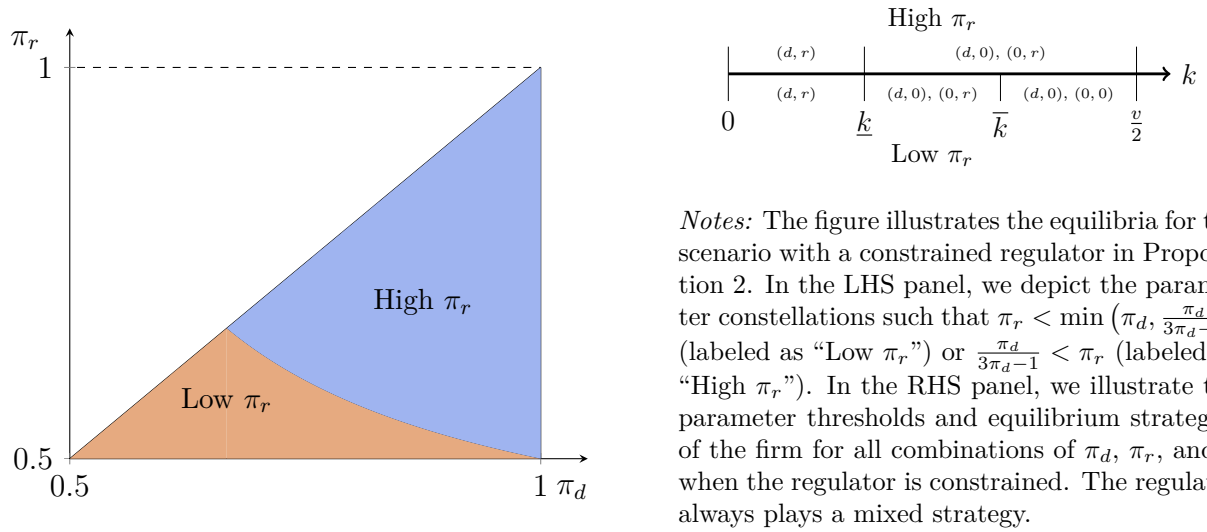
1. *if $\frac{1}{2} < \pi_r < \pi_d < 1$ and $0 < k < \underline{k}$, there is a continuum of payoff-equivalent equilibria in which the regulator mixes between a_d and a_r and the firm complies with both requirements with probability one;*

2. *if $\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right) \leq \frac{2}{3}$ and $\underline{k} < k < \overline{k}$, in the unique equilibrium the regulator mixes following $p^r_{a_d}$, and the firm complies with either the content or the readability requirement (but not both) according to $p_d = \frac{\pi_d}{\pi_d + \pi_r}$;*

3. *if $\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right) \leq \frac{2}{3}$ and $\overline{k} < k < \frac{v}{2}$, in the unique equilibrium the regulator mixes following $p^0_{a_d}$, and the firm either complies with the content requirement or does not comply with either of the requirements according to $p_d = 1 - \frac{\pi_d}{\pi_r}$;*

4. *if $\frac{\pi_d}{3\pi_d - 1} < \pi_r < \pi_d < 1$ and $\underline{k} < k < \frac{v}{2}$, in the unique equilibrium the regulator mixes following $p^r_{a_d}$, and the firm complies with either the content or the readability requirement according to $p_d = \frac{\pi_d}{\pi_d + \pi_r}$.*

Figure 1: Equilibria for Constrained Regulator



Notes: The figure illustrates the equilibria for the scenario with a constrained regulator in Proposition 2. In the LHS panel, we depict the parameter constellations such that $\pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right)$ (labeled as "Low π_r") or $\frac{\pi_d}{3\pi_d - 1} < \pi_r$ (labeled as "High π_r"). In the RHS panel, we illustrate the parameter thresholds and equilibrium strategies of the firm for all combinations of π_d, π_r, and k when the regulator is constrained. The regulator always plays a mixed strategy.

Figure 1 illustrates the parameter space and the respective equilibria. Proposition 2 states that depending on k, the game either has a unique equilibrium for $k \geq \underline{k}$ or a continuum of payoff-equivalent equilibria for $k < \underline{k}$. This multiplicity arises because the cheaper it is to comply with the requirements, the easier it is for the regulator to induce the firm to play (d, r).

The latter collection of equilibria leads to the same outcome: The regulator inspects either the disclosure or the readability requirement with positive probability in a way that makes the firm strictly better off complying with both than deviating to any other strategy. The easier it is to comply (lower k), the wider the range of mixed strategies that lead to this outcome. When $k = \underline{k}$, strategy $p^r_{a,d} = \frac{\pi_r}{\pi_d + \pi_r}$ makes the firm play (d, r) with probability one and generates the only mixed-strategy equilibrium. In contrast, as k approaches 0, any $p^r_{a,d} \in (0, 1)$ leads to the same outcome.[11]

2.3 Varying the Regulator's Resources

Propositions 1 and 2 reflect the ability of the regulator to induce compliance through inspection of the transparency requirements. Recall that k^u denotes the threshold below which an unconstrained regulator can induce full compliance from the firm. This value satisfies:

$$0 < \underline{k} < k^u < \overline{k} < \frac{v}{2}. \tag{6}$$

[11] Proposition 2 accounts for all equilibria except for knife edge scenarios in which the agents are indifferent between two of the above equilibria. We show that no mixed-strategy equilibria except those characterized in Proposition 2 exist in Appendix A.

The above ordering reveals that an unconstrained regulator can induce full compliance for a more extensive set of values k. This is because if $\underline{k} < k < k^u$, a constrained regulator is unable to induce full compliance no matter the relative value of π_r and π_d (see Proposition 2). Similarly, equilibrium compliance is lower for very high levels of compliance costs.

Proposition 3. *Suppose a tightening of the regulator's budget that renders a once unconstrained regulator constrained. With a now constrained regulator, compliance with the disclosure and readability requirements is weakly lower in equilibrium and strictly lower if* $\underline{k} < k < k^u$ *or* $\overline{k} < k < \frac{v}{2}$. *If* $k > k^u$, *the firm complies with at most one requirement: only* d *if the regulator is unconstrained, either* d *or* r *(or neither) if the regulator is constrained.*

In equilibrium, the firm never focuses on readability more than on disclosure. If the regulator can inspect both requirements, the firm either complies fully or ignores readability, depending on how costly compliance is. If the regulator is constrained, instead, the firm fully complies only if it is very cheap to do so (low k). Otherwise, she invests in disclosure or readability and focuses relatively more on disclosure if costs of compliance are intermediate and readability is not easily enforced (that is if π_r is relatively low). If compliance costs are very high and readability is not easily enforced, the firm either focuses on disclosure or chooses not to comply with either requirement. If readability is easily enforced (that is, if π_r is relatively high), the firm once again invests in disclosure or readability and relatively more on disclosure.

2.4 Discussion of Results

Our model yields several empirical predictions. First, more stringent regulation (i.e., regulation of the disclosure and readability requirements) weakly encourages compliance. To see why, suppose that the regulator in our framework cannot audit any requirements.[12] The firm will only ever choose not to comply with requirements that cannot be audited. In this sense, more stringent regulation being introduced will have a positive effect on compliance with both disclosure and readability if regulators act optimally in their role of auditors and a positive effect on disclosure alone if they do not.

Prediction 1. *Privacy policies become longer and, to a lesser extent, more readable after the introduction of more stringent regulation on disclosure and readability requirements for privacy policies.*

The model highlights the three factors that govern which of the arising equilibria we ought to expect: cost of compliance k, enforceability π_j, and the regulator's budget con-

[12] In terms of the model, this regulator could only play action a_0.

straint. While firms generally tend to comply with disclosure in equilibrium, these three factors determine the environments in which we will see relatively more compliance with the readability requirement.

The cost of compliance and enforceability are closely related. On the one hand, for low costs of compliance, the predicted level of compliance for all π_j and levels of the constraint of the regulator is higher. On the other hand, the higher the perceived risk of scrutiny by regulators (that is, the higher π_j is, for all $j \in \{d, r\}$), the more we expect compliance to arise. We expect firms that anticipate more thorough regulatory scrutiny or for which the drafting of legal documents is relatively cheap to be more likely to comply with both requirements.

Prediction 2. *With more stringent regulation, firms expecting stricter regulatory scrutiny, or operating in markets subject to stricter scrutiny, draft more readable privacy policies compared to firms that do not.*

We expect firms to react to the threat of more stringent scrutiny with better compliance. In the model, this stringency comes from the likelihood of non-compliance being detected (that is, π_j, conditional on requirement j being inspected). It is not, however, ex ante obvious which firms should expect more stringent scrutiny. We provide a detailed discussion of our empirical measures for regulatory scrutiny in Section 5.

The model also shows that the budget constraint of regulators plays an important role in their ability to incentivize compliance. Intuitively, the more resources are available to a regulator, the more thorough he can be in his audits. This thoroughness translates to different effects for the two requirements when interacting with the cost of compliance and expected scrutiny. In general, however, a firm facing scrutiny by an unconstrained regulator always complies more with both requirements than a firm facing a constrained regulator:

Prediction 3. *With more stringent regulation, firms operating in jurisdictions with less budget-constrained regulators draft more readable privacy policies than firms operating in the jurisdiction of more constrained regulators. The same firms also draft longer policies, but the effect on disclosure is smaller compared to the effect on readability.*

Finally, our equilibrium analysis reveals that a regulator facing budget constraints should focus more on readability than on disclosure in its audits. This is immediate from the equilibrium results of Proposition 2. Suppose, however, that a constrained regulator were not to inspect readability at all, perhaps because of how difficult it is to properly evaluate it. Anticipating this, firms would optimally disregard the readability requirement and comply only with disclosure. Because an unconstrained regulator would always inspect both requirements, increasing the budget of the regulatory agency would lead to a higher level of

compliance with the readability requirement. However, because firms with relatively high costs of compliance never comply with both requirements when facing an unconstrained regulator, the level of compliance with disclosure might actually decrease when the regulator has a larger budget at his disposal. While this follows from off-equilibrium (and, therefore, sub-optimal) behavior, it is worth highlighting because of the strategic considerations behind firms' drafting of privacy policies and because of the inherent limitations of enforcing the unclear legal standard of transparency through readability.

3 Data and Measurement

3.1 Estimation Sample

For our empirical analysis, we construct an unbalanced quarterly panel with the texts of some 580,000 privacy policies posted by some 75,000 firms between 2014 and 2021. We complement this information with firm-level and industry-level information to obtain our main estimation sample.

3.1.1 Privacy Policy Panel

We use an unbalanced quarterly panel of the texts of privacy policies of German firms posted between 2014 and 2021. We constructed the panel by first web-scraping the Internet Archive (via the Wayback Machine) to obtain the historical versions of the policies of a large sample of German firms.[13]

For the construction of our final estimation sample, we impose a number of restrictions. First, we consider only German-language policies posted between Q1 2014 and Q2 2021. To further eliminate pages that are likely too short to contain privacy policies or too long to contain the privacy policies but nothing else, we drop observations that are shorter than the 2nd percentile and longer than the 98th percentile (measured in simple word tokens). Moreover, to ensure observations over the entire sample period (and to partially balance our panel), we restrict our sample to policies by firms for which we observe at least one observation (1) prior to the enforcement of the GDPR (May 25, 2018), and (2) after the enforcement of the GDPR.[14]

[13] We provide a detailed description of the various steps of this web-scraping exercise in the Online Appendix D. Our initial sample is drawn from the 2019 wave of the Mannheim Web Panel (Kinne and Axenbeck, 2019). The web panel includes unique firm identifiers that allow us to match our privacy policy data with the Mannheim Enterprise Panel (*Mannheimer Unternehmenspanel*, MUP) (Bersch et al., 2014).

[14] As we will show later, we do not observe any effects around the 2016 passage of the GDPR.

Our final sample comprises 585,329 privacy policies by 75,683 firms from Q1 2014 to Q2 2021. The average number of observations per firm is 4.4 pre-GDPR enforcement and 3.3 post-GDPR enforcement.[15]

3.1.2 Firm and Industry-Level Characteristics

We complement the privacy policy panel with information on firms' employees and sales from the Mannheim Enterprise Panel (MUP). The average firm in our sample has 36 employees and sales of 15 million Euros; we classify 61.6% as micro firms (less than 10 employees), 36.3% as small and medium-sized enterprises (between 10 and 250 employees), and 2% as large firms (more than 250 employees). In our sample, micro and large firms are underrepresented, and SMEs are overrepresented, relative to the Mannheim Enterprise Panel in 2017. We use firm-level sales data to calculate annual numbers for the Herfindahl-Hirschman Index (HHI) for all 4-digit NACE industries. We further obtain from the MUP the four-digit NACE Rev. 2 codes of the industry of firms' primary business activities. Our largest sector is the services sector, with 58.6% of all firms in 2017, followed by trade with 22.3%, manufacturing with 9.6%, construction with 7.0%, utilities with 1.5%, and agriculture/mining with 1%. In our estimation sample, services, manufacturing, and utilities are over-represented, whereas trade, construction, and agriculture/mining are underrepresented. We report these sample characteristics in Table 2.

3.2 Measuring Compliance: Disclosure and Readability

We use the text of firms' privacy policies to construct measures of their compliance. For our measures of disclosure, we use the number of *topics*, capturing the breadth (of content) of a privacy policy, and *topic-weighted words*, capturing the volume of disclosure. For readability, we borrow from linguists and use two measures of readability based on their regulatory use and their ability to explain differences in readability from a small gold-standard sample.

3.2.1 Disclosure

To measure how well policies disclose information, we determine the "main topic" (or distinct theme) of a given paragraph of the policy and then tally the number of distinct main topics for each policy. We use probabilistic topic models to find these topics in our policy corpus. More specifically, we apply the *Latent Dirichlet Allocation* (LDA) model (Blei et al., 2003).

[15] In Figure B.1 in the Online Appendix, we show the number of observations per quarter (i.e., the number of privacy policies by as many firms).

Table 2: Sample Characteristics

	Obs.	Mean	Std.	Min	Max
Number of observations per firm	75683	7.734	4.67	2	30
… in pre-GDPR enforcement phase	75683	4.446	3.69	1	18
… in post-GDPR enforcement phase	75683	3.288	2.17	1	13
Employees (firm-level means)	65863	36.446	408.48	1	48300
… Micro	40578	3.72	2.54	1	10
… Small and medium-sized (SME)	23920	39.222	42.13	10	249.6
… Large	1365	960.678	2671.81	250	48300
Sales (in million; firm-level means)	55656	14.942	351.78	0	62379.6
Herfindahl-Hirschman Index (HHI; in 2017)	44883	551.131	1178.23	1.5	10000
Economic Sector (2017)	**Estimation sample**		**MUP**		
Agriculture/Mining	688	1.03%	1.96%		
Manufacturing	6387	9.56%	6.72%		
Utilities	1028	1.54%	0.92%		
Construction	4679	7.01%	10.69%		
Trade	14907	22.32%	23.89%		
Services	39105	58.55%	55.82%		
	66794				

Notes: We report sample size and firm-level characteristics for the estimation sample. The number of employees and sales figures (firm-level means) are from the Mannheim Enterprise Panel (MUP), waves 47 to 61. Micro firms have less than 10 employees; small and medium-sized enterprises (SMEs) have between 10 and 250 employees; large firms have 250 employees or more. The reported numbers are the averages of all of a given firm's observations. Market concentration information (Herfindahl-Hirschman Index) is calculated from the Mannheim Enterprise Panel for 2017 (using the four-digit NACE industry classification). Economic sectors are based on a firm's primary NACE Rev. 2 code (as reported in 2017): Agriculture are sections A and B; manufacturing is section C; utilities are sections D and E; construction is section F; trade is section G; and services are sections H, I, J, K, L, M, N, P, Q, R, and S.

We follow a two-step approach:[16] To obtain the main topic for a given paragraph, we first estimate the topic model with $K = 50$ topics on the corpus of paragraphs, following Brody and Elhadad (2010). Assuming that each paragraph was written to cover a single topic, we define the main topic of a paragraph as the topic with the highest topic density. To count the main topics for each policy, we look at all the main topics from every paragraph and list each different topic that shows up at least once as a main topic.

For our second measure, we use the results from the LDA topic models to identify paragraphs that are more or less likely to contain information related to a firm's disclosure (by Art. 13 and 14 GDPR). Using higher weights for the word counts of more relevant paragraphs (those more likely disclosing relevant information) and lower weights for those of less relevant paragraphs, we thus construct a measure of the *topic-weighted informational volume*. We take a multi-step approach: First, from the per-paragraph assignments of main topics, we calculate a topic distribution where Θ_k represents the fraction of paragraphs with topic k

[16]For more details, see the description in Online Appendix C.

as their main topic. Second, we identify all paragraphs (both pre-GDPR and post-GDPR) that contain information related to disclosures per Art. 13 and 14, using simple text parsing techniques.[17] The total word count of disclosing paragraphs is the total number of *disclosed words*. For the subset of disclosing paragraphs taken from post-GDPR policies, we calculate the topic distribution with respective densities $\tilde{\Theta}_k$. Using the main-topic distributions for all paragraphs (Step 1) and for disclosing paragraphs, we calculate a topic weight factor $\phi_k = \frac{\tilde{\Theta}_k}{\Theta_k}$ for each k. We interpret a topic k with $\phi_k > 1$ (or $\tilde{\Theta}_k > \Theta_k$) as one that is more likely capturing information required by Art. 13 and 14 than an alternative topic k' with $\phi_{k'} < 1$. Third, we obtain the word count $w_{c|k}$ for each paragraph c of a given main topic k. We multiply the paragraph word counts by the paragraph's respective topic weight factor to obtain the number of *topic-weighted words* (i.e., $\sum_c \phi_k w_{c|k}$) as our measure of disclosure.[18]

We provide descriptive statistics for the disclosure proxies in panel (a) if Table 3. We observe a significant heterogeneity of topic-weighted words across policies, with some policies not containing any disclosing paragraphs.[19] Also, the average policy covers around 12 distinct topics. The shortest policy is quite minimalist (one distinct topic), whereas the paragraphs in the longest policy contain 47 distinct topics.

3.2.2 Readability

Many factors determine how readers comprehend written texts. For example, the use of common words will make texts more accessible to a wider audience, whereas the use of specialized terms or jargon will render texts more difficult to understand. Similarly, shorter sentences or simpler and shorter words will increase the readability of a text and the transparency of its content. To assess the reading ease (or difficulty of texts), readability indices and scores have been developed and used (e.g., in the United States) in regulatory contexts.[20] These indices or scores are typically constructed as weighted averages of a set of different readability factors.[21]

[17] For a list of terms, see Table B.1 in the Online Appendix.

[18] Table B.2 in the Online Appendix illustrates this construction of topic-weighted words using two simple examples.

[19] The distribution of total word count (see the data table in Figure B.4 in the Online Appendix) has a smaller variance than the distribution of topic-weighted words.

[20] In Michigan and Massachusetts, an insurance contract must have a Flesch Reading Ease (FRE) Score of at least 50 (Michigan Compiled Laws, Section 500.2236 (2020); General Laws of Massachusetts, Title XXII, Chapter 175 Section 2B. (2014)); in Texas, the minimum score of the FRE is 40 (Texas Insurance Code, Section 2301.053 (2019)) (Wagner, 2023). Similar guidelines (with a minimum score of 45) exist in Florida (Florida Statute §627.4145, Readable language in insurance policies; available at https://flsenate.gov/Laws/Statutes/2021/0627.4145).

[21] The literature, however, knows a large number of scores and indices developed for different languages and purposes that also vary in their popularity and use. See Table B.4 in the Online Appendix for a list of readability scores and their use in the literature.

Table 3: Disclosure and Readability

	Mean	Std.	Min	Max
Panel (a): Disclosure				
Distinct topics	12.36	9.66	1	47.1
Topic-weighted words	1006.42	1078.84	0	10998.2
Panel (b): Readability				
German Flesch Reading-Ease score	35.98	5.64	-185.8	86.7
LIW	56.13	3.94	22	260.3

Notes: This table reports our measures of disclosure (distinct topics and topic-weighted words, using Grün and Hornik (2011)) and readability (German FRE and LIW, using Benoit et al. (2018)) for all 585,329 privacy policies in our estimation sample.

As a standard to assess compliance with Art. 12, the Art. 29 Working Party (2018), a former advisory body within the EU's data protection framework, has alluded to readability scores and indices by proposing mechanisms such as "readability testing."[22] The Working Party, however, does not provide guidance on which index or score is the most suitable for the analysis of legal documents, such as privacy policies, which comprise a special text category.

We follow the Working Party's lead and use two scores—developed by linguists—for our analysis. First, we use the German version of the Flesch Reading Ease Score (Flesch, 1948; Amstad, 1978) because of its established use in a regulatory context.[23] It is defined as

$$\text{German FRE} = 180 - ASL - (58.5 \times AWL) \tag{7}$$

where ASL and AWL denote the average sentence and average word length (in syllables), respectively. Higher values of the German FRE indicate better readability.

For our second readability measure, we take a data-driven approach. We follow Benoit et al. (2019) who evaluate the textual complexity in political communication, employing the Bradley-Terry model for pair-wise comparisons (Bradley and Terry, 1952) of text snippets. To implement this approach, we first hand-collected about 4,000 pair-wise comparisons of portions of privacy policies (taken from our sample), asking subjects to rank the two text snippets in a given pair by their readability.[24] With the help of the Bradley-Terry model,

[22]Paragraph 9 reads: "If controllers are uncertain about the level of intelligibility and transparency of the information and effectiveness of user interfaces/ notices/ policies etc., they can test these, for example, through mechanisms such as user panels, readability testing, formal and informal interactions and dialogue with industry groups, consumer advocacy groups and regulatory bodies, where appropriate, amongst other things."

[23]For academic work, see Lin and Osnabrügge (2018) or Wojahn et al. (2015).

[24]We used the results from our LDA model to identify paragraphs that are central to understanding the processing of personal data. We then selected a random sample of paragraphs, each 60–80 words long (ruling out multiple paragraphs from the same firm), and constructed 700 text pairs. In batches of 100, we assigned

we then determine the readability score that best explains these pair-wise comparisons.

The best readability score (as best predictor of the data) is the *läsbarhetsindex* (LIW) (Björnson, 1968):

$$\text{LIW} = ASL + \frac{100 \times n_{wsy \geq 7}}{n_w} \tag{8}$$

with ASL the average sentence length (in words), $n_{wsy \geq 7}$ the number of words with at least seven syllables, and n_w the total number of words. Higher values indicate lower readability (unlike for the German FRE). Note that the patterns for both the German FRE and the LIW align well with our sample of pair-wise text comparisons.[25] While the LIW is not hugely popular in the literature, it outperforms all other scores typically used in research.[26]

We provide summary statistics of the German FRE and LIW index for our estimation sample in panel (b) of Table 3. For comparison, simple-language news pages are easier to read than privacy policies, and privacy policies are similar to political speeches, German constitutional court decisions, or Wikipedia pages. Ironically, the German text of the GDPR itself (the Datenschutzgrundverordnung/DS-GVO, a seven-syllable word) is highly unreadable. We provide scores of these other text corpora in Table B.3 in the Online Appendix.

3.3 Additional Data Sources

We use various other data sources to construct additional variables that help us capture regulatory scrutiny and capacity.

3.3.1 Enforcement Data from the UK Information Commissioner's Office

From Koutroumpis et al. (2022), we obtain case counts (at the three-digit NACE industry level) of the UK Information Commissioner's Office (UK ICO) for the time period from 2012 to Q2 2018 (the enforcement quarter of the GDPR). The UK ICO enforces data privacy laws, and industry-level case counts serve as a measure of a data protection agency's enforcement activities with varying degrees across industries. We observe at least one case in the above time period in 62% of all industries. We scale the per-industry case counts using industry-

the text pairs to 14 human subjects. Each batch was evaluated by at least six subjects.

[25] In Figure B.2 in the Online Appendix, we illustrate that, as the difference in the German FRE (RHS panel) and the LIW (LHS panel) increase, the percentage of pairs for which the human assessment aligns with the score-based ranking increases as well.

[26] It is of note that less popular scores tend to outperform the more popular ones. In Figure B.3 in the Online Appendix, we juxtapose the performance of readability scores (measured in the increase in node purity) and their popularity (measured in the number of Google Scholar citations). The figure illustrates that popularity in the literature and performance are not positively correlated.

level counts of all private companies in the UK in 2017.[27] The median industry with positive case counts as 1.8 cases per 1,000 firms (with an interquartile range of $[0.6, 5.7]$).

3.3.2 ZEW Business Survey in the Information Economy

The ZEW Business Survey in the Information Economy (ZEW, 2017) is a quarterly survey reaching out to German companies in the information economy (ICT service providers, ICT hardware manufacturers, media service providers, and knowledge-intensive service providers). From the Q4 2027 wave of the survey, we obtain answers to a question related to the importance of data for a business.[28] We take simple means of all responses at the 3-digit NACE industry level (for 34 industries with at least two responses; values range from 2.5 to 5.0 with a mean of 3.65) and use this industry-level variable as a measure of data intensiveness of an industry (within the information economy).

3.3.3 Budgets of German State Data Protection Authorities

Germany uses a federal system for data protection regulation. Each state has its own data protection authority (DPA) that regulates the compliance of firms located in the respective state. We use the official budget of the DPA in the firm's home state to proxy regulatory capacity: an authority with more resources is less likely to be a constrained regulator in the sense of our theoretical model.

We obtain state-level budget information from state governments' websites. We collect information on a DPA's overall budget (i.e., budgeted expenditures) and planned staff numbers.[29] We further use the total number of firms per state (from the MUP) to scale DPA budget and staff numbers. Using firm-level home state information (by firm headquarters) from the MUP, we can match each firm to the budget of its respective DPA.[30]

Figure 2 summarizes the budget situation of German DPAs and highlights variation both across states and over time. We plot the total budget per firm (in Euros) and staff numbers (per 1000 firms) for all 16 states. We see significantly higher per-firm figures (in panels (a)

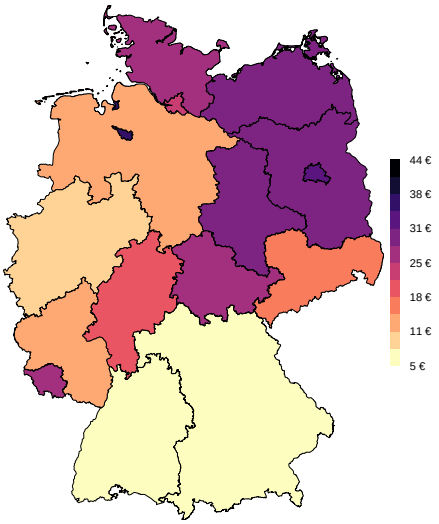
[27] We obtain the firm counts (UK Business Counts - enterprises by industry and employment size band) from UK Office of National Statistics's Nomis platform at https://www.nomisweb.co.uk/datasets/idbrent.

[28] The question in the survey: "How important is the use of personal data, such as that of customers or business partners, for your company?" Answers are on a 5-point Likert scale with 1 being "entirely unimportant" and 5 being "absolutely necessary."

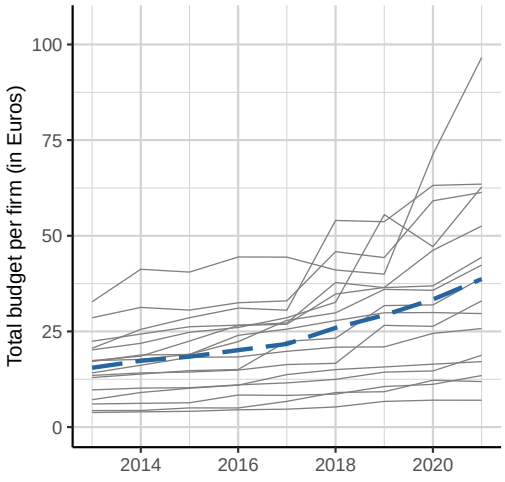
[29] We consider only the original budget numbers and disregard any revised budgets.

[30] The MUP provides address information for most years. For missing observations, we extrapolate forward and backward, using a firm's first address information for all prior observations and a firm's last address information for all following observations. We also interpolate missing observations between two observations for which the firm's state has not changed. If, between two observations, the firm has moved to another state, we do not fill the in-between observation (and retain missing values).

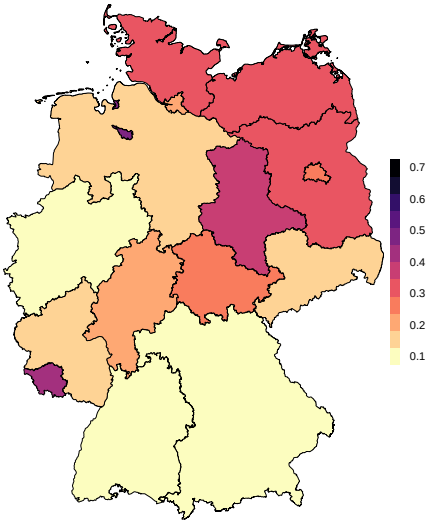
Figure 2: Budgets of German State Data Protection Authorities



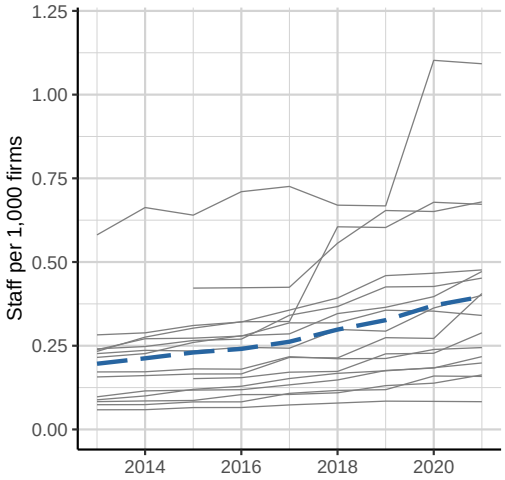
(a) Total Budget (2018)



(b) Total Budget (2013–2022)



(c) Staff (2018)



(d) Staff (2013–2022)

Notes: This figure presents (a) a map of the state-level budgeted total expenditure per firm (in 2018); (b) the individual states' (gray) and the average (blue) budgeted total expenditure per capita (2013–2022); (c) a map of the state-level (planned) staff numbers per 1000 firms; and (d) the individual states' (gray) and the average (blue) staff numbers for 1000 firms (2013–2022).

and (c), for 2018) in small states (Berlin, Bremen, and Hamburg) but also in states in the northeast of the country. In panels (b) and (d), we see an increase in the average DPA per-firm budget and staff (in blue), with significant heterogeneity of the development over time for individual states (in gray).

4 Firms' Responses to the GDPR

In this section, we document GDPR-associated changes in the amount of disclosure and the readability of privacy policies. Following our theoretical framework (Prediction 1), we expect that the increased stringency of the transparency requirement following the introduction of the GDPR in Q2 2018 leads to longer privacy policies that disclose more information to users. We further expect that firms write better privacy policies that are easier to read for users. This latter effect, if it exists, should be weaker than the effect on disclosure.

4.1 Disclosure Before and After the GDPR

In panel (a) of Figure 3, we plot the quarterly averages of our disclosure measures. Both figures paint a similar picture: the content breadth of privacy policies (topics) has doubled, and the disclosure in policies has more than tripled with the enforcement of the GDPR.[31] For quarter Q2 2018, we plot average values before and after the enforcement of the GDPR; the documented gap is, therefore, within-quarter. The count of distinct topics does not continue to increase after Q2 2018, suggesting that the breadth of policies remains relatively constant, whereas the details of the documents (and the amount of disclosed information) increase as firms continue to adapt to the new regulatory regime.

Following the line of reasoning in Johnson et al. (2023) or Peukert et al. (2022), we attribute the sudden change in our disclosure outcome variables (depicted in Figure 3) to the GDPR-induced change in regulatory stringency itself. In the top panel of Table 4, we present fixed-effects OLS regression results, accounting for observed and unobserved heterogeneity. We estimate the following model:

$$disclosure_{it} = \beta_0 + \beta_1 Post\text{-}GDPR_t + \beta_2 \mathbf{X}_{ky} + \eta_i + \nu_y + \varepsilon_{it} \tag{9}$$

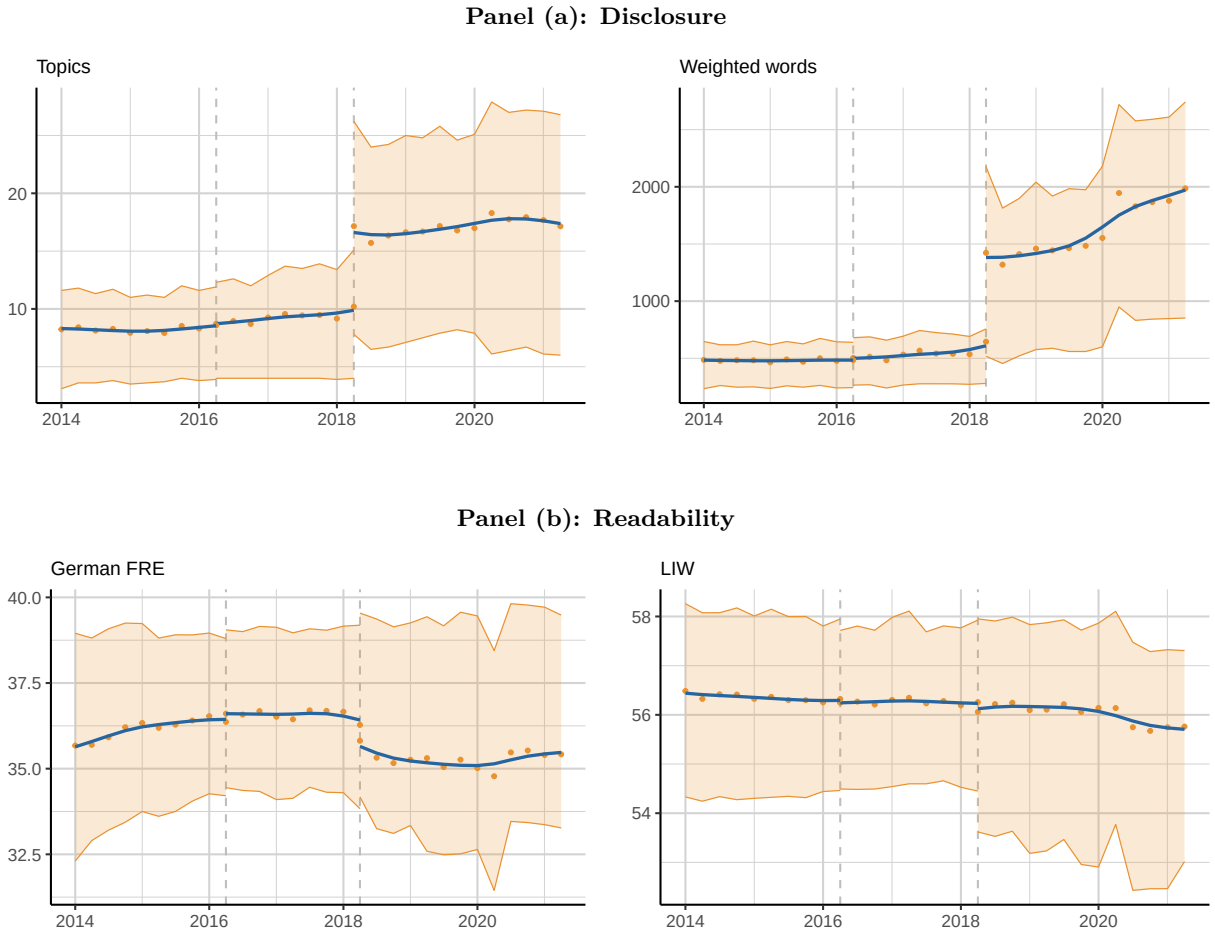
where $disclosure_{it}$ is the disclosure outcome variable by firm i (in an industry k) in quarter-year t (of year y), $Post\text{-}GDPR_t = 1$ for all policies after May 25, 2018, and zero otherwise; $\mathbf{X}_{ky}$ is a vector of time-varying (over years) industry-level and firm-level characteristics (market concentration and firm size); and η_i and ν_y are firm and year fixed effects to capture unobserved heterogeneity across firms and time.[32]

Our main variable of interest is $Post\text{-}GDPR_t$. Because our dependent variables are in log, we can interpret the post-GDPR effect as a percentage change of our dependent variable. The results align well with our descriptive evidence in Figure 3. Policies in the post-GDPR

[31] We see no such effect of the passage of the GDPR in Q2 2016 (i.e., no announcement effect).

[32] Our unit of analysis is the privacy policy of a firm-quarter observation.

Figure 3: Disclosure and Readability in Response to the GDPR



Notes: This figure presents quarterly averages of policy-level measures for disclosure (panel (a)) and readability (panel (b)). Dots represent quarterly averages; the curves are fitted to the data (spline); the lower and upper bounds represent the 25th and 75th percentiles, respectively (interquartile range). The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018.

period are, on average, 50% broader than in the pre-GDPR period. Moreover, they disclose almost 80% more content. All estimation coefficients (for the GDPR dummy) are statistically significant at the 1% level.

Our results suggest that firms redrafted their privacy policies to comply with the new rules in the GDPR. The obligation requiring firms to inform users about the processing of data, of course, is not an entirely new concept in the EU legal order. Prior to the GDPR, the Data Protection Directive (DPD) already required firms to inform data subjects about the identity of the data controller as well as the purposes of the processing of the data (Art. 10

Table 4: GDPR-Induced Changes

Dependent variable (in log):	**Topics**		**Weighted words**	
	(1)	(2)	(3)	(4)
Post GDPR (=1)	0.4574***	0.4885***	0.7490***	0.7775***
	(0.0061)	(0.0074)	(0.0060)	(0.0073)
Concentration (HHI in '00)		-0.00002		0.0002**
		(0.0002)		(0.0001)
log Employees		0.0076		0.0186***
		(0.0087)		(0.0064)
# Firm FE	75,677	64,600	75,683	64,609
R^2	0.659	0.696	0.757	0.782
Observations	585,141	409,377	585,329	409,527
Dependent variable (in log):	**German FRE**		**LIW**	
	(1)	(2)	(3)	(4)
Post GDPR (=1)	-0.0392***	-0.0418***	-0.0039***	-0.0041***
	(0.0014)	(0.0018)	(0.0005)	(0.0006)
Concentration (HHI in '00)		0.00006**		-0.00001
		(0.00003)		(0.00001)
log Employees		-0.0030*		0.0004
		(0.0015)		(0.0006)
# Firm FE	75,680	64,606	75,683	64,609
R^2	0.592	0.624	0.612	0.648
Observations	585,145	409,433	585,329	409,527

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE (8)). Dependent variables are measures of disclosure (topics and weighted words) in the top panel and readability (German FRE and LIW) in the bottom panel. All dependent variables are in log. Additional control variables are HHI (as a measure of market concentration) and log Employees (as a measure of firm size). Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

DPD).[33] With the entry into force of the GDPR, however, legal requirements of privacy policies have been fundamentally transformed. The GDPR introduces new categories a data subject has to be informed about. Examples include the legal basis for the processing of the data (Art. 13(1)(c) GDPR) and information about the rights of data subjects, such as the right to rectification, data portability, or the erasing of personal data. The changes we observe, therefore, capture additional information privacy policies now contain.

[33]The DPD entered into force in 1995 and was implemented into German law via the *Telemediengesetz* (TMG), which codified the duty to inform the data subject about the nature, scope, and purpose of the collection and use of personal data (§13(1) TMG).

4.2 Readability Before and After the GDPR

Panel (b) in Figure 3 depicts our readability scores. The effect of the GDPR on readability is ambiguous. While the German FRE decreases by about 3% post-GDPR, implying a decrease in the readability of privacy policies, the decrease of the LIW, albeit weak (but more pronounced in later periods), means an increase in readability. In the bottom panel of Table 4, we present fixed-effects regression results for the following model:

$$readability_{it} = \beta_0 + \beta_1 Post\text{-}GDPR_t + \beta_2 \mathbf{X}_{ky} + \eta_i + \nu_y + \varepsilon_{it}. \tag{10}$$

The results do not change with the inclusion of firm-level and industry-level characteristics (models (2) and (4)). The German FRE is 4% lower for post-GDPR policies, implying a *decline* in readability. The LIW is 0.4% lower for post-GDPR policies, implying a (small) *improvement* of readability. Both effects are statistically significant at the 1% level.

The reported coefficients for market concentration and firm size suggest that policies by firms in more concentrated markets are more readable (higher German FRE), whereas larger firms have less readable privacy policies (lower German FRE). As readability is meant as a measure of accessibility by users online, higher readability can be considered inherently pro-consumer. This interpretation allows us to draw a direct parallel with the measure of consumer friendliness of end-user license agreements as studied, for instance, in Marotta-Wurgler (2007, 2008).[34]

4.3 Heterogeneity

Small firms may not have the capacity or capability to pay attention to the accuracy of the disclosures and the readability of their privacy policies. Large companies, on the other hand, with teams of lawyers and significantly larger fines when found non-compliant, might be the firms driving our baseline results in Table 4. We find that large firms indeed exhibit the strongest compliance effects of the GDPR on readability. However, small firms respond, too. In fact, small firms add relatively more topics than large firms, and the addition of disclosure-related words is similar to that of large firms.

There is also considerable heterogeneity across industries. Changes in the LIW are entirely driven by firms in the trade and services sectors (making up about 80% of our sample). For the German FRE, on the other hand, we observe statistically significant effects for all industries, with firms in construction exhibiting the least compliance and firms in services

[34]Marotta-Wurgler (2007) reports a negative correlation between firm revenue and pro-consumer bias in the contracts' terms, whereas Marotta-Wurgler (2008) finds no significant correlation between HHI and consumer friendliness.

and agriculture/mining the most compliance (i.e., the smallest decline in readability). We observe similar degrees of heterogeneity for our disclosure, with curious patterns. Utilities, for instance, exhibit the smallest increase in content breadth (topics) but the largest increase in disclosure-related words.[35]

5 Regulatory Exposure, Scrutiny, and Capacity

In the previous section, we presented results on firms' average responses to the GDPR in line with Prediction 1. We further discussed the heterogeneity of these results with respect to firm size and industry sector. In this section, we explore this heterogeneity further, introducing variation in firms' exposure to the GDPR and the regulator's attention and capability. In a first step, we take a closer look at the response of firms that experience different degrees of exposure. We then take an empirical look at Predictions 2 and 3 when we explore the response of firms that (i) anticipate different degrees of attention (or scrutiny) by the regulator and (ii) are adjudicated by regulators with different budgetary constraints.

Our model builds on an assumption of limited (and asymmetric) enforceability of the transparency principle. Regulators face more challenges when enforcing the readability requirement, and firms respond with under- or non-compliance. We explore the interaction between compliance and enforceability through several different angles and conclude with a brief discussion of alternative explanations of the baseline results in Table 4 and how the results in this section help us refute them.

5.1 GDPR Exposure as Treatment Intensity

Firms with highly readable privacy policies (prior to the GDPR) will find it easier to comply with the readability requirement than firms with policies of low readability, as the latter will have more catching up to do. The same reasoning applies to firms whose policies are already very detailed (with high levels of disclosure) relative to firms with shorter, less informative policies. Firms that are already in compliance with their pre-GDPR policies thus experience a lower GDPR treatment intensity than firms with non-compliant policies.

In Table 5, we present disclosure and readability results for firms with different treatment intensities (see, e.g. Chen et al., 2022). We estimate the following model:

$$\begin{aligned} outcome_{it} &= \beta_0 + \beta_1 Post\text{-}GDPR_t + \beta_2 Post\text{-}GDPR_t \times exposure_i + \\ &\quad \beta_3 exposure_i + \beta_4 \mathbf{X}_{ky} + \eta_i + \nu_y + \varepsilon_{it}. \end{aligned} \tag{11}$$

[35] We report these results in Table B.5 in the Online Appendix.

Table 5: Exposure ("Treatment Intensity")

Dependent variable (in log):	**Disclosure**		**Readability**	
	Topics	**Weighted words**	**German FRE**	**LIW**
	(1)	(2)	(3)	(4)
High disclosure (Topics)	0.0771***			
	(0.0084)			
Low disclosure (Topics)	0.9292***			
	(0.0101)			
High disclosure (Weighted words)		0.4507***		
		(0.0075)		
Low disclosure (Weighted words)		1.144***		
		(0.0092)		
High readability (German FRE)			-0.1095***	
			(0.0021)	
Low readability (German FRE)			0.0196***	
			(0.0021)	
High readability (LIW)				0.0244***
				(0.0007)
Low readability (LIW)				-0.0320***
				(0.0007)
# Firm FE	64,583	64,609	64,606	64,609
R^2	0.722	0.805	0.645	0.679
Observations	409,320	409,527	409,433	409,527

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE (8)). Dependent variables are measures of disclosure (topics and weighted words) and readability (German FRE and LIW). We report the coefficients for the Post GDPR dummy for firms with high and low pre-GDPR values of the dependent variable. All dependent variables are in log. Additional control variables are HHI (as a measure of market concentration) and log Employees (as a measure of firm size). Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

where *outcome* is one of our disclosure or readability measures and $exposure_i = 1$ if the mean of disclosure or readability for firm i from all pre-GDPR observations is below the median of these observations (*low pre-GDPR disclosure* or *low pre-GDPR readability*), and $= 0$ if it is above the median (high pre-GDPR disclosure or high pre-GDPR readability).

We find that higher treatment intensity (or higher GDPR exposure) triggers stronger effects both for disclosure and readability. Firms with low pre-GDPR disclosure (weighted words) increased their disclosure by more than 100%, whereas firms with high pre-GDPR disclosure (weighted words) increased disclosure by less only 45%. Likewise, the number of topics increased by more than 90% for low pre-GDPR disclosure firms vs. less than 10% for high pre-GDPR disclosure firms. We find a positive effect of the GDPR on disclosure regardless of the level of treatment intensity (measured as above or below median).

For firms with high pre-GDPR readability of their privacy policies, readability worsens after the GDPR. Conversely, firms that had low pre-GDPR readability saw an improvement after the GDPR. While the GDPR may not be effectively increasing average readability

(see the bottom panel of Table 4), it is effective for those firms that needed to improve the most. This result holds true for both readability measures. The German FRE decreases (readability worsens) by 10% (or about 7/10 of a standard deviation) for high pre-GDPR readability firms and increases (readability improves) by 2% (or about 1/8 of a standard deviation). The LIW increases (readability worsens) by 2.4% (or about 1/3 of a standard deviation) vs. decreases by 3.2% (or a bit less than 1/2 of a standard deviation).

The results in Table 5 identify a GDPR-induced effect, and we do not find evidence that they are driven by a general convergence or reversion to a mean. When plotting quarterly averages of our outcome variables, we find a relatively constant gap between the mean readability of above-median and below-median firms. If general convergence or mean reversal were driving our results, this gap should be narrowing. Our results are also not an artifact of how we split the pre-GDPR sample. When plotting the conditional effects of the GDPR on our outcomes variables pre-GDPR percentile, we see a downward trend.Firms with low GDPR values exhibited stronger (positive) effects than firms with high pre-GDPR values.[36]

5.2 Regulatory Scrutiny and Attention

We now explore the role of regulatory scrutiny. Prediction 2 states that firms under higher anticipated regulatory scrutiny will respond to the GDPR with more readable policies. To test this prediction, we re-estimate the models in equations (9) and (10) by interacting the Post GDPR dummy with proxies for regulatory scrutiny, $scrutiny_k$:

$$\begin{aligned} outcome_{it} &= \beta_0 + \beta_1 Post\text{-}GDPR_t + \beta_2 Post\text{-}GDPR_t \times scrutiny_k + \\ &\quad \beta_3 scrutiny_k + \beta_4 \mathbf{X}_{ky} + \eta_i + \nu_y + \varepsilon_{it}. \end{aligned} \tag{12}$$

where $outcome_{it}$ is a disclosure or readability measure, and $scrutiny_k$ is at the industry level.

We take three different approaches to construct our measures of scrutiny. First, we use enforcement data from the UK Information Commissioner's Office for the time period from 2012 to the second quarter of 2018 (Koutroumpis et al., 2022). We calculate case counts at the 3-digit industry level and scale them by the number of UK firms in those industries. For our empirical analysis, we use an index for industry k with four levels: *no* enforcement if there are no UK ICO actions in that industry, *low*, *medium*, and *high* enforcement if the per firm UK ICO actions are in the first, second, or third tercile of industries with non-zero case counts. Our identifying assumption is that industries that were scrutinized by a privacy regulator before the GDPR were also primary targets after GDPR and that this variation

[36] We provide the respective graphs in Figures B.6 and B.7 in the Online Appendix.

of enforcement also applies to Germany.[37] This measure for scrutiny is time-invariant and based on pre-GDPR information.

For our second proxy of regulatory scrutiny, we use survey evidence on the importance of personal data for firms from the ZEW Business Survey in the Information Economy (ZEW, 2017) conducted in Q4 of 2017. We take the average of all responses at the 3-digit industry level (for 34 industries). Values range from 2.5 (low importance) to 5 (high importance). This measure for scrutiny is time-invariant and based on pre-GDPR information.

For our third proxy of regulatory scrutiny, we use the Herfindahl-Hirschman Index as a measure of (sales-based) market concentration. This proxy is time-varying. We believe that industries with higher concentration are more likely a primary target for a number of reasons. First, regulators looking for the largest impact (in terms of affected users) will likely focus on concentrated industries with (relatively) large firms. Second, when regulators respond to complaints by the public (either users or consumer advocacy groups), we ought to expect more complaints aimed at larger firms and those in concentrated industries.[38] Third, firms in concentrated industries may not be exposed to competitive pressures that can induce better compliance, and regulators are more likely to step in to correct this imbalance.[39]

Table 6 reports our results for all three measures of regulatory scrutiny. We find evidence in support of Prediction 2 for both disclosure and readability. The amount of disclosure does not increase much (or even decreases) with higher levels of regulatory scrutiny. We see this, for instance, when comparing the coefficient for no enforcement and high enforcement (panel (a)) (notice, though, the patterns are non-monotonic). The increase in the number of topics differs by 14% (relative to the no-enforcement baseline). The increase in topic-weighted words decreases by 2%. The latter result is in line with high compliance costs. We also see a weaker increase of topics and weighted words for firms in industries with higher levels of concentration (panel (c))—with an imprecisely estimated negative coefficient on the interaction term for weighted words. We see a positive effect on the firms' disclosure response to the GDPR for higher importance of data. This result is likely because when the use of personal data is important, firms will have more content to disclose in their policies.

We further find evidence in support of Prediction 2 for readability. In panel (a), we see a

[37] We do not need to assume that German firms pay attention to the enforcement activities of the data protection agency in the UK. We simply assume that German firms expect UK enforcers to pay attention to the same industries as German enforcers.

[38] Individuals can complain to the relevant data protection authorities if they believe that their rights have been violated. When responding to complaints, data protection authorities can levy fines as a punitive measure. Alongside the enforcement of data protection authorities, individuals also have the right to bring to court GDPR claims against private entities and pursue damages under Art. 82 GDPR.

[39] Related to this point is the emergence of a software monoculture (or IT monoculture) in highly concentrated industries that can generate cyber-security risks. See, for instance, Geer et al. (2003) or Whittaker (2003).

Table 6: Scrutiny and Compliance

Dependent variable (in log):	**Disclosure**		**Readability**	
	Topics	**Weighted words**	**German FRE**	**LIW**
	(1)	(2)	(3)	(4)
Panel (a): Enforcement history (UK ICO)				
UK ICO: No enforcement	0.4110***	0.7918***	-0.0450***	-0.0030*
	(0.0239)	(0.0189)	(0.0047)	(0.0016)
UK ICO: Low enforcement	0.5499***	0.7964***	-0.0428***	-0.0044***
	(0.0107)	(0.0096)	(0.0025)	(0.0008)
UK ICO: Medium enforcement	0.4395***	0.7504***	-0.0481***	-0.0026***
	(0.0125)	(0.0105)	(0.0027)	(0.0009)
UK ICO: High enforcement	0.4715***	0.7767***	-0.0321***	-0.0051***
	(0.0141)	(0.0117)	(0.0028)	(0.0010)
# Firm FE	63,740	63,749	63,746	63,749
R^2	0.697	0.782	0.624	0.648
Observations	403,302	403,452	403,358	403,452
Panel (b): ICT firms (ZEW information economy survey)				
Post GDPR (=1)	0.2641**	0.6035***	-0.1049***	0.0037
	(0.1078)	(0.0850)	(0.0236)	(0.0080)
Importance of data	-0.0484	0.0728	-0.0050	0.0158*
	(0.1127)	(0.0881)	(0.0207)	(0.0096)
Post GDPR (=1) × Importance of data	0.0542*	0.0582***	0.0153**	-0.0011
	(0.0285)	(0.0224)	(0.0063)	(0.0021)
# Firm FE	11,799	11,801	11,800	11,801
R^2	0.694	0.783	0.644	0.647
Observations	70,163	70,203	70,174	70,203
Panel (c): Market concentration				
Post GDPR (=1)	0.4924***	0.7783***	-0.0424***	-0.0041***
	(0.0075)	(0.0074)	(0.0018)	(0.0006)
Concentration (HHI in '00)	0.0006**	0.0004*	-0.00004	-0.00002
	(0.0003)	(0.0002)	(0.00006)	(0.00002)
Post GDPR (=1) × Concentration	-0.0008**	-0.0002	0.0001*	0.000010
	(0.0004)	(0.0003)	(0.00007)	(0.00003)
# Firm FE	64,600	64,609	64,606	64,609
R^2	0.696	0.782	0.624	0.648
Observations	409,377	409,527	409,433	409,527

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE (8)). Dependent variables are measures of disclosure (topics and weighted words) and readability (German FRE and LIW). In panel (a), we report the coefficients for the Post GDPR dummy for industries with different levels of UK ICO enforcement. In panel (b), we report the results for a subsample of firms in the ICT sector from a regression in which we interact the Post GDPR dummy with a measure of the firm's self-reported importance of personal data (low = 1, high = 5). In panel (c), we report the Post GDPR dummy interacted with our measure of market concentration. All dependent variables are in log. In panels (a) and (b), additional control variables are HHI (as a measure of market concentration) and log Employees (as a measure of firm size); in panel (c), the additional control variable is log Employees. Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

stronger improvement of LIW readability when pre-GDPR enforcement is high (relative to no pre-GDPR enforcement). The difference is 70% (relative to the no-enforcement baseline). Moreover, we see a weaker decline in German FRE readability for higher pre-GDPR enforcement, with a difference of close to 30%. The results in panel (b) and (c) paint the same picture. Positive coefficients on the interaction terms indicate a weaker decline in German FRE readability. The negative coefficient on the interaction term in column (4) for the LIW, indicating a stronger increase in readability, is statistically insignificant.

5.3 Regulatory Capacity at State-Level Agencies

Prediction 3 states that firms who are more likely to face an unconstrained regulator (with higher regulatory capacity) exhibit better compliance with the readability requirement. The effect on disclosure, if any, is likely much weaker (because firms comply with the disclosure requirement regardless of the regulator's capacity).

We use budget numbers for 16 state DPAs to measure the resources and capacity of state regulators that oversee a given firm's compliance with the GDPR.[40] We thus leverage a considerable degree of variation across states and over time (see Figure 2). The underlying assumption is that DPAs with larger (per firm) budgets are less likely constrained, and the theoretical implications from our model summarized in Prediction 3 apply. We estimate the following model and show the results in Table 7:

$$\begin{aligned} outcome_{it} &= \beta_0 + \beta_1 Post\text{-}GDPR_t + \beta_2 Post\text{-}GDPR_t \times budget_{i,y-1} + \\ &\quad \beta_3 budget_{i,y-1} + \beta_4 \mathbf{X}_{ky} + \eta_i + \nu_y + \varepsilon_{it}. \end{aligned} \tag{13}$$

where *outcome* is one of our disclosure or readability measures and $budget_{i,y-1}$ is a lagged (by one year) firm-level variable (that varies by the firm's state). For the budget variable, we use the data protection authority's (DPA) total budget per firm in that state (in panel (a)) and the number of staff positions per 1,000 firms (in panel (b)).

First, fully in line with Prediction 3, we do not see a statistically significant effect of regulatory capacity on firms' disclosure compliance. The point estimates are positive for topics and negative for weighted words. The latter, if anything, hints at weaker compliance with the disclosure requirement in states with higher-budget regulators. These results are well in line with our theory model, which predicts negative effects on disclosure compliance

[40] As outlined in Art. 4 (16) lit. a GDPR, the relevant data protection authority is determined based on the location of a firm's central administration. A firm's central administration is the establishment in which its main management activities are taking place and does not require that the data processing is actually carried out in this location (Recital 36 GDPR).

Table 7: Capacity and Compliance

Dependent variable (in log):	Disclosure		Readability	
	Topics	**Weighted words**	**German FRE**	**LIW**
	(1)	(2)	(3)	(4)
Panel (a): DPA Budget – Total budget per firm				
Post GDPR (=1)	0.4484***	0.7576***	-0.0419***	-0.0042***
	(0.0105)	(0.0090)	(0.0022)	(0.0007)
Total budget (per firm, lagged)	-0.0006	-0.0004	-0.0005***	0.00003
	(0.0010)	(0.0008)	(0.0002)	(0.00007)
Post GDPR (=1) × Total budget	0.0006	-0.0006	0.0002*	0.00002
	(0.0006)	(0.0005)	(0.0001)	(0.00004)
# Firm FE	75,484	75,490	75,487	75,490
R^2	0.659	0.757	0.592	0.612
Observations	583,410	583,598	583,414	583,598
Panel (b): DPA Budget – Staff per 1000 firms				
Post GDPR (=1)	0.4557***	0.7579***	-0.0426***	-0.0038***
	(0.0112)	(0.0093)	(0.0023)	(0.0008)
Staff (per 1000 firms, lagged)	0.0764	-0.0210	-0.0761***	0.0124*
	(0.1097)	(0.0782)	(0.0215)	(0.0071)
Post GDPR (=1) × Staff	0.0090	-0.0575	0.0219*	-0.0005
	(0.0579)	(0.0428)	(0.0112)	(0.0038)
# Firm FE	75,484	75,490	75,487	75,490
R^2	0.660	0.757	0.593	0.612
Observations	578,253	578,436	578,252	578,436

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE (8)). Dependent variables are measures of disclosure (topics and weighted words) and readability (German FRE and LIW). We report the interaction term of the Post GDPR (=1) dummy and a budget variable (budgeted total expenditure per firm in panel (a) and total staff per 1000 firms in panel (b). All dependent variables are in log. Clustered (firms) standard errors in parentheses. Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

for firms with sufficiently high compliance costs. Second, we find results for the readability requirement in support of our theoretical prediction. The results for the German FRE are well in line with our prediction. Firms in higher-budget states exhibit better readability compliance than other firms—the interaction terms are positive and statistically significant at the 10% level. The coefficients for the LIW are statistically insignificant, with a point estimator in line with Prediction 3 for the DPA's staff in panel (b).

Overall, the effects of state regulators' budgets on firms' compliance support our theoretical predictions. Because firms' disclosure compliance is at a high baseline level (see the top panel of Table 4), additional regulatory capacity has little effect on firm behavior. For readability, firms facing a higher-budget regulator anticipate stronger enforcement of the readability requirement. We see evidence of improved readability compliance (for the German FRE) in response to stronger regulatory capacity.

5.4 Discussion of Alternative Explanations

Our model builds on an assumption of limited (and asymmetric) enforceability of the transparency principle. Regulators face more challenges when enforcing the readability requirement, and firms respond with under- or non-compliance. Correspondingly, we find a small (and ambiguous) readability response by firms to the GDPR in the bottom panel of Table 4. Other explanations, however, may yield empirically indistinguishable results. In this section, we introduce three such explanations that may potentially explain the baseline results in Table 4 but cannot explain those in Tables 5, 6, and 7 in Section 5.

Ambivalence: Regulators may derive no value from the enforcement of (and compliance with) the readability requirement. They neglect it in their enforcement because of this ambivalence rather than because of enforcement difficulties; and rational firms will respond with little or no compliance (Table 4).

The results in Section 5, however, do not support this explanation. If firms anticipated that regulators did not care about readability, then we would not see any differences in changes in readability for high vs. low pre-GDPR compliance (Table 5). Moreover, firms anticipating more stringent scrutiny from regulators would not change readability more or less than firms anticipating lax scrutiny—because the regulator's attention is not focused on readability (Table 6). Last, for the results in Table 7 to be explained by regulators' ambivalence, we would need to observe lower values of readability for low-budget states and higher values for high-budget states. This is not reasonable. Because of the same set of rules to be enforced, we are more likely to see the same valuation of readability across state regulators but differential enforcement (and eventually compliance) because of budget constraints.

Cost Differentials: A second alternative explanation relies on compliance cost differences. For firms, compliance with readability might simply be prohibitively costly. Regulators, in fact, enforce the readability requirement just as effectively as the disclosure requirement (no enforcement asymmetry), but because of asymmetric compliance costs, we observe under- or non-compliance.

If compliance cost differentials are the main driver of our results, then our results for exposure (i.e., treatment intensity) in Table 5 should point in the opposite direction. For instance, firms with low compliance costs (and high pre-GDPR compliance) ought to become even better and improve compliance more than high-compliance cost firms. We see this in Marcus (1988), for instance, which shows that firms that performed poorly continued that path after a strengthening of regulatory rules, whereas better-performing firms were able to

improve even further on their strong performance.

Moreover, compliance-cost differences would be able to explain our results in Tables 6 and 7 only if these differences are correlated with the respective industries or states. Particularly for readability, there is little reason to believe that the costs of drafting readable legal documents differ across industries or states (with the same rules in place). For disclosure, however, one may argue that for firms in industries with higher importance of personal data, the amount of disclosure for compliance—and thus the compliance cost—is also higher. We take this explanation, discussed above, as the most plausible.

Measurement: Last, the enforcement difficulties are inherently related to measurement issues, and our metrics for readability may simply not capture firms' compliance with the readability requirement. Because measurement difficulties are the same across firms and industries, Tables 5 and 6 do not support this explanation of our results. Neither do differences in measurement difficulties across states, as all policies are in German.

6 Conclusion

In this paper, we study compliance with asymmetric enforceability, with a particular focus on the transparency principle of the GDPR, compelling firms to disclose information about the nature of their data collection, processing, and use in a "concise, transparent, intelligible and easily accessible form, using clear and plain language" (Art. 12(1) GDPR). Disclosure is objective and easy to verify. Readability, on the other hand, is subjective and vague, rendering compliance difficult to enforce. We show in a simple theoretical framework that this asymmetry in enforceability will lead to differential dynamics in firms' compliance. Firms will anticipate regulators to enforce what is indeed enforceable and then comply accordingly.

We apply these theoretical insights to the data, using a sample of more than 585,000 privacy policies posted by more than 75,000 German firms between 2014 and 2021. We find strong evidence for disclosure compliance but weak evidence for readability compliance for the average firm. However, we also find evidence for firms responding to higher exposure to regulation: Firms with low-readability policies prior to the GDPR improved their policies, while firms with high pre-GDPR readability experienced a decline in readability.

Our model predicts a stronger response in readability than in disclosure for firms that expect to see more attention from regulators. Using information on enforcement activities by the UK Information Commissioner's Office (Koutroumpis et al., 2022), the self-reported importance of the use of personal data for businesses in the information economy (ZEW, 2017), and market concentration measures as industry-level proxies for regulatory scrutiny, we con-

firm this prediction. We document stronger effects of scrutiny on readability compliance than on disclosure compliance, intuiting that firms already exhibit high disclosure compliance and more regulatory scrutiny should not have a meaningful effect on their disclosure (relative to readability).

Finally, we leverage the unique regulatory landscape in Germany, where 16 data protection authorities, each with its own budget, are responsible for enforcing EU data protection law. We exploit variation across states and time in the authorities' budgets to examine the effect of a regulator's budget constraint (and the impact that has on its enforcement activities) on the respective firms' compliance. Our data confirms our theoretical prediction that a regulator's constraint does not affect firms' disclosure compliance. However, we find evidence that firms in states with higher-budget data protection authorities exhibit better readability compliance.

Our results have immediate implications for the enforcement activities of agencies and can explain why the GDPR falls short of its potential (European Commission, 2019). Moreover, understanding the role of regulators' incentives and constraints on compliance is important beyond the specific predictions of our model and data. It is crucial to understand under which conditions enforcement of legal requirements can be carried out effectively and efficiently. The limitations that come with reductions in regulatory agency budgets affect enforceability and, therefore, compliance more generally and beyond the context of the GDPR.

Recent EU legislation uses language similar to that of the GDPR to define its transparency standards, and our results speak more generally to the effectiveness (or lack thereof) of regulatory tools that are based on difficult-to-verify information. Article 3 of the Platform-to-Business (P2B) Regulation (2019) requires firms to draft their terms and conditions in "plain and intelligible language." Article 14 of the Digital Services Act (2022) mentions "clear, plain, intelligible, user-friendly and unambiguous language […] in an easily accessible and machine-readable format." Similarly, in the U.S., in the absence of federal privacy regulation, we see state-level privacy laws mushrooming that also include provisions targeting readability. For instance, the California Consumer Privacy Act (CCPA) requires that information be made available in a "format that is easily understandable to the average consumer" (1798.130. (B) (iii) CCPA), and the Colorado Privacy Act (CPA) mandates that a privacy notice should be "reasonably accessible, clear, and meaningful" (§6-1-1308 (1)(a) CPA).[41] Given the results of our study, these standards will most likely not have the impact the legislator might have hoped for.

[41] §501.711 (1) Florida Digital Bill of Rights, §6 (c) of the Connecticut Privacy Act, §541.102. (a) of the Texas Data Privacy and Security Act, and §9 (a) of the proposed Massachusetts Data Privacy Act are further examples that encompass readability requirements with various versions of similarly vague language.

References

Amos, R., G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer (2021): "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset," in *Proceedings of The Web Conference 2021*, Association for Computing Machinery, WWW '21, 22.

Amstad, T. (1978): "Wie verständlich sind unsere Zeitungen?" Ph.D. thesis, University of Zurich.

Armstrong, M. and D. E. M. Sappington (2006): "Regulation, Competition, and Liberalization," *Journal of Economic Literature*, 44, 325–366.

Art. 29 Working Party (2018): "Article 29 Working Party: Guidelines on Transparency Under Regulation 2016/679 (wp260rev.01)," The Working Party on the Protection of Individuals with Regard to the Processing of Personal Data, available for download at https://ec.europa.eu/newsroom/article29/items/622227.

Ash, E. and S. Hansen (2023): "Text Algorithms in Economics," *Annual Review of Economics*, 15, 659–688.

Bakos, Y., F. Marotta-Wurgler, and D. R. Trossen (2014): "Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts," *Journal of Legal Studies*, 43, 1–35.

Bardsley, P. (1996): "Tax Compliance Games with Imperfect Auditing," *Public Finance*, 51, 473–489.

Becher, S. and U. Benoliel (2021): "Law in Books and Law in Action: The Readability of Privacy Policies and GDPR," in *Consumer Law and Economics*, ed. by K. Mathis and A. Tor, Cham, Switzerland: Springer, 179–204.

Bellstam, G., S. Bhagat, and J. A. Cookson (2021): "A Text-Based Analysis of Corporate Innovation," *Management Science*, 67, 4004–4031.

Benoit, K., K. Munger, and A. Spirling (2019): "Measuring and Explaining Political Sophistication Through Textual Complexity," *American Journal of Political Science*, 63, 491–508.

Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018): "quanteda: An R Package for the Quantitative Analysis of Textual Data," *Journal of Open Source Software*, 3, 774–777.

Bersch, J., S. Gottschalk, B. Müller, and M. Niefert (2014): "The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany," ZEW Discussion Paper 14-104, ZEW – Leibniz Centre for European Economic Research, Mannheim, Germany.

Björnson, C.-H. (1968): "Läsbarhet [Readability]," *Stockholm: Liber*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Bradley, R. A. and M. E. Terry (1952): "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, 39, 324–345.

Brody, S. and N. Elhadad (2010): "An Unsupervised Aspect-Sentiment Model for Online Reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Association for Computational Linguistics, 804–812.

Chen, C., C. B. Frey, and G. Presidente (2022): "Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally," The Oxford Martin Working Paper Series on Technological and Economic Change Working Paper No. 2022-1, Oxford Martin School, Oxford, UK.

Degeling, M., C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz (2019): "We Value Your Privacy … Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy," *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS'19)*.

Dresher, M. (1962): *A Sampling Inspection Problem in Arms Control Agreements: A Game-Theoretic Analysis*, Santa Monica, Cal.: RAND Corporation.

European Commission (2019): "Data Protection Rules as a Trust-Enabler in the EU and Beyond – Taking Stock," Communication from the Commission to the European Parliament and the Council, European Commission.

Fellingham, J. C. and P. Newman (1985): "Strategic Considerations in Auditing," *Accounting Review*, 60, 634–650.

Flesch, R. (1948): "A New Readability Yardstick," *Journal of Applied Psychology*, 32, 221–233.

Frankenreiter, J. (2022): "Cost-Based California Effects," *Yale Journal on Regulation*, 39, 1155–1217.

Ganglmair, B. and M. I. Wardlaw (2017): "Complexity, Standardization, and the Design of Loan Agreements," Unpublished manuscript, University of Georgia, available at https://ssrn.com/abstract=2952567.

Geer, D., R. Bace, P. Gutmann, P. Metzger, C. P. Pfleeger, J. S. Quarterman, and B. Schneier (2003): "Cyber*In*security: The Cost of Monopoly," Report, Computer & Communications Industry Association, available at http://www.ccianet.org/papers/cyberinsecurity.pdf.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574.

Goldberg, S. G., G. A. Johnson, and S. K. Shriver (2024): "Regulating Privacy Online: An Economic Evaluation of the GDPR," *American Economic Journal: Economic Policy*, 16, 325–58.

Graetz, M., J. Reinganum, and L. Wilde (1986): "The Tax Compliance Game: Toward an Interactive Theory of Law Enforcement," *Journal of Law, Economics, and Organization*, 2, 1–32.

Greenberg, J. (1984): "Avoiding Tax Avoidance: A (Repeated) Game-Theoretic Approach," *Journal of Economic Theory*, 32, 1–13.

Griffiths, T. L. and M. Steyvers (2004): "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101, 5228–5245.

Grün, B. and K. Hornik (2011): "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, 40, 1–30.

Heyes, A. G. (1994): "Environmental Enforcement when 'Inspectability' is Endogenous: A Model with Overshooting Properties," *Environmental and Resource Economics*, 4, 479–494.

Jensen, C. and C. Potts (2004): "Privacy Policies as Decision-Making Tools," in *Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI '04)*, ed. by E. Dykstra-Erickson and M. Tscheligi, New York, N.Y.: ACM, 471–478.

Johnson, G. (forthcoming): "Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond," in *The Economics of Privacy*, ed. by A. Goldfarb and C. Tucker, Chicago, Ill.: University of Chicago Press.

Johnson, G., S. Shriver, and S. Goldberg (2023): "Privacy and Market Concentration: Intended and Unintended Consequences of the GDPR," *Management Science*, 69, 5695–5721.

Kinne, J. and J. Axenbeck (2019): "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," ZEW Discussion Paper 18-033, ZEW – Leibniz Centre for European Economic Research, Mannheim.

Koski, H. and N. Valmari (2020): "Short-Term Impacts of the GDPR on Firm Performance," ETLA Working Papers 77, ETLA Economic Research.

Koutroumpis, P., F. Ravasan, and T. Tarannum (2022): "Under Investment in Cyber Skills and Data Protection Enforcement: Evidence from Activity Logs of the UK Information Commissioner's Office," Available at SSRN: https://ssrn.com/abstract=4179601.

Laffont, J.-J. (1994): "The New Economics of Regulation Ten Years After," *Econometrica*, 62, 507–537.

——— (2005): *Regulation and Development*, Cambrdge, UK: Cambridge University Press.

Larsen, V. H. and L. A. Thorsrud (2019): "The Value of News for Economic Developments," *Journal of Econometrics*, 210, 203–218.

Lin, N. and M. Osnabrügge (2018): "Making Comprehensible Speeches When Your Constituents Need It," *Research and Politics*, 5, 1–8.

Linden, T., R. Khandelwal, H. Harkous, and K. Fawaz (2020): "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, 2020, 47–64.

Livermore, M. A., A. Riddell, and D. Rockmore (2017): "The U.S. Supreme Court and the Judicial Genre," *Arizona Law Review*, 59, 837–901.

Loughran, T. and B. McDonald (2016): "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 54, 1187–1230.

Macho-Stadler, I. and D. Perez-Castrillo (2006): "Optimal Enforcement Policy and Firms' Emissions and Compliance with Environmental Taxes," *Journal of Environmental Economics and Management*, 51, 110–131.

Marcus, A. A. (1988): "Responses to Externally Induced Innovation: Their Effects on Organizational Performance," *Strategic Management Journal*, 9, 387–402.

Marotta-Wurgler, F. (2007): "What's in a Standard Form Contract? An Empirical Analysis of Software License Agreements," *Journal of Empirical Legal Studies*, 4, 677–713.

——— (2008): "Competition and the Quality of Standard Form Contracts: The Case of Software License Agreements," *Journal of Empirical Legal Studies*, 5, 447–475.

McCallum, A., X. Want, and A. Corrada-Emmanuel (2007): "Topic and Role Diversity in Social Networks with Experiments on Enron and Academic Email," *Journal of Artificial Intelligence Research*, 30, 249–272.

Milne, G. R., M. J. Culnan, and H. Greene (2006): "A Longitudinal Assessment of Online Privacy Notice Readability," *Journal of Public Policy & Marketing*, 25, 238–249.

Peukert, C., S. Bechtold, M. Batikas, and T. Kretschmer (2022): "Regulatory Spillovers and Data Governance: Evidence from the GDPR," *Marketing Science*, 41, 318–340.

Ruckman, K. and I. P. McCarthy (2017): "Why Do Some Patents Get Licensed While Others Do Not?" *Industrial and Corporate Change*, 26, 667–688.

Stern, J. (2000): "Electricity and Telecommunications Regulatory Institutions in Small and Developing Countries," *Utilities Policy*, 9, 131–157.

Wagner, I. (2023): "Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021," *ACM Transactions on Privacy and Security*, 26, 1–32.

Whittaker, J. A. (2003): "No Clear Answers on Monoculture Issues," *IEEE Security & Privacy*, 1, 18–19.

Wojahn, O., S. Geister, and J. Richter (2015): "The Impact of Analyst Report Complexity on Trading Decisions in an Experimental Setting," *Journal of Behavioral and Experimental Finance*, 7, 29–32.

Yuan, B. and J. Li (2019): "The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation," *International Journal of Environmental Research and Public Health*, 16, 1–15.

ZEW (2017): "ZEW Business Survey in the Information Economy, December 2017," ZEW – Leibniz Centre for European Economic Research, Mannheim, www.zew.de/WS380-1.

A Appendix: Formal Proofs

Proof of Proposition 1 The game is solved by iterated dominance. First, $a_{d,r}$ is the regulator's (weakly) dominant strategy: Given Assumption 2, the regulator strictly prefers to play $a_{d,r}$ if the firm chooses $(0,0)$, $(d,0)$, or $(0,r)$. Moreover, the regulator is indifferent (between all its actions) if the firm plays (d,r). By iterated dominance, in equilibrium, the firm must play a best response to the regulator's dominant strategy. It is sufficient to compare the payoffs of the firm when the regulator plays $a_{d,r}$. First, both $(0,0)$ and $(0,r)$ are dominated by $(d,0)$ in the reduced game. Strategy $(d,0)$ dominating $(0,r)$ follows once again from Assumption 2 where $\pi_d > \pi_r$ implies $(1-\pi_r)(v-k) > (1-\pi_d)(v-k)$. To see why $(d,0)$ dominates $(0,0)$, note that

$$(1-\pi_r)(v-k) > (1-\pi_d)(1-\pi_r)v \quad \Longleftrightarrow \quad k < \pi_d v.$$

The right-hand side holds by Assumptions 1 ($0 < k < \frac{v}{2}$) and 2 ($\pi_d > \pi_r > \frac{1}{2}$). Last, depending on the value of k, the firm chooses either $(d,0)$ or (d,r). It holds:

$$v - 2k > (1-\pi_r)(v-k) \quad \Longleftrightarrow \quad k < \frac{\pi_r}{1+\pi_r} v = k^u.$$

Proof of Proposition 2 No pure strategy equilibria can exist when the regulator is constrained. The regulator's undominated strategies are a_d and a_r; none of the firm's strategies are dominated. Suppose the regulator played a_j with probability one. The firm best response would then be to play $(d,0)$ if $j = d$, or $(0,r)$ if $j = r$. Then, the regulator would want to deviate from his strategy. We look for mixed strategy equilibria: Each agent plays

their undominated strategies with some probability in a way that makes the other indifferent between their undominated strategies.

To make the regulator indifferent between a_d and a_r, the firm can play $(d,0)$ with probability p_d, $(0,r)$ with probability p_r, $(0,0)$ with probability $1-p_d-p_r$. Alternatively, she can play (d,r) with probability one. Notice that no mixed strategy involving (d,r) can exist since the regulator's best response would be to optimally reply to the other strategy with probability one, which would make the firm want to deviate. We find p_d, p_r that satisfy:

$$(1-p_d-p_r)(1-\pi_d)(-\gamma)-p_d\gamma-p_r(1-\pi_d)\gamma=(1-p_d-p_r)(1-\pi_r)(-\gamma)-p_d(1-\pi_r)\gamma-p_r\gamma.$$

That is:

$$p_r\in\left[0,\frac{\pi_r}{\pi_d+\pi_r}\right],\quad p_d=\frac{\pi_d-(1-p_r)\pi_r}{\pi_d},\quad \text{and}\quad 1-p_d-p_r=(1-p_r)\frac{\pi_r}{\pi_d}-p_r$$

Since $p_d>0$, the regulator must make the firm indifferent between $(d,0)$ and either $(0,r)$ or $(0,0)$.

Suppose first that the regulator wanted to make the firm indifferent between $(d,0)$ and $(0,r)$; he must play a_d with probability $p^r_{a_d}$ that solves:

$$p^r_{a_d}(v-k)+\left(1-p^r_{a_d}\right)(1-\pi_r)(v-k)=p^r_{a_d}(1-\pi_d)(v-k)+\left(1-p^r_{a_d}\right)(v-k),$$

or:

$$p^r_{a_d}=\frac{\pi_r}{\pi_d+\pi_r}.$$

Suppose now that the regulator wanted to make the firm indifferent between $(d,0)$ and $(0,0)$ instead; then, the regulator plays a_d with probability $p^0_{a_d}$ that solves:

$$p^0_{a_d}(v-k)+\left(1-p^0_{a_d}\right)(1-\pi_r)(v-k)=p^0_{a_d}(1-\pi_d)v+\left(1-p^0_{a_d}\right)(1-\pi_r)v,$$

or:

$$p^0_{a_d}=\frac{(1-\pi_r)k}{\pi_d v-\pi_r k}.$$

With these probabilities, we have three candidate equilibria in which both players mix between two strategies; further, we must check when, if ever, the firm wants to deviate to (d,r). To do so, we obtain the utility of the firm when she mixes between $(d,0)$ and either $(0,r)$ or $(0,0)$ to determine which mixed equilibrium would emerge given parameters k, π_d, and π_r. Then, we compare the resulting utilities with the utility of full compliance, $v-2k$.

Suppose first that the regulator played $p^r_{a_d}=\frac{\pi_r}{\pi_d+\pi_r}$: We check for which parameters

playing the firm does not want to deviate from the corresponding mixed strategy that would form an equilibrium, that is, mixing between $(d,0)$ and $(0,r)$ according to $p_d = \frac{\pi_d}{\pi_d+\pi_r}$. By plugging in $p^r_{a_d}$ in the expected utility of the firm under the various strategies, we obtain:

$$\begin{aligned} E[(d,0)]|_{p^r_{a_d}} = E[(0,r)]|_{p^r_{a_d}} &= \frac{\pi_r}{\pi_d+\pi_r}(v-k) + \left(1-\frac{\pi_r}{\pi_d+\pi_r}\right)(1-\pi_r)(v-k) \\ &= \frac{(v-k)[\pi_d(1-\pi_r)+\pi_r]}{\pi_d+\pi_r} \\ E[(0,0)]|_{p^r_{a_d}} &= \frac{\pi_r}{\pi_d+\pi_r}(1-\pi_d)v + \left(1-\frac{\pi_r}{\pi_d+\pi_r}\right)(1-\pi_r)v \\ &= \frac{v[\pi_d+\pi_r-2\pi_d\pi_r]}{\pi_d+\pi_r} \end{aligned}$$

Direct comparison reveals that, subject to the regulator mixing in $p^r_{a_d}$, $E[(d,0)]|_{p^r_{a_d}} = E[(0,r)]|_{p^r_{a_d}} > E[(0,0)]|_{p^r_{a_d}}$ if and only if one of two conditions are satisfied:

$$\begin{gathered} \frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d-1}\right) \leq \frac{2}{3} \quad \wedge \quad k < \frac{\pi_r\pi_d}{\pi_r+\pi_d-\pi_r\pi_d}v, \\ \frac{\pi_d}{3\pi_d-1} < \pi_r < \pi_d < 1 \quad \wedge \quad k < \frac{v}{2}. \end{gathered}$$

Furthermore, $E[(d,0)]|_{p^r_{a_d}} > E[(d,r)]|_{p^r_{a_d}} = v-2k$ if and only if:

$$k < \frac{\pi_r\pi_d}{\pi_r+\pi_d+\pi_r\pi_d}v.$$

Suppose now that the regulator played $p^0_{a_d} = \frac{(1-\pi_r)k}{\pi_d v-\pi_r k}$; we repeat the same exercise to check for which parameters, in equilibrium, the firm mixes between $(d,0)$ and $(0,0)$:

$$\begin{aligned} E[(d,0)]|_{p^0_{a_d}} = E[(0,0)]|_{p^0_{a_d}} &= \frac{(1-\pi_r)k}{\pi_d v-\pi_r k}(1-\pi_d)v + \left(1-\frac{(1-\pi_r)k}{\pi_d v-\pi_r k}\right)(1-\pi_r)v \\ &= \frac{(v-k)v\pi_d(1-\pi_r)}{\pi_d v-\pi_r k} \\ E[(0,r)]|_{p^0_{a_d}} &= \frac{(1-\pi_r)k}{\pi_d v-\pi_r k}(1-\pi_d)(v-k) + \left(1-\frac{(1-\pi_r)k}{\pi_d v-\pi_r k}\right)(v-k) \\ &= \frac{(v-k)[k\pi_d-v\pi_r+k\pi_r(1-\pi_d)]}{\pi_d v-\pi_r k} \end{aligned}$$

Again by direct comparison, it holds that $E[(d,0)]|_{p^0_{a_d}} = E[(0,0)]|_{p^0_{a_d}} > E[(0,r)]|_{p^0_{a_d}}$ if

and only if:

$$\frac{1}{2} < \pi_r < \min\left(\pi_d, \frac{\pi_d}{3\pi_d - 1}\right) \leq \frac{2}{3} \quad \wedge \quad k > \frac{\pi_r \pi_d}{\pi_r + \pi_d - \pi_r \pi_d} v.$$

Combining the conditions above, we immediately obtain the equilibria described in bullet points 2, 3, and 4 of Proposition 2. No other equilibria can exist for $k > \underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v$ since no other deviations are available to the firm and no other undominated strategy is available to the regulator. For $k < \underline{k}$, there cannot be any equilibrium in which the firm plays $(0,0)$ with positive probability from the above calculations. We must then only compare pure compliance, (d,r), and mixing between $(d,0)$, $(0,r)$.

Pure compliance dominates the latter for $k < \underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} v$. This holds when the regulator mixes according to $p^r_{a_d}$. Moreover, infinite payoff equilibria exist for $k < \underline{k}$. In these equilibria, the firm plays (d,r) with probability one; the regulator mixes between a_d and a_r with different probabilities. To characterize them all, we find the highest and lowest probability of playing a_d as a function of k that makes the firm weakly better off playing (d,r) than deviating:

$$\begin{aligned} E[(d,0)]|_{p^r_{a_d}} &= p^r_{a_d}(v-k) + \left(1 - p^r_{a_d}\right)(1-\pi_r)(v-k) \\ E[(0,r)]|_{p^r_{a_d}} &= p^r_{a_d}(1-\pi_d)(v-k) + \left(1 - p^r_{a_d}\right)(v-k) \end{aligned}$$

We are interested in $p^r_{a_d}$ that makes the firm weakly better off selecting (d,r) over either $(d,0)$ or $(0,r)$. The former case arises when: $v - 2k > p^r_{a_d}(v-k) + \left(1 - p^r_{a_d}\right)(1-\pi_r)(v-k)$, which is equivalent to:

$$p^r_{a_d} \geq \overline{p}^r_{a_d} = \frac{v[\pi_d(3-\pi_r) + \pi_r]\pi_r - k(3-\pi_r)(\pi_d + \pi_r + \pi_c \pi_r)}{v[\pi_d(3-\pi_r) + \pi_r]\pi_r - k(2-\pi_r)(\pi_d + \pi_r + \pi_c \pi_r)}$$

The latter arises when: $v - 2k \geq p^r_{a_d}(1-\pi_d)(v-k) + \left(1 - p^r_{a_d}\right)(v-k)$, which is equivalent to:

$$p^r_{a_d} \geq \underline{p}^r_{a_d} = \frac{k}{(v-k)\pi_d}$$

For all $k \in (0, \underline{k})$, then, any strategy in which the regulator plays a_d with probability $p^r_{a_d} \in [\underline{p}^r_{a_d}, \overline{p}^r_{a_d}]$ induces the firm to play (d,r) with probability one. These are the infinite payoff equivalent equilibria referred to in bullet point 1 of Proposition 2.

No other equilibria exist: suppose a mixed-strategy equilibrium exists that involves the firm playing (d,r) with some positive probability different from one. Then, the regulator will

optimally play the best response to the other action with probability one, to which the firm's best response is something other than (d, r). Because there is no mixed strategy employed by the firm that makes the regulator indifferent between a_d and a_r where $p_d = 0$, this exhaust all candidate equilibria.

Proof of Proposition 3 The proof of Proposition 3 follows immediately from the proofs of Proposition 1 and Proposition 2 if it holds that $0 < \underline{k} < k^u < \overline{k} < \frac{v}{2}$. The outer conditions are satisfied under Assumptions 1 and 2. It is then sufficient to show that:

$$\underline{k} = \frac{\pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d} < \frac{\pi_r}{1 + \pi_r} = k^u$$

and:

$$k^u = \frac{\pi_r}{1 + \pi_r} < \frac{\pi_r \pi_d}{\pi_r + \pi_d - \pi_r \pi_d} = \overline{k}$$

The former is equivalent to $1 > \frac{\pi_d + \pi_r \pi_d}{\pi_r + \pi_d + \pi_r \pi_d}$, the latter is equivalent to $\pi_d > 1 - \pi_d$. Both conditions are satisfied under Assumption 2 as well.

Online Appendix

— Not for publication —

Regulatory Compliance with Limited Enforceability: Evidence from Privacy Policies

Bernhard Ganglmair* Julia Krämer† Jacopo Gambato‡

July 8, 2024

Abstract

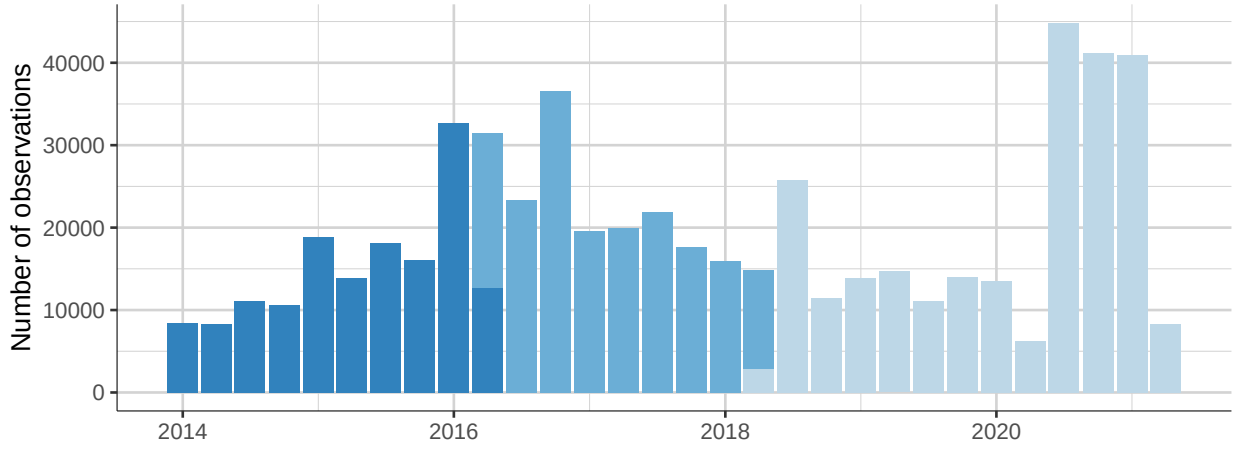
This is the online appendix (not for publication) of the paper "Regulatory Compliance with Limited Enforceability: Evidence from Privacy Policies" by B. Ganglmair ⓡ J. Krämer ⓡ J. Gambato.

Contents

*University of Mannheim and ZEW Mannheim, b.ganglmair@gmail.com
†Erasmus University Rotterdam, j.k.kramer@law.eur.nl
‡University of Mannheim and ZEW Mannheim, ja.gambato@gmail.com

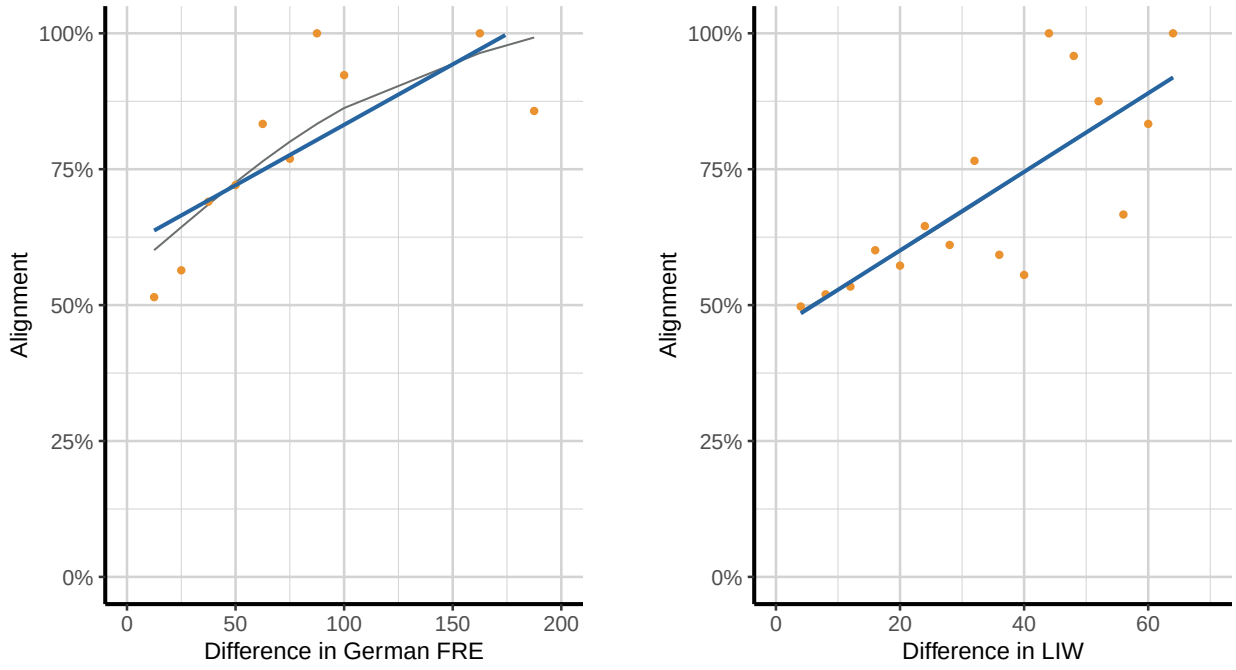
B Additional Figures and Tables

Figure B.1: Observations by Quarter



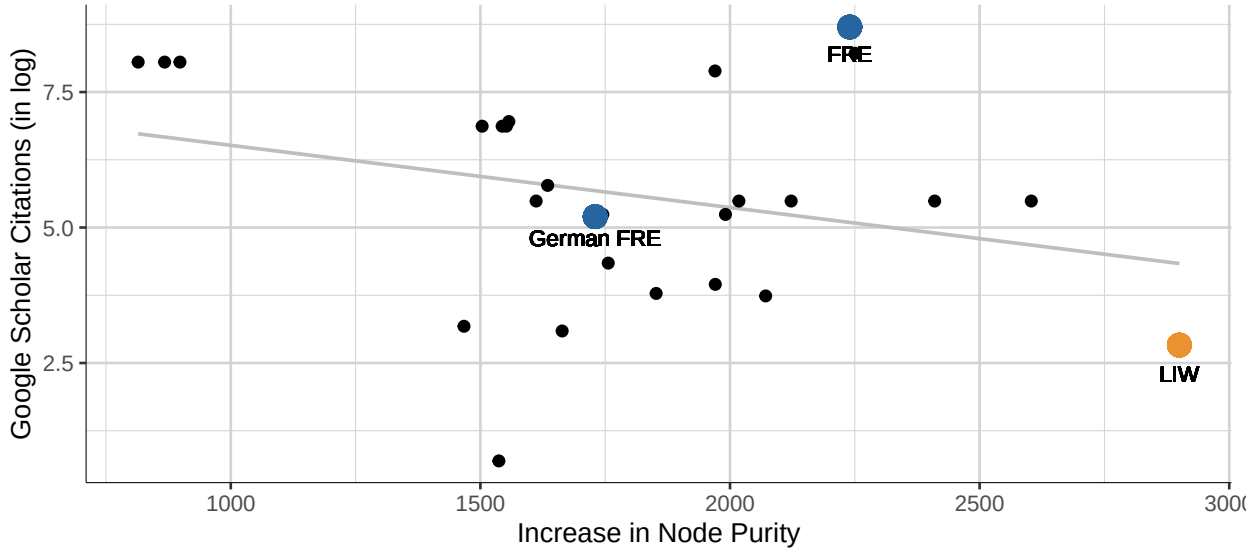
Notes: The figure presents the number of observations (one policy per firm) per quarter for our estimation sample (Q1 2014 to Q2 2021). The different shades of blue indicate three time phases: pre-GDPR passage (May 4, 2016; Q2 2016), pre-GDPR enforcement (May 25, 2018, Q2 2018), and post-GDPR enforcement. In Q2 2016, and Q2 2018, we have observations in two phases (before and after the respective cut-off dates).

Figure B.2: Human vs. Factor-Based Assessment of Readability (German FRE and LIW)



Notes: The figures depict the percentage of text pairs for which the human assessments align with the ranking based on the text pairs' absolute differences in German FRE (LHS) and LIW (RHS). Values on the horizontal axis are binned for visual ease. Average alignment for each bin (dots); fitted spline (grey thin line); and linear fit (blue thick line). Dot size does not reflect the number of observations in each bin. An increase in the difference in German FRE of one standard deviation (5.64) increases the alignment by 1.25 percentage points (OLS coefficient: 0.0022, t-statistic: 3.27). An increase in the difference in LIW of one standard deviation (3.94) increases the alignment by 2.85 percentage points (OLS coefficient: 0.0072, t-statistic: 4.62).

Figure B.3: Ability to Predict Data vs. Popularity of Readability Scores



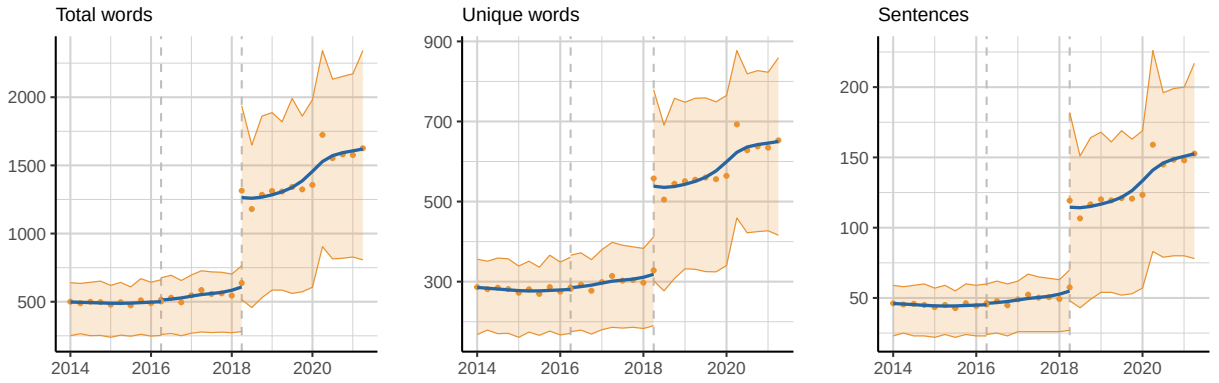
Notes: The figure depicts the performance of readability indices in predicting pair-wise comparisons (following Benoit et al. (2019)) on the horizontal axis and their popularity (Google Scholar Citations) on the vertical axis. We mark our preferred index, the LIW, (in orange), Flesch's Reading Ease Score (FRE) and Amstad's Verständlichkeitsindex (German FRE) (in blue) that has been used in the regulation of insurance contract language in the U.S. states of Florida, Massachusetts, Michigan, and Texas. An increase in Google Scholar Citations (in log) is associated with a decrease in Node Purity (OLS coefficient: -69.18, t-statistic: -1.43).

Table B.1: Disclosure-Related Terms

GDPR Article	**Terms**
13/14(1)(a)	`(?i)(verantwortliche|verantwortlich[a-z]+ f[u?]r die datenverarbeitung|kontakt|kontaktdaten)`
13/14(1)(b)	`(?i)(datenschutzbeauftragte|kontakt|kontaktdaten)`
13/14(1)(c)	`(?i)(art[.]?|arti[a-z]+|?)[ ]+6((?:.*){0,8})(ds[-]*g[-]*vo|datenschutzgrundverordnung|gdpr)`
13/14(1)(d)	`(?i)(zweck|rechtsgrundlage|datenverarbeitung|berechtigte[a-z]? interesse)`
13/14(1)(e)	`(?i)(empf[?a]nger)`
13/14(1)(f)	`(?i)(drittland|[u?]bertragung|usa|privacy[ -]shield|datenschutzschild|angemessenheitsbeschluss|schutzniveau)`
13/14(2)(a)	`(?i)(speicherdauer|dauer|speicherung|speicherfrist)`
13/14(2)(b)	`(?i)(recht auf auskunft|auskunftsrecht|auskunft)`
13/14(2)(c)	`(?i)(recht auf widerruf|einwilligung zu widerrufen|widerruf)`
13/14(2)(d)	`(?i)(recht auf beschwerde|beschwerderecht|aufsichtsbeh[?o]rde|landesbeauftragte[a-z]? f[?u]r den datenschutz|recht auf daten[?u]bertragbarkeit|daten[u?]bertragbarkeit)`
13/14(2)(e)	`(?i)(bereitstellung|vertraglich vorgeschrieben|vertrag)`
13/14(2)(f)	`(?i)(automatisierte entscheidungsfindung|profiling)`

Notes: This table summarizes the terms used to identify paragraphs that disclose information pertaining to Art. 13 or 14. The information to be disclosed as listed in Art. 13 is the same as in Art. 14.

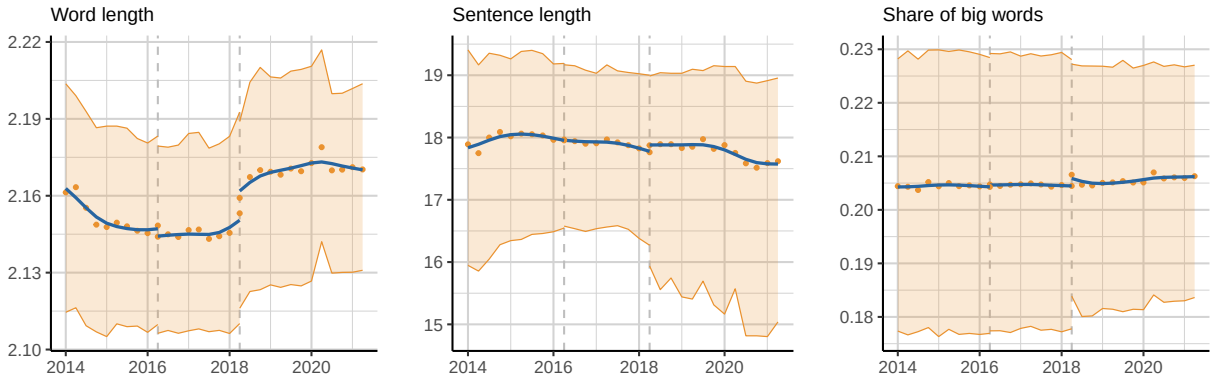
Figure B.4: Informational Volume



Notes: This figure presents quarterly averages of policy-level measures for informational volume (total word count, unique words, and number of sentences). Dots represent quarterly averages; the curves are fitted to the data (spline); the lower and upper bounds represent the 25th and 75th percentiles, respectively (interquartile range). The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018. The following table provides basic descriptive statistics:

	Mean	**Std.**	**Min**	**Max**
Total word count	917.21	839.2	61	5614
Unique words	419.55	270.93	22	2039
Number of sentences	84.24	75.18	2	569

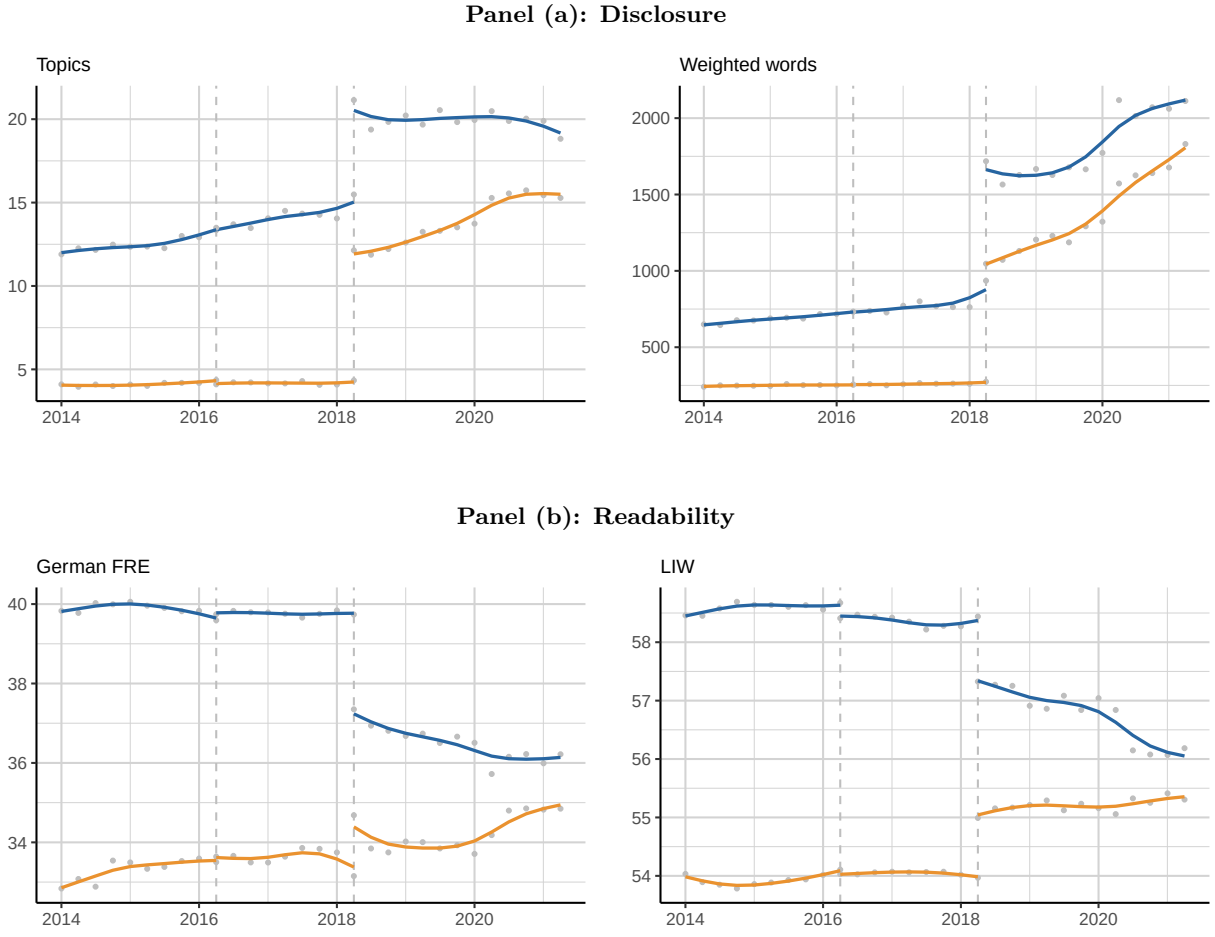
Figure B.5: Readability Components



Notes: This figure presents quarterly averages of policy-level measures for readability components (average word length, average sentence length, and share of words with five syllables or more). Dots represent quarterly averages; the curves are fitted to the data (spline); the lower and upper bounds represent the 25th and 75th percentiles, respectively (interquartile range). The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018. The following table provides basic descriptive statistics:

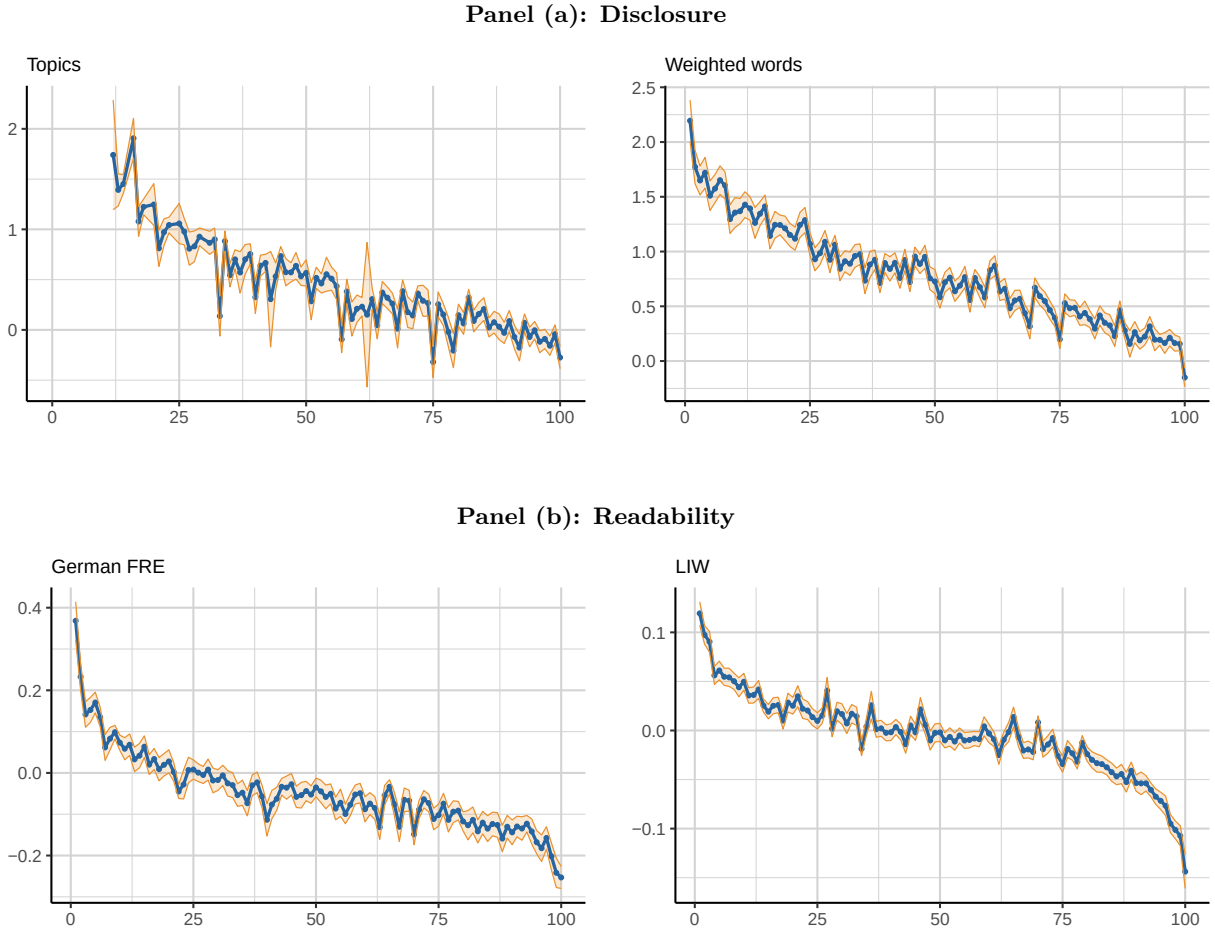
	Mean	**Std.**	**Min**	**Max**
Word length (in syllables)	2.16	0.07	1.4	3.4
Sentence length (in words)	17.84	3.26	4.2	222
Share of big words (5+ syllables)	0.21	0.04	0	0.5

Figure B.6: Disclosure and Readability by Pre-GDPR Exposure



Notes: This figure presents quarterly averages of policy-level measures for disclosure (panel (a)) and readability (panel (b)). Dots represent quarterly averages; the curves are fitted to the data (spline). The blue curve is for firms with pre-GDPR policies above the median of the respective variable; the orange for firms with pre-GDPR policies below the median. The vertical dashed lines indicate the GDPR passage in Q2 2016 and GDPR enforcement in Q2 2018.

Figure B.7: GDPR-Induced Changes by Pre-GDPR Percentiles



Notes: This figure presents the coefficients for the Post GDPR dummy for firms in a given percentile (on the horizontal axis) of the pre-GDPR distribution of the respective dependent variable. Regression results are from fixed-effects OLS regressions (firm FE and year FE). All dependent variables are in log. Additional control variables are HHI (as a measure of market concentration) and log Employees (as a measure of firm size). The orange shaded area depicts the 99% confidence interval.

Table B.2: Disclosure as the Topic-Weighted Information Volume

		Example 1			Example 2		
	Words	**Topic**	**Factor ϕ_k**	$\phi_k w_{c\|k}$	**Topic**	**Factor ϕ_k**	$\phi_k w_{c\|k}$
Paragraph 1	10	A	2.0	20	A	2.0	20
Paragraph 2	20	B	1.0	20	B	1.0	20
Paragraph 3	30	C	0.5	15	C	0.5	15
Paragraph 4	40	C	0.5	20	A	2.0	80
Total word count	**100**	**Disclosure (Ex. 1)**		**75**	**Disclosure (Ex. 2)**		**135**

Notes: This table illustrates the topic-weighted information volume, using a privacy policy with four paragraphs and their respective word counts. For the overall distinct-topic distribution, we assume $(0.25, 0.25, 0.50)$. For the distinct-topic distribution of disclosing paragraphs, we assume $(0.50, 0.25, 0.25)$. The topic factors are therefore $(\phi_A, \phi_B, \phi_C) = (2, 1, 0.5)$. The two examples differ in the distinct topic for Paragraph 4 ('C' in Example 1, 'A' in Example 2). The unweighted word count of the policy is 100. In Example 1, Paragraph 4 is unlikely a disclosing paragraph: the topic-weighted word count is 75. In Example 2, Paragraph 4 is likely a disclosing paragraph: the topic-weighted word count is 135.

Table B.3: Comparison of Readability with Other Text Corpora

	Obs.	**Word length**	**Sentence length**	**Big words**	**German FRE**	**LIW**
Privacy policy panel	585329	2.16	17.84	0.21	35.98	56.13
		(0.07)	(3.26)	(0.04)	(5.64)	(3.94)
Simple-language news	1594	1.74	10.74	0.04	67.5	39.11
(nachrichtenleicht.de)		(0.12)	(1.8)	(0.03)	(7.28)	(5.42)
Speeches and statements:	1128	1.83	18.16	0.3	54.84	48.05
Angela Merkel		(0.07)	(2.3)	(0.03)	(4.47)	(3.1)
Decisions by German Consti-	9358	1.96	16.35	0.15	49.27	50.17
tutional Court (BVerfG)		(0.09)	(2.91)	(0.03)	(6.75)	(4.91)
Wikipedia (German)	10000	1.9	20.63	0.12	48.48	53.51
		(0.2)	(14.48)	(0.04)	(18.23)	(15.48)
Wikipedia (English)	10000	1.71	19.78	0.05	60.33	47.8
		(0.16)	(6.57)	(0.03)	(11.58)	(9.31)
GDPR/DS-GVO (Wikipedia)	1	2.1	18.63	0.12	38.35	57.1
GDPR/DS-GVO (official)	1	2.24	40.39	0.18	8.83	81.39

Notes: We report text characteristics for our estimation sample and various German-language corpora. *Simple-language news* are all news articles published on nachrichtenleicht.de between November 2019 and June 2023. The *Speeches and statements* by Angela Merkel are from Barbaresi (2019). The *Decisions by the German Constitutional Court* are from Möllers et al. (2021). The Wikipedia pages are from a random sample of German pages and their English-language counterparts (accessed in June 2022).

Table B.4: Popularity of Readability Scores

	Google Scholar	
Readability score/index	**Search**	**Citations**
Flesch's Reading Ease Score	~25,000	6069
Gunning's Fog Index	~13,500	2669
Simple Measure of Gobbledygook (SMOG)	~10,600	3143
Lexile Measure	~5300	69
Anderson's Readability Index	4950	242
Automated Readability Index (ARI)	4400	323
Fry Readability	4210	1744
Flesch-Kincaid Readability Score	3990	3698
Simplified Automated Readibility Index	3190	323
Coleman's Readability Formula	2420	134
Coleman-Liau Index	2020	963
The Old Dale-Chall Readability Formula	1050	2473
Fucks' Stilcharakteristik	928	22
The New Dale-Chall Readability Formula	868	1246
Björnsson's Läsbarhetsindex (LIW/LIX)	684	17
Linsear Write	441	1049
Neue Wiener Sachtextformeln (1–4)	366	242
Easy Listening Formula	186	77
Atos Readability	154	2
Wheeler & Smith's Readability Measure	141	44
Farr-Jenkins-Paterson's Simplification of Flesch Reading Ease Score	113	326
EFLAW Readability	109	24
Amstad Verständlichkeitsindex (German FRE)	9	189
Coleman-Liau Estimated Cloze Percent	4	963
Dickes-Steiwer Index	4	42
Danielson-Bryan´s Readability Measure	3	52

Notes: The table reports the number of Google Scholar search results and the number of Google Scholar citations for a variety of readability scores and indices. Numbers are hand collected, accessed March 31, 2023.

Table B.5: GDPR-Induced Changes by Firm Size and Industry

Dependent variable (in log):	**Disclosure**		**Readability**	
	Topics	**Weighted words**	**German FRE**	**LIW**
	(1)	(2)	(3)	(4)
Panel (a): Size				
Micro firms	0.4751***	0.7409***	-0.0406***	-0.0040***
	(0.0091)	(0.0084)	(0.0020)	(0.0007)
Small/Medium sized firms	0.5184***	0.8180***	-0.0458***	-0.0040***
	(0.0106)	(0.0092)	(0.0024)	(0.0008)
Large firms	0.3715***	0.7743***	-0.0051	-0.0065**
	(0.0386)	(0.0298)	(0.0072)	(0.0027)
# Firm FE	65,854	65,863	65,859	65,863
R^2	0.698	0.784	0.627	0.650
Observations	413,099	413,249	413,154	413,249
Panel (b): Industry				
Industry: Agriculture/Mining	0.4409***	0.7021***	-0.0361***	-0.0045
	(0.0710)	(0.0526)	(0.0131)	(0.0042)
Industry: Manufacturing	0.4846***	0.7960***	-0.0547***	-0.0018
	(0.0193)	(0.0157)	(0.0043)	(0.0014)
Industry: Utilities	0.2810***	0.8754***	-0.0457***	-0.0024
	(0.0563)	(0.0370)	(0.0095)	(0.0035)
Industry: Construction	0.5638***	0.7757***	-0.0612***	-0.0027
	(0.0210)	(0.0190)	(0.0056)	(0.0018)
Industry: Trade	0.5001***	0.7709***	-0.0421***	-0.0055***
	(0.0143)	(0.0116)	(0.0029)	(0.0010)
Industry: Services	0.4858***	0.7744***	-0.0374***	-0.0041***
	(0.0090)	(0.0084)	(0.0020)	(0.0007)
# Firm FE	64,596	64,605	64,602	64,605
R^2	0.696	0.782	0.625	0.648
Observations	409,332	409,482	409,388	409,482

Notes: We report the results of fixed-effects OLS regressions (firm FE and year FE). Dependent variables are measures of disclosure (topics and weighted words) and readability (German FRE and LIW). In panel (a), we report the coefficient for the Post GDPR dummy by firm size category; in panel (b), we report the coefficient for the Post GDPR dummy by industry. For definitions of size category and industry, see the table notes in Table 2. All dependent variables are in log. In panel (b), additional control variables (not reported) are HHI (as a measure of market concentration) and log Employees (as a measure of firm size). Signif. levels: ***: 0.01, **: 0.05, *: 0.1.

C LDA Topic Models

Probabilistic topic models uncover the latent topical structure of a document by analyzing the co-occurrence of tokens (i.e., words, terms, or phrases) used in the document. The underlying idea is that authors first decide which topics to cover before drafting the document. A document thus becomes a collection of multiple topics. The LDA topic model (Blei et al., 2003) describes such a topic k as a *per-topic word distribution* $\vec{\beta}_k$ over the vocabulary of N tokens (i.e., a $(1 \times N)$ vector for each topic k). Moreover, for our corpus of privacy policies, holding documents that cover K topics, each document d will exhibit these K topics with different proportions according to a *per-document topic distribution* $\vec{\theta}_d$. The data we observe are the documents in a text corpus $\mathcal{D}$ and the tokens $\vec{w}_d$ used in each document. The topics, however, are not observed. We apply LDA to reverse this process of topic generation and automatically discover the latent topical structure. This means that we obtain estimates for $\vec{\beta}_k$ (for $k = 1, \ldots, K$) and $\vec{\theta}_d$ (for $d \in \mathcal{D}$). In Figure C.1, we provide a stylized depiction of this process for two documents ($d = 1, 2$) and three topics ($k = 1, 2, 3$).

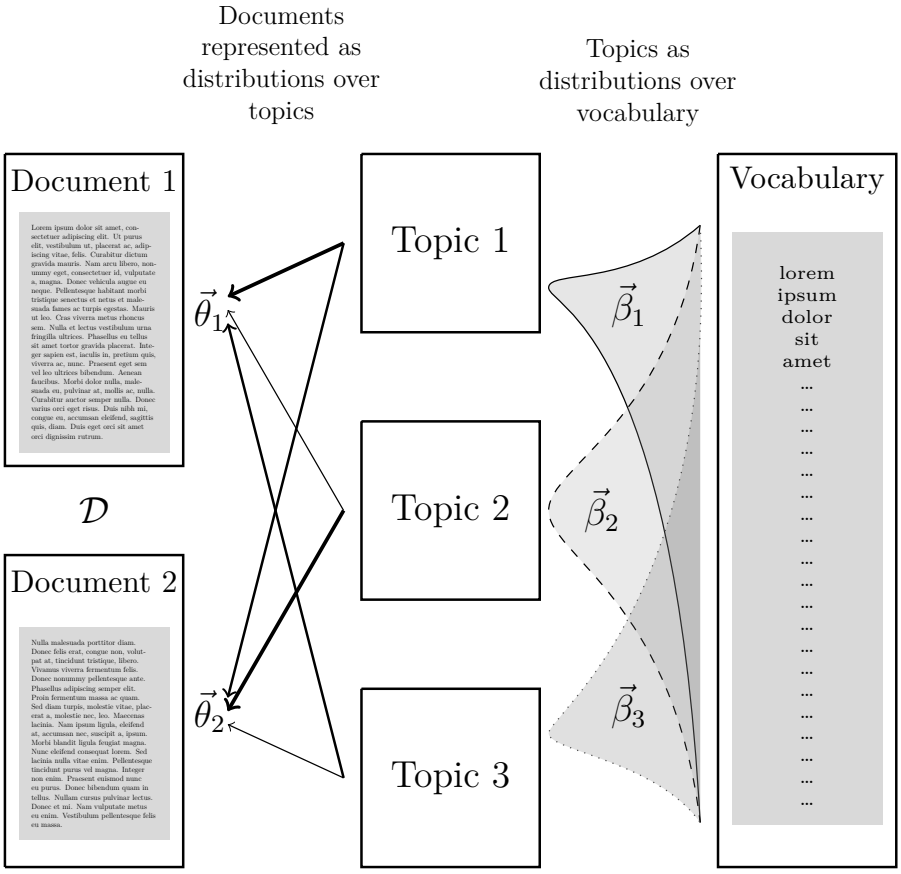
To obtain the number of distinct topics, we follow a two-step approach (similar to Ganglmair and Wardlaw (2017)). We use the `topicmodels` package (Grün and Hornik, 2011) to estimate topic models.

1. To obtain the main topic for a given paragraph, we first perform standard pre-processing steps and define tokens as unigrams. We then estimate the topic model with $K = 50$ topics on the corpus of paragraphs, following Brody and Elhadad (2010). Estimating topic models is computationally intensive. We, therefore, limit the number of tokens to 3,000 (by frequency) and estimate our topic model on paragraphs of 5,000 randomly drawn policies, predicting the topic assignments for *all* paragraphs in our corpus. The number of topics K is chosen to balance the additional granularity of higher K with the computational burden.

   For each paragraph c in a policy d, we obtain the per-document topic distribution $\vec{\theta}_{cd} = \left(\theta_{1|cd}, \ldots, \theta_{K|cd}\right)$ over K topics, with $\sum_{k=1}^{K} \theta_{k|cd} = 1$. Each $\theta_{k|cd}$ represents the weight with which a topic k is covered in a paragraph c.

2. We assume that, in practice, each paragraph was written to cover a single topic. We call this topic the *main topic* k_{cd} of a paragraph and define it as the topic with the highest topic density $\theta_{k|cd}$, so that $k_{cd} = \arg\max_{k=1,\ldots,K} \theta_{k|cd}$. We obtain a $(1 \times K)$ vector $\vec{k}_d$ with $K = 50$ elements, each being equal to 1 if topic k is a main topic at least once, and zero otherwise. For a count of main topics for each policy, the union of main topics gives us the set of distinct topics that are a main topic for at least one paragraph (i.e., the number of main topics is then $\sum_{k=1}^{K} \vec{k}_d$.)

Figure C.1: Probabilistic Topic Models



Source: Ganglmair and Wardlaw (2017)

D Construction of the Privacy Policy Panel

The construction of the Privacy Policy Panel begins with an initial sample of 570,000 firm IDs and URLs of the firms' websites containing the privacy policies, taken from the 2019 wave of the Mannheim Web Panel (Kinne and Axenbeck, 2019). We select the privacy policy pages by sampling those URLs from the web panel that contain the term "datenschutz" (the German word for data protection) or "privacy". This section discusses the individual steps of our data construction. The data construction process is as follows:

1. From the 2019 wave of the Mannheim Web Panel, we determine the most common URL patterns used by firms to store their privacy policies. The resulting list (Table D.1) cumulatively takes up 52% of all patterns in the referenced wave. For each firm, we extract the registered URL and look for the archived correspondent page for each quarter between 2014 and Q2 of 2021 on the Internet Archive through the Wayback Machine. The Wayback Machine is a part of the Internet Archive, an organization

founded in 1996 with the intent to preserve the history of the Internet by archiving important websites. The organization repeatedly visits websites and stores snapshots of their content for potential future use. A user accessing the Wayback Machine can then search for a specific website and "visit" its historical versions, which can then be scraped and collected as any real-time site would.

2. For all available pages of a given URL, we download and store the full HTML page and record the respective date. If, in a given quarter, a page is not found for that specific URL, the scraper circles through the list of common-pattern URLs. If this second step does not recover a page, we set the observation to missing. If, in a given quarter, we find multiple pages in the Internet Archive, we store the first page recorded (by date) in that quarter. The reason a page may not be found is that a website may have moved the location of its privacy policy over time, e.g., from `/datenschutz/` to `/datenschutz.html`.

Table D.1: Data Construction: Most Common URL Patterns

No.	Pattern	Frequency	Share
1	datenschutz/	109642	0.17
2	datenschutz.html	68805	0.28
3	datenschutz	66367	0.38
4	datenschutzerklaerung/	27113	0.43
5	j/privacy	25163	0.47
6	datenschutz.php	12987	0.49
7	datenschutzerklaerung	11554	0.51
8	datenschutzerklaerung.html	9453	0.52

3. From each downloaded page (in HTML format), we extract the text of the respective privacy policy. We use a simple parser (manually calibrated and optimized using a viewer app) to capture the relevant text portions while ignoring other portions of the HTML pages (such as headers, pictures, or external links). The parser relies on the *readability-lxml* package in Python.[1] An adapted version of the *doc.summary()* function of this package extracts text from the HTML page. We also delete empty pages or error pages from our sample.

The Internet Archive restricts what we are able to capture by what was visited and saved in the Wayback Machine in the first place. It is important that we briefly touch upon the way this process takes place. The Archive uses a series of webcrawlers (both directly

[1] https://github.com/buriy/python-readability

controlled by the Archive and by third parties, e.g. Alexa crawls) to visit a large amount of websites and save the content of the pages they visit. The webcrawlers are programmed in such a way that, starting from any page, they follow any links contained in it to enrich the collection.[2] The active collection by webcrawlers happens in programmed waves or "crawls" starting from a list of URLs as initial targets.[3]

The data contained in the Wayback Machine has some inherent bias. Crawlers follow links contained on a visited page: the resulting data might suffer from over-representation of large, public, and well-connected sites compared to smaller economic agents with lower visibility. The end result of a crawl systematically depends on its starting point: some websites might not appear in different crawls. Older firms might furthermore be over-represented since crawlers often revisit sites already seen in the past. At the same time, more recent crawls appear to be more thorough and widespread than older ones. All of the above shape the composition of our final estimation sample. Overall, we expect our sample to be biased towards larger firms (because they are more likely to be mentioned on other sites) and older firms (because the Wayback Machine might occasionally revisit already stored sites). Furthermore, we expect a bias toward more consumer-facing industries and especially companies where a website is a part of their core product.

References

Barbaresi, A. (2019): "German Political Speeches Corpus (Version v4.2019) [Data set]," Zenodo. https://doi.org/10.5281/zenodo.3611246.

Benoit, K., K. Munger, and A. Spirling (2019): "Measuring and Explaining Political Sophistication Through Textual Complexity," *American Journal of Political Science*, 63, 491–508.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Brody, S. and N. Elhadad (2010): "An Unsupervised Aspect-Sentiment Model for Online Reviews," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, Association for Computational Linguistics, 804–812.

Ganglmair, B. and M. I. Wardlaw (2017): "Complexity, Standardization, and the

[2]This process is usually combined with policies that control and limit the number of links a crawler will follow on a given page, i.e., how "deep" it will go into any single page. This prevents crawlers from ending up in an infinite loop or "getting lost" in one site.

[3]A list of these crawls and respective starting URLs can be found at https://archive.org/details/web.

Design of Loan Agreements," Unpublished manuscript, University of Georgia, available at https://ssrn.com/abstract=2952567.

Grün, B. and K. Hornik (2011): "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, 40, 1–30.

Kinne, J. and J. Axenbeck (2019): "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," ZEW Discussion Paper 18-033, ZEW – Leibniz Centre for European Economic Research, Mannheim.

Möllers, C., A. Shadrova, and L. Wendel (2021): "BVerfGE-Korpus (1.0) [Data set]," Zenodo. https://doi.org/10.5281/zenodo.4551408.