Electronic Markets (2023) 33:33
https://doi.org/10.1007/s12525-023-00622-x

RESEARCH PAPER

Promise not fulfilled: FinTech, data privacy, and the GDPR

Gregor Dorfleitner[1,2] · Lars Hornuf[3,4] · Julia Kreppmeier[1]

Received: 8 March 2022 / Accepted: 6 January 2023 / Published online: 20 July 2023

Abstract
This article analyzes how the General Data Protection Regulation (GDPR) has affected the privacy practices of FinTech firms. We study the content of 276 privacy statements respectively before and after the GDPR became binding. Using text analysis methods, we find that the readability of the privacy statements has decreased. The texts of privacy statements have become longer and use more standardized language, resulting in worse user comprehension. This calls into question whether the GDPR has achieved its original goal—the protection of natural persons regarding the transparent processing of personal data. We also link the content of the privacy statements to FinTech-specific determinants. Before the GDPR became binding, more external investors and a higher legal capital were related to a higher quantity of data processed and more transparency, but not thereafter. Finally, we document mimicking behavior among FinTech industry peers with regard to the data processed and transparency.

Keywords Data privacy · FinTech · General Data Protection Regulation · Privacy statement · Textual analysis · Financial technology

JEL Classification: K200 · L810 · M13

Introduction

Data have become a critical resource for many business models as a result of digitalization and globalization. Individuals disclose personal information intentionally and unintentionally over the Internet and when using their smartphones (Lindgreen, 2018; World Bank, 2021). Because of the international location of servers and cloud-computing services, the processing of data often takes place under different jurisdictions and does not stop at national borders. On May 25, 2018, the General Data Protection Regulation (GDPR) became binding in the European Economic Area (EEA)[1] to address the increasing challenges of data security and privacy. The GDPR extends its territorial reach even outside the EEA if European data are involved. The financial sector and, in particular, the recently emerging Financial Technology (FinTech) industry process large amounts of sensitive data. Payment data, for example, can entail information about racial or ethnic origin, political opinions, religious beliefs, trade-union membership, health or sex life. The different FinTech business models, which frequently rely on artificial intelligence, big data, and cloud computing, thus represent an important and relevant industry to examine the impact of the GDPR on data privacy practices.

Companies are not required by law to have a privacy statement; however, they often comply with the requirement to inform their users (art. 13–15 GDPR), by publishing such statements, about the personal data they process. Therefore, privacy statements serve as research objects for many studies that analyze privacy. For example, Ramadorai et al. (2021) study a signalling model of firms engaging in data extraction. They analyze a sample of 4,078 privacy statements of

Responsible Editor: Rainer Alt

✉ Lars Hornuf
lars.hornuf@tu-dresden.de

1 Department of Finance, University of Regensburg, Regensburg, Germany

2 Hanken Centre for Accounting, Finance and Governance, Hanken School of Economics, Helsinki, Finland

3 Faculty of Business and Economics, Technische Universität Dresden, Dresden, Germany

4 CESifo Research Network, Munich, Germany

[1] Thus, it applies in the European Union (EU) and countries of the European Free Trade Association except Switzerland.

U.S. firms and find significant differences in accessibility, length, readability and quality between and within the same industries. Large companies with a medium level of technical sophistication appear to use more legally secure privacy statements and are more likely to share user data with third parties. Other studies analyze the effect of privacy regulation by comparing privacy-statement versions before and after the GDPR became binding. Becher & Benoliel (2021), for instance, focus on the "clear and plain language" requirement in the GDPR (art. 12 GDPR). By analyzing the readability of 216 privacy statements of the most popular websites in the United Kingdom and Ireland after the GDPR became binding, they conclude that privacy statements are hardly readable. For a small sub-sample of 24 privacy statements before and after the GDPR became binding, they document a small improvement in readability. In another study, Degeling et al. (2019) periodically examine, from December 2017 to October 2018, the 500 most popular websites of all EU member states, gathering a final sample of 6,579 privacy statements, and find that the number of sites with privacy statements increased after the GDPR became binding. When focusing on cookie consent libraries, they conclude that most cookies do not fulfill the legal requirements. Linden et al. (2020) study 6,278 privacy statements inside and outside the EU. They underline that the GDPR was a main driver of textual adjustments and that many privacy statements are not yet fully compliant regarding disclosure and transparency. This article extends the previous research by focusing on the FinTech industry in its entirety, which is characterized by the presence of companies in different growth stages ranging from startup companies to established global corporations. Data privacy is particularly important for FinTechs who find themselves caught between the pressure to innovate for future business success and the privacy aspects that result from the highly sensitive data processed in financial services. To address the peculiarities of the companies within the FinTech industry and the data they process, we link the analysis of privacy statements to company- and industry-specific factors.

The guiding principle of the processing of personal data according to the GDPR is transparency (art. 5(1)a GDPR). In this paper, we analyze 276 privacy statements published by German FinTech firms before and after the GDPR became binding. We analyze the readability of the privacy statements, their standardization as a basic requirement for transparency, the amount of data processed, and transparency of data processing in the true sense. We then examine how FinTech company and industry specific factors influence these metrics. We perform textual analysis on the privacy statements and provide evidence that their readability has worsened since the GDPR became binding. Specifically, the texts have become longer and more time-consuming to read. In a next step, we find an increase in the use of standardized text. Further, we study the quantity of data processed as stated in the privacy statements and the related level of transparency. We study whether FinTech-specific factors such as the number of external investors and the existence of bank cooperation predict privacy practices respectively before and after the GDPR became binding. Finally, peer pressure among FinTechs and industry standards might induce mimicking behaviour. We find that *ex-ante* industry-wide privacy practices influence FinTechs' privacy practices after the GDPR became binding. Our results remain robust when excluding more mature FinTechs and when using alternative model specifications.

The rest of this article is organized as follows. The "Institutional Background: The GDPR" section describes the institutional background of the GDPR and the theoretical framework of this study. The "Literature and Hypotheses" section examines the related literature and develops the hypotheses that will be tested. The "Data and Method" section outlines the data and method. The "Results" section presents our results. The "Robustness" section provides robustness checks, and the "Conclusion" section concludes.

Institutional background: The GDPR

The European Parliament passed the GDPR on April 14, 2016. After a transition period, the regulation became binding on May 25, 2018. The regulation is intended to harmonise data protection legislation in the EU. According to its territorial scope (art. 3 GDPR), data of EU citizens are subject to the regulation, independent of whether the data are processed inside or outside the EU. After the GDPR became binding, many jurisdictions outside the EU adopted data protection regulations with a scope and provisions similar to those in the GDPR.[2] In addition to questions of data security, the GDPR distinguishes between four main actors in the field of privacy: the data subject, who is a natural person and whose personal data are processed; the data controller, as the entity offering products or services for which the data are needed; the data processor, supporting the data controller to process the data; and third parties that might process data not directly related to the product or service provision (e.g., companies evaluating a user's credit-worthiness) (Linden et al., 2020). To give the GDPR bite, fines of up to 4% of a company's yearly global revenue or 20 million euros can be imposed in cases of non-compliance (art. 83 GDPR).

This article builds on art. 5 GDPR, which describes the key principles of the processing of personal data, in particular the overarching principle of transparency.[3] Art. 5 GDPR is

[2] Specific examples of privacy regulations similar to the GDPR are the California Consumer Privacy Act of 2018, the Personal Data Protection Act 2019 in Thailand, the Brazilian General Data Protection Law of 2020, the Swiss Federal Act on Data Protection of 2020, and the Chinese Personal Information Protection Law of 2021.

[3] "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject" (art. 5(1)a GDPR).

further specified in the rec. 39 GDPR which demands *inter alia* that natural persons should transparently know about the form of processing of their personal data and the extent of data processing. The basic requirement for transparency is that the information is communicated in understandable language.[4] In addition, our analysis is based on the more concertising statements by the Article 29 Working Party on the Protection of Individuals with Regard to the Processing of Personal Data (2018). Based on the aforementioned legislation regarding transparency, we further investigate in this study the theoretical concepts of readability, standardization, quantity of data processed and finally transparency which we subsume under the term privacy practices.

An important EU directive that pertains directly to the GDPR and which deals with data protection in the FinTech sector—especially in payment services—is the Payment Services Directive 2 (PSD2). Focusing on payment services, the PSD2 regulates practices related to the processing of payment data and lawful grounds for granting access to bank accounts. The PSD2 also deals with the processing of silent party data. Silent party data is personal data of a data subject who is not a user of a specific payment service provider, but whose personal data is processed by that payment service provider for the performance of a contract between the provider and a payment service user. Similar to the GDPR, the PSD2 also addresses issues of user consent, data minimization, data security, data transparency, data processor accountability, and user profiling. Although the PSD2 affects some of the FinTechs studied in this article, we focus below on the more general GDPR, which is equally applicable to all FinTechs.

Literature and hypotheses

Related literature

The theoretical foundation of this study is embedded in the economics of privacy literature investigating economic trade-offs that reveal people's considerations in terms of privacy.[5] The economics of privacy literature is embedded in the broader context of information economics (Posner, 1981) and is substantially affected by the advances in digital information technology.

The GDPR as a new data protection regulation affects nearly every area of life where natural persons claim a service or product with or in exchange for personal data. Therefore, the encompassing consequences and the economic impact of the GDPR are quantified in several studies and highlight a decrease in web traffic, page views and revenue generated as a result of the consent requirement on the part of the data subject (art. 7 GDPR) or limitations in marketing channels (Aridor et al., 2020; Goldberg et al., 2021).

Privacy statements are the essential source of information about how companies put privacy into practice and process personal data. These statements are the standard way to promote transparency to users (Martin et al., 2017) and to balance the equity of power between data subjects and data processors (Acquisti et al., 2015). Therefore, privacy statements are often used in the literature to analyze privacy-related aspects of companies as outlined in the "Introduction." Computer and information science scholars have developed tools that help researchers analyze privacy statements on a large scale (Contissa et al., 2018; Harkous et al., 2018; Tesfay et al., 2018). Contissa et al. (2018), for example, apply their tool to the privacy statements of large-platform and BigTech companies as an exploratory inquiry and conclude that none fully comply with the GDPR, as the formulations are partially unclear, potentially illegal or insufficiently informative.

Privacy and security aspects of FinTech companies have been studied in a variety of contexts. Stewart & Jürjens (2018) survey the German population regarding FinTech adoption and identify data security, consumer trust and user-design interface as the most important determinants. Gai et al. (2017) provide a theoretical construct for future FinTech industry development to ensure sound security mechanisms based on observed security and privacy concerns and their solutions. Other studies emphasize the specificity and importance of the data processed by FinTechs. Ingram Bogusz (2018) describes and distinguishes the data that FinTechs process between *content data*, directly related to the identification of a person, and *metadata*, usually left unintentionally by users but useful for the data processor. Berg et al. (2020) demonstrate the large opportunities to use data collected during 250,000 purchases on a German e-commerce website. Among other things, such data has significant explanatory power to determine creditworthiness. Dorfleitner & Hornuf (2019) provide a descriptive analysis of privacy statements of German FinTechs before and after the GDPR became binding to derive policy recommendations. However, apart from Dorfleitner & Hornuf (2019), the preliminary research does not analyze the privacy statements of FinTech companies specifically regarding privacy regulation and the GDPR. In this study, we go well beyond the simple descriptive statistics of Dorfleitner & Hornuf (2019) and examine the readability and standardization of privacy statements using text analysis. Furthermore, we link the content of the FinTechs' privacy statements to company- and industry-specific factors in a multivariate con-

[4] "The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used." (rec. 39 GDPR).

[5] For a literature review on the economics of privacy, see Acquisti et al. (2016).

text in order to account for the diversity and specificity of business models within the FinTech industry.

Derivation of hypotheses

Readability The GDPR requires that information and communication be transmitted to users in clear and plain language (art. 5, 7, 12 GDPR, rec. 39, 42, 58 GDPR) in order to achieve transparency. This objective corresponds to the linguistic concept of readability, i.e. the reader's ease with and ability to understand a text. Apart from the legislative requirements of the GDPR, companies also have an economic incentive to provide readable privacy statements, which in turn can increase user trust in their business conduct (Ermakova et al., 2014) and thereby create a competitive advantage (Zhang et al., 2020). While these arguments seem to suggest that companies should have increased the readability of their privacy statements after the GDPR became binding, there are also severe counterarguments. Many users do not read disclosures such as privacy statements (Omri & Schneider, 2014), even for products and services they use daily (Strahilevitz & Kugler, 2016). Firms provided their users, often within a very short time frame, updated privacy statements after the GDPR became binding (Becher & Benoliel, 2021). It appears unlikely that such a large number of new privacy statements has triggered additional engagement with these texts by data subjects. Indeed, several studies state that privacy statements are difficult and time-consuming to read and often require an understanding of complex legal or technical vocabulary (Fabian et al., 2017; Lewis et al., 2008; Sunyaev et al., 2015). Second, and in line with this observation, Earp et al. (2005) and Fernback & Papacharissi (2007) find that privacy statements often aim to protect companies from contingent lawsuits rather than address the privacy needs of data subjects. Thus, while firms know that their customers tend to ignore privacy statements, especially if they are technical to read, they may have emphasized their own interests with respect to avoiding lawsuits when updating these statements with respect to the GDPR. Indeed, as long as there is no need for companies to fear that the requirement of clear and plain language will become the subject of legal proceedings, they have few incentives to improve the readability of their privacy statements.

This theoretical argumentation is supported by empirical evidence. Two years after the GDPR became binding, the penalties imposed on companies remain relatively low, and none traces back to the clear and plain language requirement (Wolff & Atallah, 2021). For a sample of 24 privacy statements from the most popular websites in the United Kingdom and Ireland, Becher & Benoliel (2021) finds that many of the privacy statements before the GDPR were barely readable and have improved only slightly since the GDPR became binding. Linden et al. (2020) study 6,278 privacy statements before and after the GDPR became binding using different text metrics like syllables, word count or passive voice and state that the policies became significantly longer but that there was no change in sentence structure.

Summarizing this reasoning, we expect that companies may not have significantly improved the readability of their privacy statements after the GDPR became binding in May 2018.

> **Hypothesis 1:** *The readability of FinTech privacy statements has not improved since the GDPR became binding.*

Standardization The standardization of legal text is often deemed uninformative for the reader and is therefore referred to as *boilerplate* in academic literature. Boilerplate language is characterized by very similar uses of language and wording across legal documents from different issuers (Peacock et al., 2019) and little company-specific information (Brown & Tucker, 2011). For a user, boilerplate text requires much effort to read, and details might appear to be irrelevant (Bakos et al., 2014).

Boilerplate language in legal text brings cost advantages for companies. First, the costs of adopting the specific legal requirements such as the GDPR are lower for all market participants. Second, reduced legal uncertainty due to the use of established and proven text passages, which have yet to cause legal violations, promises fewer future penalties (Kahan & Klausner, 1997). For many companies, the GDPR provided an incentive to intensively address and spend resources on data privacy compliance (Martin et al., 2019). During the period of transition to the GDPR, organizations looked for external information and support regarding the implementation of its legal requirements. Companies often rely on compliance assessment tools to audit their business processes for legal compliance (Agarwal et al., 2018; Biasiotti et al., 2008). In the related literature of requirements engineering, boilerplate language is often proposed to reduce text ambiguities (Arora et al., 2014). For example, Agarwal et al. (2018) provide a tool specifically designed for assessing GDPR compliance, including one process step that allows the user to incorporate boilerplate language. Other sources of information are websites or online policy generators, which deliver guidance on implementing and interpreting the GDPR or even templates for generating privacy statements.[6] The mentioned advantages of applying boilerplate language as well as the examples of assistance to GDPR compliance underpin that we can expect an increase in boilerplate language in the privacy statements since the GDPR became binding.

[6] A template for privacy statements funded within the Horizon 2020 Framework Program of the European Union is provided at https://gdpr.eu/privacy-notice/, last access: 31 August 2021.

Hypothesis 2: *The standardization of FinTech privacy statements has increased since the GDPR became binding.*

Quantity of data processed and transparency For a comprehensive analysis of the FinTechs' transparency beyond readability and standardization, we investigate the content of the privacy statements. While the mere quantity of data processed is important in a first step, we also consider the actual level of transparency.

At the core of the GDPR are principles related to the processing of personal data (art. 5 GDPR), in particular the articles related to lawful, fair and transparent data processing as well as data minimization (art. 5 (1a, c), rec. 39 GDPR). An increase in transparency ensures that consumers provide better-informed consent with respect to the data processed (art. 4, 11 GDPR) (Betzing et al., 2020). An imprecise statement about which and how much personal data are processed violates the provisions of the GDPR, which in turn can result in high penalties. Thus, with regard to the expected costs, an accurate disclosure about which data are processed outweighs the general principle of data minimization. However, the major change of the GDPR introduced compared with the previous privacy legislation in Germany is the potential for high penalties (Martin et al., 2019). This fact represents an incentive for companies to rework their privacy statements, to be precise about the quantity of data processed and to enhance transparency after the GDPR became binding.

Regarding the behavior of data subjects, we apply the theoretical considerations of the privacy calculus model. Data disclosure is the result of a consumer's individual cost-benefit analysis, referred to as a privacy calculus, according to which costs and benefits of disclosing personal data are weighed against each other (Dinev & Hart, 2006). The potential risks of data disclosure are difficult to assess and will only appear in the future, which is why benefits often outweigh costs in the short run (Acquisti, 2004). Data subjects must consent to the privacy statements that are written by companies if they are to receive immediate gratification (O'Donoghue & Rabin, 2000) or, more concretely, to obtain a desired service or product (Aridor et al., 2020). The notion behind many business models is that customers actively forsake parts of their data privacy in exchange for goods and services (Mulder & Tudorica, 2019). Therefore, the data subject's control over the data processed and transparency is limited, and companies have the upper hand.

Empirical studies evidence that it is beneficial and important for companies to ensure and enhance transparency. Li et al. (2019) show that transparency may enhance trust and reputation in a business's activities. Martin et al. (2017) find that a higher level of transparency in the case of a data breach results in a lower negative stock-price reaction.

To summarize the argumentation, we expect an increase not only in the quantity of data processed but also in transparency as companies fulfill the legal requirements of the GDPR and avoid potentially high penalties while benefiting economically.

Hypothesis 3a: *The quantity of data processed by FinTechs has increased since the GDPR became binding.*

Hypothesis 3b: *The transparency of FinTechs has increased since the GDPR became binding.*

Determinants of both the quantity of data processed and transparency In order to account for the peculiarities and diversity of the FinTech industry with regard to data privacy practices, we pay particular attention to the finance literature in developing the following hypotheses. Young companies, such as most FinTechs, prioritize the core business instead of privacy compliance when launching a seminal business. Moreover, founders are rarely experts in privacy or law. Nevertheless, when starting business operations, FinTechs inevitably process personal data and need to act in order to protect privacy sufficiently (Miller & Tucker, 2009) and to comply with current privacy regulation. Therefore, the question arises whether some FinTechs meet the legal requirements better than others. External investors contribute knowledge and experience to build a proper and future-oriented company. The advanced knowledge of external investors is based on experience in legal compliance and privacy with corresponding business contacts and cooperations (Hsu, 2006). The more external investors are involved in an investment, the more likely it is to succeed as a business because of the access to external knowledge (De Clercq & Dimov, 2008). We hypothesize that having a greater number of investors with different education, experience and background knowledge help achieve privacy compliance.

Hypothesis 4a: *External investors increase both the quantity of data processed and transparency of FinTechs.*

Another important group of stakeholders for FinTechs are the banks they may collaborate with. Within such cooperation, FinTechs receive access to financial resources, infrastructure, customers, security reputation (Drasch et al., 2018), a banking license and legal support to comply with regulation (Hornuf et al., 2021a). Moreover, banks have a strong incentive to collaborate with FinTechs in order to boost their digital transformation, which might result in more data being shared. Banks also have long-term experience managing personal data and handling data in compliant way. Banks can transfer this knowledge to FinTechs, especially if they cooperate. We therefore expect that cooperation with a bank has a positive effect on compliance with privacy regulation.

Hypothesis 4b: *Cooperations with banks increase both the quantity of data processed and transparency of FinTechs.*

Mimicking behavior Mimicking behavior often leads to standardization (Kondra & Hinings, 1998) as described in Hypothesis 2, which is particularly likely to be at work after the GDPR became binding. Prior studies evidence that companies tend to mimic the behavior of other companies in the same industry, including for stock repurchase decisions (Cudd et al., 2006), target amounts in crowdfunding (Cumming et al., 2020) or tax avoidance (Kubick et al., 2015). An industry-centric perspective with regard to privacy appears to be reasonable; as Martin et al. (2017) show, when a specific entity experiences a privacy breach, the firm performance of companies in the same segment is also affected. In our study, FinTechs operating in the same sub-segment and thus having corresponding business models should also have similar data processing practices (Hartmann et al., 2016). Consequently, there is an incentive to adopt an immediate peer's privacy statement. Mimicking an industry peer's behavior in the field of privacy is fairly easy, as the privacy statements can be accessed on the corresponding website with just a few clicks. Firms in the same segment can expect to incur similar fines and penalties in cases of non-compliance (Hajduk, 2021). Expert interviews in the context of the GDPR reveal that start-up executives have concerns that their industry peers could report their possible violations to the data protection authorities (Martin et al., 2019). Mimicking industry peers and adopting similar privacy practices prevents companies from experiencing such adversity.

We therefore expect that the industry-specific design of privacy practices stated in privacy statements has a positive influence on a single company's quantity of processed data and transparency.

Hypothesis 5: *Mimicking behavior has a positive influence on the company-specific quantity of data processed and transparency.*

Data and method

Data

Our sample consists of companies operating in financial technology in Germany.[7] Data collection before the GDPR became binding, on 25 May 2018, took place between 15 October 2017 and 20 December 2017. Data collection after the GDPR became binding occurred between 15 August 2018 and 31 October 2018. We comprehensively map the FinTech industry operating in Germany and include both FinTech start-ups and established FinTech companies in our sample. The sample consists of 276 companies with German privacy statements.

Variables

To test Hypothesis 1, we use the readability measures *SMOG German, Wiener Sachtext* and, alternatively, *No. words*. For a test of Hypothesis 2 to examine standardization, we calculate the similarity and distance metrics *Cosine similarity, Jaccard similarity, Euclidean distance* and *Manhattan distance*. We describe these text-based measures and their respective calculations in more detail in the "Methods" section.

Variables of interest

To test Hypotheses 3a and 3b, 4a and 4b and 5, we construct a *data index* to account for the quantity of data processed and the *transparency index* for actions undertaken to ensure transparency. The underlying assumption of the index construction is that we assume that when a company does not concretely state the processing of specific data or certain data-processing practices, such processing does not occur. After the GDPR became binding, this assumption seems justified given the high potential penalties for misrepresentation.

The *data index* is a measure of the quantity of data processed by a company. The data processed ranges from general personal data (e.g. name, address) to metadata (e.g. IP address, social plugins) to special categories of personal data (e.g. health, religion). Table 1 provides the full list of data categories from which the *data index* is composed. For the variable *transparency index*, we aggregate variables representing different dimensions of transparent data-protection actions undertaken by the companies. Apart from vague formulations in art. 12 and rec. 58, 60, the GDPR does not explicitly define and specify transparency or how to ensure transparency. Therefore, we combine the potential transparency vulnerabilities of Mohan et al. (2019) and Müller et al. (2019) to define our considered dimensions of transparency. The *transparency index* represents the normalized sum over eight dummy variables such as *data* (whether a company states in detail which personal data they process), *purpose* (1 if a company states for what reason or purpose personal data are processed) and *storage* (1 if it states how long data are stored or when they are deleted). Table 1 lists in detail the composition of the *transparency index*. As proposed by Wooldridge (2002, p. 661), we divide the indices data index and *transparency index* by the maximum achievable number of variables of which the respective index is

[7] Study data are kindly provided by Dorfleitner & Hornuf (2019). We reduced the original data set to 276 companies because of the non-availability of privacy statements, non-availability of privacy statements in German language, inconsistencies in company data and inactivity or insolvency during both data collection periods.

Table 1 Definition of variables

Variable	Description	Source
Bankcooperation	D: 1 if the Fintech cooperates with a bank, 0 otherwise.	Bank, FinTech websites
No. investors	Logarithm plus 1 of the number of external investment firms and individual investors	BvD Dafne, Crunchbase
Mimic Data Index	Mimicking variable for Data Index	Own calculations
Mimic Transparency Index	Mimicking variable for Transparency Index	Own calculations
Wiener Sachtext	Neuer Wiener Sachtext readability metric	Own calculations
SMOG German	SMOG readability metric (adopted to German language)	Own calculations
No. words	Logarithm of the total number of words	Own calculations
Cosine similarity	Cosine similarity	Own calculations
Jaccard similarity	Jaccard similarity	Own calculations
Euclidean distance	Euclidean distance	Own calculations
Manhattan distance	Manhattan distance	Own calculations
Controls		
Firm age	Logarithm of the age of the FinTech company.	German company register, LinkedIn
Employees	Number of employees (rank variable between 1 and 5)	BvD Dafne, Crunchbase, LinkedIn
City	D: 1 located in a city with more than one million inhabitants, 0 otherwise.	German company register, Websites
Legal capital	D: 1 if a company has a legal form that requires a legal capital of more than 1 EUR, 0 otherwise.	German company register, Websites
GDPR	D: 1 if observations are after the introduction of the GDPR on May 25th 2018, representing the post-GDPR period, 0 otherwise.	
Data index	An index aggregating the quantity of data processed. The index adds the hereafter following variables and is divided by 38.	Own calculations
Name	D: 1 if the first and last name are processed, 0 otherwise.	D and H (2019)
Gender	D: 1 if the gender or form of address are processed, 0 otherwise.	D and H (2019)
Title	D: 1 if the title is processed and 0 otherwise.	D and H (2019)
Language	D: 1 if the company processes the language, 0 otherwise.	D and H (2019)
Identifier	D: 1 if the identifier (e.g. user name or ID) is processed, 0 otherwise.	D and H (2019)
Password	D: 1 if the the password is processed, 0 otherwise.	D and H (2019)
Age	D: 1 if the age or date of birth are processed, 0 otherwise.	D and H (2019)
Place of birth	D: 1 if the place or country of birth are processed, 0 otherwise.	D and H (2019)
Address	D: 1 if the address or delivery address or billing address are, processed, 0 otherwise.	D and H (2019)
E-mail address	D: 1 if the e-mail address is processed, 0 otherwise.	D and H (2019)
Phone number	D: 1 if the phone number or mobile number are processed, 0 otherwise.	D and H (2019)
Residence city	D: 1 if the city of residence is processed, 0 otherwise.	D and H (2019)
Residence country	D: 1 if the company processes the country of residence, 0 otherwise.	D and H (2019)
Marital status	D: 1 if the company processes the marital status, 0 otherwise.	D and H (2019)
Occupation	D: 1 if the occupation or employee status are processed, 0 otherwise.	D and H (2019)
Bank	D: 1 if the bank data or account data or payment data are processed, 0 otherwise.	D and H (2019)

Table 1 continued

Variable	Description	Source
PIN	D: 1 if the PIN or TAN are processed, 0 otherwise.	D and H (2019)
Income	D: 1 if the monthly revenues or expenses are processed, 0 otherwise.	D and H (2019)
Tax residency	D: 1 if the tax residency or status are processed, 0 otherwise.	D and H (2019)
Social security number	D: 1 if the social security number is processed, 0 otherwise.	D and H (2019)
Tax ident number	D: 1 if the tax identification number is processed, 0 otherwise.	D and H (2019)
Driving license	D: 1 if driving license data is processed, 0 otherwise.	D and H (2019)
Passport, registration	D: 1 if passport and identity card data or the registration number are processed, 0 otherwise.	D and H (2019)
Graduation, qualification	D: 1 if information on graduation or qualifications are processed, 0 otherwise.	D and H (2019)
Insurance	D: 1 if information on insurance is processed, 0 otherwise.	D and H (2019)
IP-address	D: 1 if the IP-address is processed, 0 otherwise.	D and H (2019)
GPS, location	D: 1 if GPS or location data are processed, 0 otherwise.	D and H (2019)
Personal data published	D: 1 if personal data are published, 0 otherwise.	D and H (2019)
Personal data transfer	D: 1 if personal data are collected from, transferred to or disclosed with third parties, 0 otherwise.	D and H (2019)
Social Plugins, Third party	D: 1 if social plugins are used or third party services are integrated, 0 otherwise.	D and H (2019)
Behavior, usage, movement	D: 1 if behavioral, usage or movement data are processed or tracking services are used, 0 otherwise.	D and H (2019)
Google Analytics	D: 1 if Google Analytics is used, 0 otherwise.	D and H (2019)
Health	D: 1 if health-related data is processed, 0 otherwise.	D and H (2019)
Religion	D: 1 if the religious confession is processed, 0 otherwise.	D and H (2019)
Nationality	D: 1 if the nationality or citizenship is processed, 0 otherwise.	D and H (2019)
Picture	D: 1 if user or title pictures are processed, 0 otherwise.	D and H (2019)
Conversation record	D: 1 if a conversation recording is processed, 0 otherwise.	D and H (2019)
Signature	D: 1 if the signature or sample of writing is processed, 0 otherwise.	D and H (2019)
Transparency index	An index aggregating dimensions of transparency we define hereafter. The index adds the hereafter following variables and is divided by 8.	Own calculations
Data	D: 1 if the company states which personal data are processed, 0 otherwise.	D and H (2019)
Purpose	D: 1 if the company states for what reason or purpose personal data are processed, 0 otherwise.	D and H (2019)
Storage	D: 1 if the company states for how long data are stored or when they are deleted, 0 otherwise.	D and H (2019)
Avoid	D: 1 if the company states if there exists a possibility to avoid data processing, 0 otherwise.	D and H (2019)
Opt-In	D: 1 if the company states whether they have an Opt-In procedure, 0 otherwise.	D and H (2019)
Pseudo	D: 1 if the company states that data are processed pseudonymously, 0 otherwise.	D and H (2019)
Third	D: 1 if the company states which personal data are shared with third parties, 0 otherwise.	D and H (2019)
Third data	D: 1 if the company states with which third parties data are shared, 0 otherwise.	D and H (2019)

Note: List and definitions of all variables with the corresponding source. In the following table the abbreviation “D" stands for dummy variable and “D and H (2019)" for Dorfleitner & Hornuf (2019). All variables that are directly included in the following analyzes are marked in *italics*

composed to scale them between 0 and 1. We interpret a higher index value to mean respectively a higher quantity of data processed and more transparency.

Explanatory variables

To construct our explanatory and control variables, we collect detailed firm-specific variables, which we describe below with the data sources used. Accuracy of the data was validated using cross-checks with press releases, FinTech websites and other news and information online.

To test Hypothesis 4a, that a higher number of external investors positively influences the quantity of data processed and transparency, we include the variable *No. investors*, measured as the absolute number of external investment firms and individual investors who funded the company. This variable is already considered in other FinTech-related studies such as Cumming & Schwienbacher (2018) and Hornuf et al. (2018b). We derive the variable from the BvD Dafne and Crunchbase database, which was also used in other academic papers, such as Bernstein et al. (2017) and Cumming et al. (2019).

To test Hypothesis 4b, we include the dummy variable *bankcooperation*, which equals 1 if the respective company has a cooperation with a bank and 0 otherwise. For data collection, we first searched all bank websites to find indications of bank-FinTech cooperation. In a second step, we checked for cooperation from the FinTech side.

To analyze mimicking behavior as outlined in Hypothesis 5, we follow the approach of Cudd et al. (2006), who use the industry average of a measure in the year preceding the focal period for mimicking behavior. We obtain the variables *mimic data index* and *mimic transparency index* by calculating the average of the indices *data index* and *transparency index* within the same FinTech sub-segment before the GDPR became binding according to the taxonomy of Dorfleitner et al. (2017).

Control variables

To consider unobserved heterogeneity, we use the following control variables. First, we control for firm location with the variable *city*, which can be a relevant geographic determinant. This variable indicates whether a company is located in a city with more than one million inhabitants. In metropolitan areas, more customers and sources for funding (Hornuf et al., 2021a) as well as start-up incubators are within geographical reach and thus available to support a company's development. Besides, more FinTechs are located in one place in metropolitan areas, which often leads to the establishment of entrepreneurial clusters (Porter, 1998). Competition within a cluster necessitates the creation of a competitive advantage (Tsai et al., 2011), which is a quality signal of compliance with applicable privacy regulation. Gazel & Schwienbacher (2021) provide empirical evidence that location in a cluster reduces the risk of firm failure for FinTechs. We collected the data from the German company register.

Second, we consider the variable *legal capital*. This variable reflects the founder's dedication and readiness to make a notable investment in the own venture at an early stage of development (Hornuf et al., 2021b) and which can be interpreted as a quality signal of motivation and future success of business operations. In Germany, for the most common legal form of a limited liability company (the so-called *GmbH*), one needs to raise legal capital of at least 12,500 EUR at the time of incorporation. The dummy variable equals 1 if the minimum capital requirement of the underlying legal form amounts to more than 1 EUR and 0 otherwise. We derived this information from the German company register and imprints of the FinTech websites.

Third, we include number of *employees* as a proxy for FinTech companies' human capital and size (Hornuf et al., 2018a). *Employees* is a rank variable ranging from 1 to 5 and representing number of employees: 1–10, 11–50, 51–100, 101–1000 and above 1000. A larger number of employees usually means a more diversified team in terms of members' abilities and skills, resulting in venture success (Duchesneau & Gartner, 1990), which might also translate to compliance and legal aspects. For privacy-related aspects, Ramadorai et al. (2021) outline that larger firms tend to extract more data. Therefore, we proxy for firm size and human capital strength using the number of employees. We derived the data from BvD Dafne and complemented them with data from the Crunchbase database as well as LinkedIn entries.

Fourth, we control for the age of the FinTech company during the particular data-collection period since its year of incorporation with the variable *firm age*. This variable serves as a proxy for a FinTech's stage of business (Hornuf et al., 2021b). We assume that established companies pay more attention to privacy aspects because they have more experience and available resources. Bakos et al. (2014) find for contracts in boilerplate language that consumers have more confidence in larger and older companies because they seem more credible and fair. We derive the year of incorporation from the German company register and respectively calculate it as the difference of the data collection period before and after the GDPR became binding.

We further include *industry dummies* to account for the diversity of business models. Our industry classification follows the FinTech taxonomy of Dorfleitner et al. (2017) with the segments and sub-segments (in parentheses): financing (donation-based crowdfunding, reward-based crowd, crowdinvesting, crowdlending, credit and factoring), asset management (social trading, robo-advice, personal

financial management, investment and banking), payments (alternative payment methods, blockchain and cryptocurrencies, other payment FinTechs) and other FinTechs (insurance, search engines and comparison websites, technology IT and infrastructure, other FinTechs). The categorization is based on FinTechs' business models in accordance with the functions and business processes of traditional banks. The business model provides first indications about the data processing of a specific FinTech because in a digitized industry, data are often at the core of the business model.

The variables *employees, legal capital, bankcooperation* and *city* are time-invariant. We collected all variables in this paper respectively before and after the GDPR became binding.

Methods

Textual analysis: Preprocessing

We prepare the texts of the privacy statements using standard methods of text mining, including cleaning to remove white spaces, numbers, punctuation and other symbols. For the standardization analysis, we also need to consider that the language of the privacy statements is German. We therefore remove capitalization and apply stemming to the German language to reduce words to their root in order to consider different grammatical forms of the same word family. We delete stop words with the help of the German stop word list in the R package "lsa" (Wild, 2022) because stop words such as articles, conjunctions and frequently used prepositions do not convey additional meaning. Subsequently, we break the texts down into tokens that represent individual words and count their frequency within each text separately for both data-collection periods.

Readability

The GDPR refers to the comprehensibility of privacy statements in order to achieve transparency with "easy to understand, and [...] clear and plain language" (rec. 39 GDPR) and mentions "that it should be understood by an average member of the intended audience" (Working Party on the Protection of Individuals with Regard to the Processing of Personal Data, 2018). Readability is defined as the ease of understanding a text and is usually measured using formulas based on sentence length, syllables and word complexity. The most commonly used readability measures in academic literature are the Flesch reading ease score (Flesch, 1948) and the Gunning Fog Index (Gunning, 1952), both corresponding to the number of formal years of education required to comprehend a text. We investigate companies operating in Germany and because the privacy statements are often written in German, we address the variety of morphological and semantic richness by using metrics for or adapted to German.

First, we apply the Neue Wiener Sachtext formula by Bamberger & Vanecek (1984) using the formula

$$nWS = 0.1935 \cdot \frac{n_{wsy\geq 3}}{n_w} + 0.1672 \cdot ASL + 0.1297 \cdot \frac{n_{wchar\geq 6}}{n_w} - 0.0327 \cdot \frac{n_{wsy=1}}{n_w} - 0.875 \quad (1)$$

where $n_{wsy\geq 3}$ is the number of words with three syllables or more, ASL is the average sentence length (number of words / number of sentences), $n_{wchar\geq 6}$ is the number of words with 6 characters or more and $n_{wsy=1}$ is the number of words of one syllable.

Second, we calculate the simplified SMOG metric of McLaughlin (1969) adapted to the peculiarities of the German language as

$$SMOG\ German = \sqrt{Nw_{min3sy} \cdot \frac{30}{n_{st}} - 2} \quad (2)$$

where Nw_{min3sy} is the number of words with a minimum of three syllables and n_{st} is the number of sentences (Bamberger & Vanecek, 1984). While these formulas for determining readability are frequently used in the literature (Loughran & McDonald, 2016; Ramadorai et al., 2021), they are nevertheless often criticized (Loughran & McDonald, 2014). Regarding privacy statements, Singh et al. (2011) state that the measures take into account sentence complexity and word choice but no aspects that determine comprehension. To address these points of criticism, we additionally consider the variable *No. words*, defined as the logarithm of the total number of words in the privacy statements. We consider the variable as an alternative measure of the understandability and complexity of a text reflected in the time required to read the whole text.

Standardization

To test Hypothesis 2, we quantify the extent to which the texts of privacy statements are standardized by calculating common measures of text similarity and distance for dissimilarity. We apply the vector space model (VSM) of Salton et al. (1975) to convert texts into term-frequency vectors, which enables us to perform algebraic calculations. The accounting and finance literature often applies Cosine or Jaccard similarity to account for similarity (Cohen et al., 2020; Peterson et al., 2015).[8]

As a first similarity measure, we calculate the *Cosine similarity*. Because of the vector representation of the texts, we

[8] For illustrative examples of Cosine and Jaccard similarity, see Cohen et al. (2020).

can calculate the cosine of the included angle. The *Cosine similarity* between two documents is defined as the scalar product of the two term-frequency vectors divided by the product of their Euclidean norms. The values range from 0 to 1 because term-frequency vectors of texts cannot be negative. A main property of the *Cosine similarity* is that it does not consider text length. A value close to 1 indicates the presence of pure boilerplate language. The second similarity measure we calculate is *Jaccard similarity*, defined as the quotient of the size of the intersection and the size of the corresponding union of two term-frequency representations. In contrast to *Cosine similarity*, for the *Jaccard similarity* each word occurs only once in the sample, and its frequency is not accounted for. For privacy statements, Ramadorai et al. (2021) use Cosine similarity to analyze industry-specific boilerplate, whereas Kaur et al. (2018) employ Jaccard similarity to measure keyword similarity. Besides the similarity measures, we calculate the two distance metrics *Euclidean distance* and *Manhattan distance*. *Euclidean distance* is the shortest distance between the two document vectors with the corresponding term weights. In contrast to *Euclidean distance*, *Manhattan distance* is the absolute distance between the two vectors. Unlike for the similarity measures, values of distance metrics close to 1 indicate no correspondence between the analyzed texts.

We calculate all the aforementioned similarity and distance measures pairwise for the privacy statement texts D_1 and D_2 of two different companies within one data-collection period. In the next step, to obtain one average similarity or distance-measure value for one company before and after the GDPR became binding, we calculate our similarity and distance measures in relation to the average privacy statement per period, analogous to the "centroid vector [or] the average policy" of Ramadorai et al. (2021), as

$$\overline{D} = \frac{\sum_{n=1}^{N} D_n}{N} \tag{3}$$

where $\overline{D}$ is the average value per year, $\sum D_n$ the sum of the similarity respective distance of one FinTech's document in relation to every other document, and N the number of companies.

Empirical approach

To test our Hypotheses 1, 2, 3a and 3b, we use a two-sided paired t-test to examine whether the mean values of readability, standardization, quantity of data processed, and transparency are significantly different for the periods before and after the GDPR became binding.

We test Hypotheses 4a, 4b and 5 in a multivariate setting. Because our dependent variables are fractional indices in the interval between 0 and 1, we estimate fractional probit regressions using quasi-maximum likelihood (Papke & Wooldridge, 1996).

We further explore in Hypothesis 4a and 4b determinants of the *data index* and *transparency index* in separate models before and after the GDPR became binding. To compare the obtained regression coefficients of non-linear models for the same sample of companies at two different points in time, we further conduct seemingly unrelated estimations (Zellner, 1962). Then, we perform Wald chi-square tests to test whether the coefficients differ across our analyzed periods. The validity of the tests is ensured by the previously performed estimation based on the stacking method with respect to the appropriate co-variance matrix of the estimators for the standard errors (Weesie, 1999) and was formerly successfully applied by Mac an Bhaird & Lucey (2010) and Laursen & Salter (2014).

Results

Sample

Figure 1 shows the graphical distribution of the companies in their sub-segments following the detailed FinTech taxonomy of Dorfleitner et al. (2017). Table 2 provides summary statistics for all our variables. Most of the companies in the sample operate in the crowdinvesting and alternative payments, insurance respective IT, technology and infrastructure sub-segments. Crowdinvesting can be a data-intensive subsegment (Ahlers et al., 2015), whereas payment providers receive manifold payment data that can entail almost all possible information about a person. Moreover, insurance companies typically process health data, which are special categories of personal data (art. 9 GDPR). The descriptive statistics of *bankcooperation* indicate that, on average, 25.4% of FinTechs in our sample maintain a cooperation with a bank. The median of *No. investors* is 0, which indicates that less than half the companies in the sample have received external funds. The mean and median values of *employees*, around 2, indicate that most of our FinTechs are small companies employing 11 to 50 people. The variable *city* indicates that, on average, 48.6% of the analyzed FinTechs are located in a large city.

Readability

The mean and median in combination with the quantiles of the readability metrics *Wiener Sachtext* and *SMOG German* increase slightly, which indicates that the readability

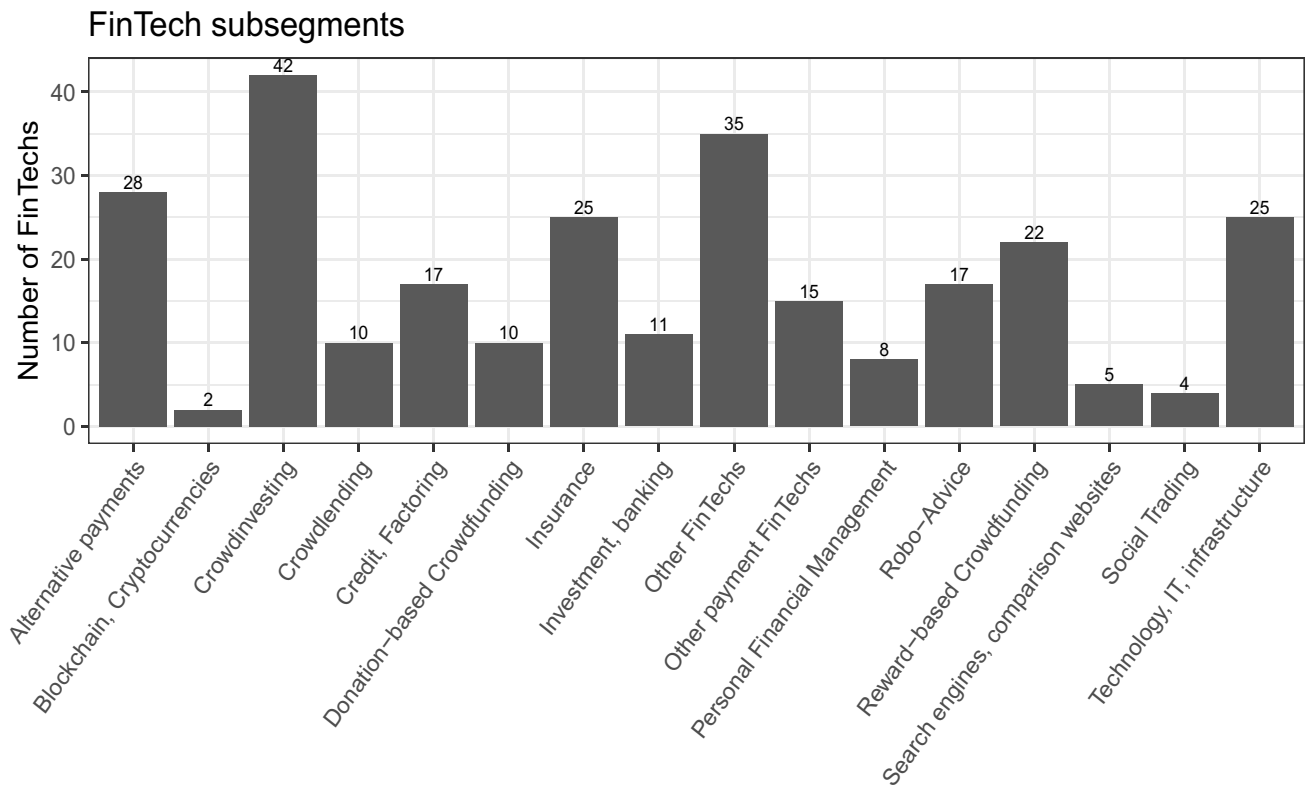

Fig. 1 Frequency of occurrence of the FinTech sub-segments following the taxonomy of Dorfleitner et al. (2017), the bars represent the number of companies in each sub-segment. N=276

of the privacy statements worsened after the GDPR became binding. In Table 3, two t-tests indicate a significant difference in means for both metrics (paired t-tests, $t = 2.569$ and $p < 0.05$, $t = 6.010$ and $p < 0.01$). The alternative readability proxy *No. words* shows a clear increase in any summary statistic, which indicates that the privacy statements contain more words and require more time to read. The increase is confirmed by a t-test on differences in means (paired t-test, $t = 15.017$, $p < 0.01$).

The cumulative distribution functions of all our variables considering readability are illustrated in Fig. 2. A shift of the graph to the right indicates a worsening in readability from before (black) to after (grey) the GDPR became binding, which is evident for all our measures. These results are contrary to the GDPR's objective of clear and plain language. A discussion of the result for *Wiener Sachtext* and *SMOG German* requires a closer look at the method. Both metrics are mainly calculated based on word complexity and sentence length. In particular, word complexity is a critical issue for technical termini, which accompanies privacy-related legalese. Because the information content and quality regarding advanced technological topics can suffer from simpler language (Wachter, 2018). It is not surprising that in the FinTech industry a more complex language has recently been used to describe the data processing of complex business models based on, for example, artificial intelligence or the blockchain technology. Our results for *No. words* are in line with (Linden et al., 2020), who find in their before and after the GDPR comparison an increase in text length but no changes in sentence structure. Thus, our evidence supports Hypothesis 1.

Standardization

In this section, we test Hypothesis 2 on the increase of boilerplate language after the GDPR became binding. The similarity measures *Cosine similarity* and *Jaccard similarity* reveal a clear increase in mean and median. Both measures indicate an increase in boilerplate language, which is confirmed by a t-test for differences in means at conventional levels (paired t-tests, $t = 8.606$ and $p < 0.01$, $t = 6.880$ and $p < 0.01$). Consistent with the similarity metrics, we identify for the distance metrics *Euclidean distance* and *Manhattan distance* a decrease in means and medians, indicating an increase in the use of boilerplate language. The means are statistically significantly different before and after the GDPR

Table 2 Descriptive statistics of all variables

Variable	Mean	S.D.	Min	Q1	Median	Q3	Max
Legal capital	0.888	0.316	0.000	1.000	1.000	1.000	1.000
Bankcooperation	0.254	0.436	0.000	0.000	0.000	1.000	1.000
Employees	2.130	1.050	1.000	1.000	2.000	2.000	5.000
City	0.486	0.501	0.000	0.000	0.000	1.000	1.000
Firm age_pre	1.534	0.529	0.000	1.099	1.498	1.792	3.091
Firm age_post	1.749	0.436	0.693	1.386	1.701	1.946	3.136
No. investors_pre	0.714	1.047	0.000	0.000	0.000	1.400	4.000
No. investors_post	0.754	1.072	0.000	0.000	0.000	1.400	4.000
Wiener Sachtext_pre	13.654	0.915	10.113	12.988	13.739	14.332	17.270
Wiener Sachtext_post	13.860	1.127	9.888	13.149	13.866	14.625	17.225
SMOG German_pre	12.266	1.221	8.955	11.321	12.255	13.222	16.974
SMOG German_post	12.992	1.749	8.191	11.809	13.111	14.070	17.310
No. words_pre	7.102	0.864	2.890	6.796	7.252	7.601	8.970
No. words_post	7.866	0.867	2.944	7.453	7.959	8.443	9.622
Cosine similarity_pre	0.533	0.095	0.132	0.495	0.559	0.601	0.659
Cosine similarity_post	0.583	0.089	0.126	0.537	0.603	0.651	0.706
Jaccard similarity_pre	0.207	0.047	0.023	0.191	0.217	0.238	0.276
Jaccard similarity_post	0.227	0.044	0.014	0.214	0.240	0.255	0.280
Euclidean distance_pre	0.096	0.024	0.074	0.083	0.090	0.101	0.303
Euclidean distance_post	0.081	0.023	0.062	0.070	0.076	0.087	0.318
Manhattan distance_pre	1.312	0.136	1.132	1.226	1.275	1.350	1.901
Manhattan distance_post	1.255	0.125	1.097	1.178	1.229	1.286	1.934
Data Index_pre	0.206	0.103	0.000	0.125	0.200	0.275	0.575
Data Index_post	0.237	0.098	0.000	0.169	0.225	0.300	0.550
Transparency Index_pre	0.303	0.175	0.000	0.125	0.375	0.375	0.875
Transparency Index_post	0.295	0.158	0.000	0.125	0.250	0.375	0.875
Mimic Data Index	0.206	0.037	0.106	0.181	0.198	0.226	0.340
Mimic Transparency Index	0.303	0.065	0.175	0.254	0.290	0.335	0.425

Note: Descriptive statistics for all our variables, the abbrevation "_pre" indicates before and "_post" after the GDPR became binding. N=276. The variables are defined in Table 1

Table 3 Paired two-sided t-test to test Hypotheses 1, 2, 3a, 3b

Variable	*Pre-GDPR*		*Post-GDPR*		Diff.	t-stat	*p*-value
	Mean	S.D.	Mean	S.D.			
Wiener Sachtext	13.654	0.915	13.860	0.206	0.206	2.569	0.011*
SMOG German	12.266	1.221	12.992	1.749	0.726	6.010	0.000***
No. words	7.102	0.864	7.866	0.867	0.764	15.017	0.000***
Cosine similarity	0.533	0.095	0.583	0.090	0.050	8.606	0.000***
Jaccard similarity	0.207	0.047	0.227	0.044	0.020	6.880	0.000***
Euclidean distance	0.096	0.024	0.081	0.023	−0.015	−12.530	0.000***
Manhattan distance	1.312	0.136	1.255	0.125	−0.057	−7.074	0.000***
Data Index	0.206	0.103	0.237	0.098	0.031	5.940	0.000***
Transparency Index	0.303	0.175	0.295	0.158	−0.009	−0.904	0.367

Note: Paired two-sided t-test (significance level of 5%) to test Hypothesis 1 regarding readability, Hypothesis 2 regarding standardization and Hypotheses 3a and 3b regarding quantity of data processed and transparency. N=276. The variables are defined in Table 1

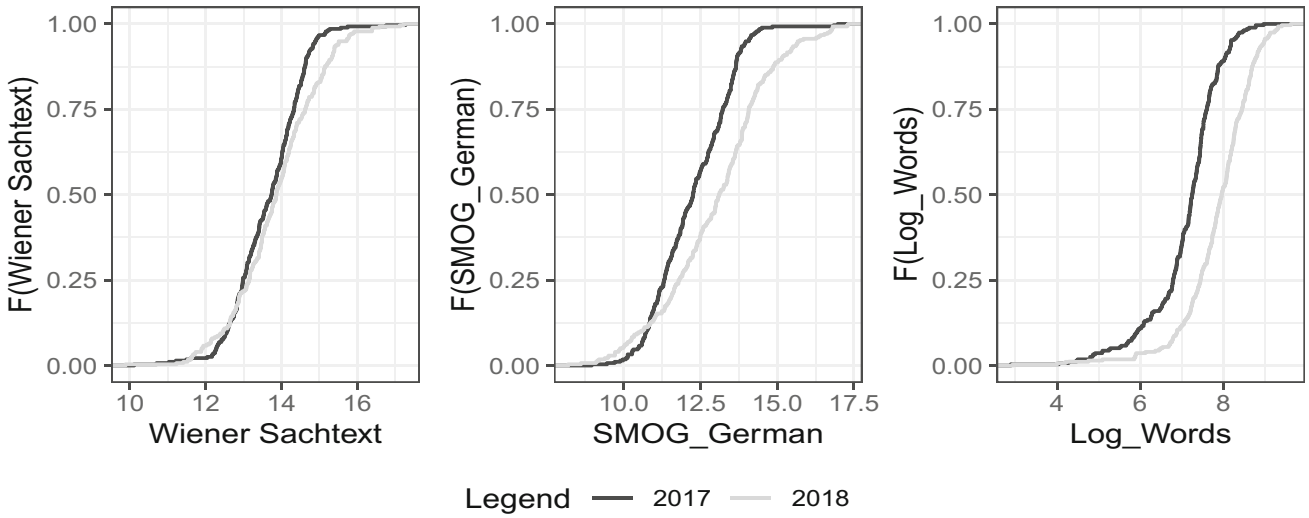

Fig. 2 Cumulative distribution function for the readability measures *Wiener Sachtext, SMOG German* and *No. words* for before (2017, black) and after (2018, grey) the GDPR became binding. N=276. The variables are defined in Table 1

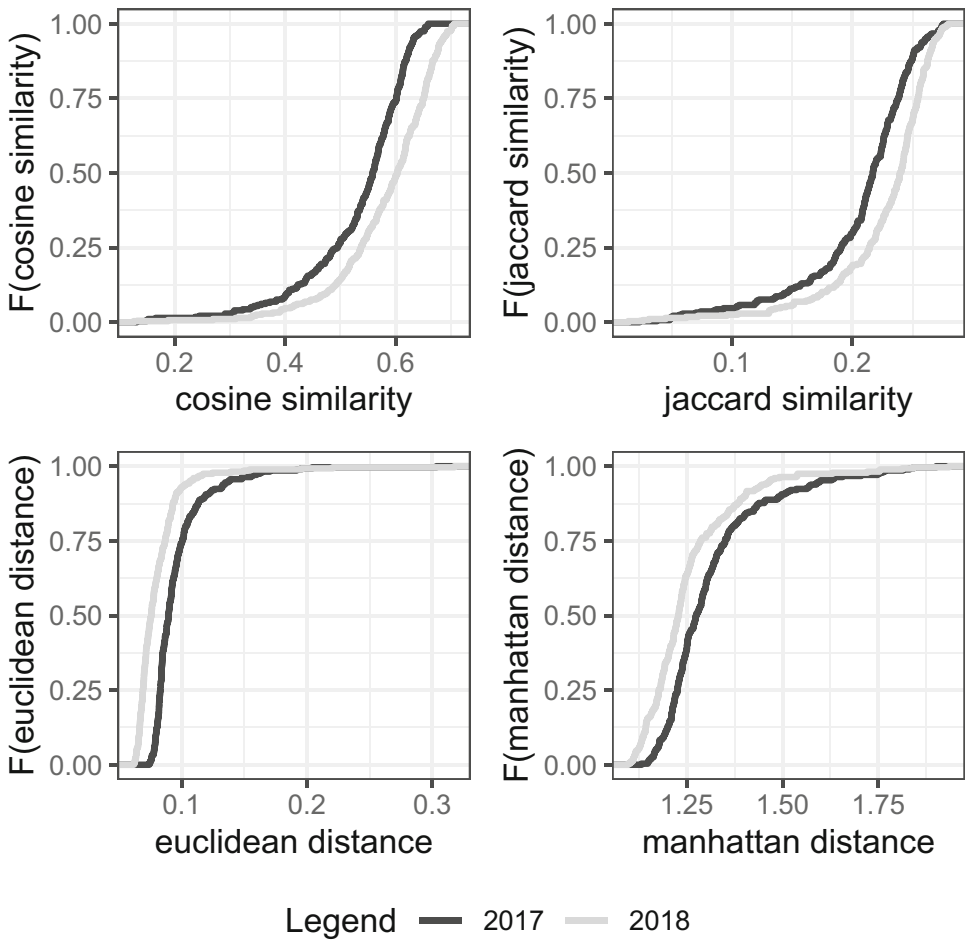

Fig. 3 Cumulative distribution function for the similarity and distance measures *cosine similarity, jaccard similarity, euclidean distance* and *manhattan distance* for before (2017, black) and after (2018, grey) the GDPR became binding. N=276. The variables are defined in Table 1

became binding (paired t-tests, $t = -12.530$ and $p < 0.01$, $t = -7.074$ and $p < 0.01$). The standard deviation for all measures remains almost the same for both sample periods. Regarding all of our similarity and distance metrics, the first and third quantiles are far from the minima or maxima, illustrating that although some outliers exist, there is a tendency towards the mean and the median. In Fig. 3, the cumulative distribution function of the similarity and distance measures illustrates a shift to more similar and therefore standardized language from before (grey) compared with after (black) the GDPR became binding.

In sum, we find an increase in privacy statements' use of standardized language after the GDPR became binding. Companies appear to have chosen the path towards legal-safeguarding boilerplate policies to the detriment of their users. Overall, Hypothesis 2 receives support.

Quantity of data processed and transparency

In this section, we move from the analysis of the readability as the basic requirement for transparency to the actual transparency in terms of content of the privacy statements.[9] For the *data index*, we find an increase in the mean and median from before to after the GDPR became binding, which illustrates that companies state more often in their privacy statements post-GDPR that they process specific data. The difference is statistically significant in a t-test (paired t-test, $t = 5.940$, $p < 0.01$). Thus, we find supportive evidence for Hypothesis 3a. A closer look at all summary statistics emphasizes large divergences in the quantity of data processed between the individual companies. The *data index* minimum of 0 indicates that some firms do not state that they process any data. The range of the actual maximum value before and after the GDPR became binding indicates that even companies that process a lot of data are far from the maximum theoretical index value of 1.

For the *transparency index*, we find a small decrease in the mean and median. This finding suggests that, contrary to our Hypothesis 3b, companies' privacy practices have not improved in terms of transparency since the GDPR became binding. Note that there are companies in both periods reaching a maximum value of 0.875 for the *transparency index*, which indicates a high level of transparency. After performing a t-test on the mean, we find no statistically significant difference (paired t-test, $t = -0.940$, $p > 0.05$). Thus, we find no empirical support for Hypothesis 3b, that transparency has increased since the GDPR became binding. However, one must bear in mind that the FinTech industry operates in a highly competitive environment and is caught between the pressure to innovate and state-of-the-art data privacy. For this reason, it can be difficult for FinTechs to be fully transparent without losing their competitive edge to competitors. In contrast to our results, Linden et al. (2020) use different but closely related transparency measures and conclude that transparency has improved since the GDPR became binding but that privacy statements are far from fully transparent.

When considering the results of both indices, we conclude that since the GDPR became binding, FinTechs state that they process more data although they have not made efforts to enhance the transparency of privacy practices. Further, we identify large differences between individual companies. A possible explanation is that the FinTech industry as a whole is highly diverse and that the different business models require different intensities of data processing. For example, crowdinvesting platforms process a lot of data. The projects and initiator data need to be assessed in detail before the funding. During the funding process, disclosure of more information about the project and the initiators has been identified as a success factor (Ahlers et al., 2015).

Determinants of the quantity of data processed and transparency

Table 4 shows the results for Hypotheses 4a and 4b on the effect of the number of investors and the existence of a bank cooperation on the quantity of data processed and transparency.

We find that before the GDPR became binding, the coefficient of *No. investors* is positive and significant at the 5% level for both indices, where a one-standard-deviation increase in *No. investors* is associated with a 55.9% increase in the *data index* in model (1) and a 41.2% increase in *transparency* in model (3). However, the effect and significance of the variable disappear for the period after the GDPR became binding in models (2) and (4). Before the GDPR became binding, the number of external investors had a positive effect on data-privacy compliance because it was positively related to the quantity of data processed and to transparency. Our results for *No. investors* provide partial support for Hypothesis 4a.

Further, none of our regression models yield a significant effect of *bankcooperation* on quantity of data processed or on transparency. Because of missing significances, we cannot provide further evidence for how external investors or cooperating banks influenced the implementation of the GDPR by FinTechs. Regarding *bankcooperation*, we find no empirical support for Hypothesis 4b.

The control variable *legal capital* has a significant positive influence on both indices for all models, which indicates that founders who invested more legal capital are also more dedicated to their business in terms of data privacy compliance.

[9] For detailed summary statistics of our disaggregated indices, we refer readers to Tables 13 and 14 in the Appendix.

Table 4 Seemingly unrelated fractional probit regressions to test Hypotheses 4a and 4b

	Dependent variable:					
	Data Index			Transparency Index		
	Pre-GDPR (1)	Post-GDPR (2)	Wald-Test *p*-value	Pre-GDPR (3)	Post-GDPR (4)	Wald-Test *p*-value
No. investors	0.055*	0.025	0.358	0.069*	0.042	0.568
	(0.024)	(0.022)		(0.034)	(0.031)	
Bankcooperation	0.031	−0.005	0.574	0.035	0.031	0.970
	(0.048)	(0.044)		(0.070)	(0.065)	
Legal capital	0.247***	0.107+	0.137	0.575***	0.268**	0.037*
	(0.071)	(0.061)		(0.095)	(0.092)	
City	−0.059	0.010	0.223	−0.045	0.075	0.137
	(0.042)	(0.038)		(0.058)	(0.056)	
Firm age	−0.025	−0.019	0.913	−0.061	−0.083	0.796
	(0.042)	(0.043)		(0.055)	(0.067)	
Employees	0.032	0.040+	0.814	0.006	0.033	0.604
	(0.024)	(0.022)		(0.035)	(0.037)	
Constant	−0.752***	−0.601***		−0.711***	−0.624***	
	(0.155)	(0.153)		(0.161)	(0.164)	
Industry Effects	Yes	Yes		Yes	Yes	
Observations	276	276	276	276	276	276
Log Likelihood	−97.239	−103.970		−115.858	−115.384	

Note: Seemingly unrelated fractional probit estimations to test Hypotheses 4a and 4b regarding determinants on the quantity of data processed and transparency, Wald-Test (significance level of 5%) with *p*-values to compare equality of coefficients of models (1)(2) and (3)(4), numbers in parentheses are robust standard errors. The variables are defined in Table 1. $^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Wald tests for differences in coefficients before and after the GDPR became binding only show a significant difference for legal capital as a determinant of the *transparency index*. The coefficients for the *transparency index* are significantly different and lower after the GDPR (Wald chi-square test, $\chi^2 = 4.740$, $p < 0.05$). Thus, the effect of *legal capital* on transparency is stronger before the GDPR. This may be because before the GDPR became binding, only highly dedicated founders invested time in privacy compliance, whereas the GDPR made this issue the focus of every company. We consider variance inflation factors (VIF), reported in Table 8 in the Appendix, and find no indications of multicollinearity for any of our model specifications.

Mimicking behavior

Table 5 reports the results for Hypothesis 5, which considers mimicking behavior regarding data privacy compliance among industry peers.

In Table 5 model (1), we find a positive significant effect of the *mimic data index* on the 1% significance level, in which a one-standard-deviation increase in *mimic data index* is associated with a 58.6% increase in the *data index* relative to the average. In model (2), we find a highly significant impact of the *mimic transparency index* on the *transparency index*, in which a one-standard-deviation increase in the explanatory variables leads to a 75.27% increase in the dependent variable relative to the average. The results indicate a strong mimicking behaviour among industry peers in terms of data privacy compliance, because a higher industry average for both indices before the GDPR became binding accompanies more data processed and greater transparency for a specific company.[10] Thus, the conjecture that FinTechs mimic the privacy statements of their industry peers is supported by our evidence. As for our control variables, we find a weak statistically positive effect for *legal capital* for both indices. In sum, we find supportive evidence for Hypothesis 5 on mimicking behavior.

[10] In unreported analysis, we estimate the same model using a mimicking variable based on segment-level averages of finance, asset management, payments and other FinTechs. Interestingly, we find for that specification no statistically significant coefficients and thus conclude that the less detailed categorization fails to depict commonalities in business models, data processing and consequently mimicking behavior.

Table 5 Fractional probit regression to test Hypothesis 5

	Dependent variable:	
	Data Index Post-GDPR (1)	Transparency Index Post-GDPR (2)
No. investors	0.018	0.024
	(0.023)	(0.033)
Bankcooperation	0.026	0.053
	(0.045)	(0.061)
Legal capital	0.110+	0.174+
	(0.064)	(0.102)
City	0.021	0.079
	(0.039)	(0.054)
Firm age	−0.030	−0.044
	(0.034)	(0.052)
Employees	0.027	0.026
	(0.022)	(0.037)
Mimic Data Index	1.613**	
	(0.554)	
Mimic Transparency Index		2.027***
		(0.427)
Constant	−1.189***	−1.373***
	(0.145)	(0.176)
Industry Effects	No	No
Observations	276	276
Log pseudolikelihood	−150.6126	−165.5007

Note: Fractional probit regression to test Hypothesis 5 regarding mimicking behavior, numbers in parentheses are robust standard errors. The variables are defined in Table 1. $^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Robustness

Finally, we perform robustness checks and estimate alternative specifications to test the validity of our results.

Sub-sample: Exclusion of mature FinTechs

To test for the influence of more mature FinTechs, we exclude companies, like Hornuf et al. (2021a), that employ more than 1000 people or that were founded at least 10 years before our first data-collection period. More experienced and larger companies have more free resources to address legal issues. Especially regarding boilerplate and mimicking behavior, it could be argued that larger or older firms are role models for their immediate industry peers and whose privacy practices are mimicked. When excluding these FinTechs, 249 companies remain in the sample. In Table 9 in the Appendix, we report summary statistics for the text-feature analysis and find patterns remarkably similar to those for the whole sample analyzed in the "Results" section. For the regression estimates in Tables 10 and 11 in the Appendix, we observe no changes in signs and only small changes in significance of the coefficients. Therefore, we note that it is unlikely that more mature FinTechs drive our results.

Pooled OLS with GDPR interaction

To verify our results for the year-wise estimations and post-estimation tests in the seemingly unrelated estimations in the "Results" section, we run an OLS regression with the interaction dummy variable *GDPR*. We estimate the OLS regression to simplify the regression model for the link function in the prior probit specification and pool our observations in a single model with the GDPR interaction to evaluate the effect of the policy intervention simultaneously. The results are reported in Table 12 in the Appendix and mostly show similar patterns in terms of signs and significance of the coefficients compared with the prior model specifications. Additionally, we find that the dummy variable *GDPR* itself has a positive significant influence on level of transparency.

Causality

Endogeneity problems in empirical studies can come in a variety of forms. Reverse causality and simultaneity are

among the most relevant. In this study, the results in Tables 4 and 5 could, for example, be affected by reverse causality. However, when considering the significant variable *legal capital*, it can be argued that the decision for the *legal capital* is made at the moment the company is founded, while the decision for the dependent variable, the *data and transparency index*, is made at a later stage when the company begins operations. As for simultaneity, variables that are potentially missing should, for example, correlate with *firm age*, which is not significant though. This gives us some confidence that endogeneity is not an obvious problem in our analysis.

Conclusion

The theoretical framework of this study is embedded in the general legal principle for data processing, namely transparency (art. 5(1)a GDPR). We empirically study the degree of implementation of the GDPR by FinTech companies operating in Germany. For this purpose, we analyze the privacy statements of 276 FinTechs before and after the GDPR became binding. We use methods from text analysis, extend our findings using a content-based approach, and link this to FinTech company- and industry-specific determinants.

With regard to the text-feature analysis, we document a decrease in readability in conjunction with substantially longer texts and more time required to read the privacy statements. The FinTechs appear to safeguard themselves with exact technical and legal termini and comprehensive statements instead of the user comprehension required by the GDPR. We further find indications of an increase in standardized legal language built on the literature of boilerplate after the GDPR became binding, reducing the informational content that users can draw from the texts. These findings contradict the basic requirements for transparency of the GDPR. Further, we analyze the quantity of data processed, the actual transparency of privacy statements, and their determinants. We document a significant increase in the quantity of data processed but find no significant changes in the level of transparency. The number of external investors positively influences the quantity of data processed and transparency solely before the GDPR became binding. Regarding cooperation with a bank, we find no significant effects in any specification. Legal capital that we interpret as *ex-ante* founder team dedication is positively related to the level of privacy and is particularly relevant for transparency before the GDPR became binding. These results underline that before the GDPR became binding, externally induced pressure of investors and internal engagement of the founders resulted in better privacy practices. However, the results vanish after the GDPR became binding, as the GDPR made all FinTechs act to ensure data privacy.

We ask whether it is possible for a user to give informed consent (art. 7 GDPR) if they cannot transparently capture the language respective to the content of privacy statements. This raises the question of whether FinTech companies have implemented the essential provisions of the GDPR and whether the regulation has achieved its goal. The answer is broadly no. Looking at the question from a company perspective, however, one has to consider whether a company can ever be fully GDPR compliant without seriously restricting its business activities. This is particularly relevant for a data-intensive and competitive industry like the FinTech industry. From the perspective of regulators, one might ask whether the GDPR is deficient in the sense that the financial industry needs to simplify the language of privacy statements so that laypeople can understand what information is being processed and how. We do not assume that laypersons will actually read privacy statements and enforce their rights (Strahilevitz & Kugler, 2016), which would be associated with far too high transaction costs. Rather, as with securities prospectuses, professional market participants such as data protection authorities are usually the addressees of privacy statements. They have the task of preparing the information and communicating it to the broader audience of customers (Firtel, 1999). So far, however, no comprehensive measures are known in which European or national data protection authorities have carried out extensive benchmarking of privacy statements. Tools supported by artificial intelligence in particular could help here, enabling consumers to have privacy statements checked online. They could examine the privacy statements for content and summarize them in simplified language.

We also provide evidence that mimicking behavior in terms of FinTech industry pressure positively influences data-privacy compliance after the GDPR became binding, which indicates that the GDPR gave companies an incentive to adopt their direct industry peers' data-processing or privacy statements. This raises the question of whether FinTech companies can gain a competitive advantage over their peers by improving their privacy policies. The current literature is inconclusive about whether high quality and easy to read privacy statements lead to a competitive advantage. Even if privacy statements are read by the customers, the one-sidedness of privacy statements will, however, most likely not trigger a race to the top (Marotta-Wurgler, 2008; Marotta-Wurgler & Chen, 2012). This would perhaps only be the case if professional market participants make privacy statements easily comparable and accessible to a broad public. Even in

such a scenario, an inferior standard could also prevail if network effects support the demand for a common, potentially inferior standard agreement (Engert & Hornuf, 2018).

Despite FinTech companies' imperfect implementation of the GDPR, our results nevertheless point to managerial recommendations. Our analysis of mimicking behavior shows, among other things, that companies take heed of the data privacy behavior of others. If data protection authorities and the media make the quality of privacy statements indeed transparent and easily accessible in the future, this could eventually lead to competition and a race to the top in privacy statement content. To excel in this competition, companies not only need to be compliant with the GDPR, but may also need to innovate in how privacy statements are agreed on. For example, users could actively give up parts of their data privacy in exchange for better prices or more usage rights, and conversely pay more to maintain greater data privacy. As is well known from the literature (Hillebrand et al., 2023), more transparency also leads to more trust and reputation gains for companies. For example, easy-to-click menus could help users prevent companies from sharing personal information with certain other companies when it is not strictly necessary for the performance of a contract. Here technical possibilities could help to enable FinTechs and consumers with a corresponding implementation.[11] Finally, the processing and forwarding of data could also be prepared and standardized in tabular form. However, standardization would require coordination among the companies in the FinTech industry and possibly new legislative initiatives.

Our article has limitations. We mainly refer to the privacy practices that companies declare in their privacy statements, and thus to the supply side of privacy (Ramadorai et al., 2021). Consumers must accept the terms for data processing if they want to use a service or product (Aridor et al., 2020). One avenue for further research is to compare what companies state in their privacy statements with the privacy practices they actually pursue. The results regarding transparency rely on our variable construction. Other approaches and methods can therefore yield different outcomes and insights. Similarly to Goldberg et al. (2021), we can only provide early evidence relating to our data-collection period shortly after the GDPR became binding in May 2018 and how the analyzed companies implemented the regulation at this point.

Finally, our article has practical implications. Legislators as well as policymakers in the EU and other countries that have adopted a privacy regulation related to the GDPR can now see the implications and the unintentional consequences of the regulation. This may pave the way towards future readjustment of the GDPR or give more practical guidance on how to create privacy statements to ensure compliance with the applicable legal standards. Further, our study emphasizes the importance of companies making greater efforts to implement effective privacy practices and communicate them to users in order to benefit from the opportunity to build a competitive advantage.

Acknowledgements We thank Fabio Bertoni, Engin Iyidogan, Youngjin Yoo, and the participants of the 7th Crowdinvesting Symposium (HU Berlin), the 5th European Alternative Finance Research Conference (University of Utrecht), the 3rd Machine Lawyering Conference (Chinese University of Hong Kong), the European Privacy InfoSec & Compliance Summit (EPIC) Summit (Munich), and the SKEMA-ESSEC 2022 Finance Workshop: FinTech and Decentralized Finance (SKEMA Business School, Nice) for their helpful comments and suggestions. Gregor Dorfleitner gratefully acknowledges the financial support by the Deutsche Bundesbank.

Funding Open Access funding enabled and organized by Projekt DEAL.

[11] There are already numerous tools that help companies to implement data privacy. See, for example, https://www.iitr.de/index.php, https://www.circle-unlimited.com/solutions/contracts/data-protection-management, https://compliance-aspekte.de/en/solutions/dsms/, https://www.dsgvo.tools, https://www.datenschutzexperte.de/dsgvo-tool/, and https://trusted.de/dsgvo-software. The providers of these tools could also extend them in such a way that a negotiation process about data transfer between companies and customers is facilitated.

Appendix

Table 6 Correlation matrix pre-GDPR

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Data Index	1							
Transparency Index	0.528	1						
No. investors	0.199	0.086	1					
Legal capital	0.185	0.256	0.065	1				
City	−0.016	−0.017	0.320	0.047	1			
Firm age	−0.002	−0.043	0.106	0.115	−0.005	1		
Bankcooperation	0.071	0.024	0.309	0.102	0.084	0.092	1	
Employees	0.205	0.078	0.577	0.066	0.183	0.235	0.198	1

Note: Correlation matrix for the data collection period before the GDPR became binding. The included variables correspond to the regression estimations in Table 4. N=276. The variables are defined in Table 1

Table 7 Correlation matrix post-GDPR

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Data Index	1									
Transparency Index	0.501	1								
No. investors	0.150	0.097	1							
Legal capital	0.114	0.110	0.061	1						
City	0.073	0.097	0.316	0.047	1					
Firm age	−0.020	−0.047	0.079	0.116	−0.008	1				
Bankcooperation	0.072	0.052	0.308	0.102	0.084	0.083	1			
Employees	0.156	0.104	0.576	0.066	0.183	0.244	0.198	1		
Mimic Data Index	0.212	0.151	0.048	0.001	−0.022	−0.103	−0.070	0.105	1	
Mimic Transparency Index	0.141	0.262	−0.129	−0.044	−0.091	−0.137	−0.135	−0.027	0.765	1

Note: Correlation matrix for the data collection period after the GDPR became binding. The included variables correspond to the regression estimations in Tables 4 and 5. N=276. The variables are defined in Table 1

Table 8 Variance inflation factors

	VIF1	VIF2	VIF3	VIF4	VIF5	VIF6
No. investors	1.85	1.85	1.85	1.88	1.71	1.73
Legal capital	1.19	1.19	1.19	1.19	1.02	1.02
City	1.18	1.18	1.18	1.18	1.12	1.12
Firm age	1.26	1.26	1.26	1.28	1.09	1.10
Bankcooperation	1.28	1.28	1.28	1.28	1.13	1.13
Employees	1.76	1.76	1.76	1.80	1.60	1.58
Mimic Data Index					1.04	
Mimic Transparency Index						1.06

Note: Variance inflation factors, VIF1-VIF4 correspond to Table 4 and models (1)-(4), VIF5-VIF6 correspond to Table 5 and models (1) and (2). The reported VIFs provide no indications for multicollinearity. N=276. The variables are defined in Table 1

Table 9 Descriptive statistics and paired t-test without mature FinTechs

Variable	*Pre-GDPR*			*Post-GDPR*					
	Mean	SD	Median	Mean	SD	Median	Diff.	t-stat.	*p*-value
Wiener Sachtext	13.655	0.894	13.730	13.847	1.152	13.855	0.192	2.254	0.025*
SMOG German	12.283	1.221	12.247	12.973	1.770	13.0923	0.690	5.395	0.000***
No. words	7.085	0.848	7.228	7.856	0.875	7.975	0.771	14.414	0.000***
Cosine similarity	0.538	0.090	0.562	0.585	0.089	0.604	0.047	7.979	0.000***
Jaccard similarity	0.209	0.046	0.217	0.228	0.044	0.241	0.019	6.325	0.000***
Euclidean distance	0.096	0.024	0.089	0.081	0.024	0.076	−0.015	−11.744	0.000***
Manhattan distance	1.306	0.132	1.271	1.252	0.126	1.228	−0.054	−6.584	0.000***

Note: Sub-sample analysis, excluding mature FinTechs, summary statistics and paired two-sided t-tests (significance level of 5%) regarding the text-based variables. N=249. The variables are defined in Table 1

Table 10 Seemingly unrelated fractional probit regression without mature FinTechs

	Dependent variable:					
	Data Index			Transparency Index		
	Pre-GDPR (1)	Post-GDPR (2)	Wald-Test *p*-value	Pre-GDPR (3)	Post-GDPR (4)	Wald-Test *p*-value
No. investors	0.036	0.021	0.671	0.076*	0.037	0.460
	(0.025)	(0.025)		(0.035)	(0.036)	
Bankcooperation	0.011	−0.018	0.684	0.038	0.028	0.917
	(0.051)	(0.047)		(0.073)	(0.070)	
Legal capital	0.280***	0.130+	0.144	0.637***	0.286**	0.030
	(0.077)	(0.067)		(0.120)	(0.108)	
City	−0.048	0.007	0.373	−0.029	0.073	0.243
	(0.045)	(0.042)		(0.060)	(0.059)	
Firm Age	−0.012	−0.001	0.880	−0.101	−0.099	0.988
	(0.055)	(0.056)		(0.070)	(0.078)	
Employees	0.028	0.040	0.740	−0.016	0.039	0.355
	(0.026)	(0.026)		(0.042)	(0.042)	
Constant	−0.827***	−0.725***		−0.594**	−0.685***	
	(0.173)	(0.172)		(0.182)	(0.180)	
Industry Effects	Yes	Yes		Yes	Yes	
Observations	249	249	249	249	249	249
Log Likelihood	−87.899	−94.058		−104.089	−104.242	

Note: Sub-sample analysis, excluding mature FinTechs, seemingly unrelated fractional probit regression regarding determinants of the quantity of data processed and transparency, Wald-Test (significance level of 5%) *p*-values to compare equality of coefficients of models (1)(2) and (3)(4), numbers in parentheses are robust standard errors. The variables are defined in Table 1. $^{+}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 11 Fractional probit regression without mature FinTechs

	Dependent variable:	
	Data Index Post-GDPR (1)	Transparency Index Post-GDPR (2)
No. investors	0.008	0.017
	(0.026)	(0.038)
Bankcooperation	0.010	0.043
	(0.048)	(0.065)
Legal capital	0.132+	0.179
	(0.070)	(0.115)
City	0.019	0.069
	(0.042)	(0.058)
Firm age	−0.007	−0.062
	(0.045)	(0.061)
Employees	0.031	0.036
	(0.026)	(0.044)
Mimic Data Index	1.402*	
	(0.715)	
Mimic Transparency Index		1.904***
		(0.440)
Constant	−1.190***	−1.322***
	(0.179)	(0.192)
Industry Effects	No	No
Observations	249	249
Log pseudolikelihood	−136.1568	−149.3130

Note: Sub-sample analysis, excluding mature FinTechs, fractional probit regression regarding mimicking behavior, numbers in parentheses are robust standard errors. The variables are defined in Table 1. $^{+}p<0.1$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 12 Pooled OLS regression with GDPR interaction

	Dependent variable:	
	Data Index (1)	Transparency Index (2)
GDPR	0.052	0.046
	(0.036)	(0.061)
No. investors	0.016*	0.023+
	(0.007)	(0.012)
GDPR x No. investors	−0.008	−0.009
	(0.010)	(0.016)
Bankcooperation	0.001	0.002
	(0.014)	(0.023)
GDPR x Bankcooperation	0.006	0.016
	(0.019)	(0.031)
Legal capital	0.060***	0.175***
	(0.017)	(0.030)
GDPR x Legal capital	−0.025	−0.092*
	(0.025)	(0.044)
City	−0.019	−0.015
	(0.012)	(0.020)
GDPR x city	0.024	0.043
	(0.017)	(0.027)
Firm age	−0.005	−0.023
	(0.012)	(0.018)
GDPR x Firm age	0.001	0.003
	(0.017)	(0.027)
Employees	0.012+	0.003
	(0.007)	(0.012)
GDPR x Employees	−0.003	0.004
	(0.010)	(0.017)
Constant	0.240***	0.248***
	(0.041)	(0.046)
Industry Effects	Yes	Yes
Observations	552	552
R^2	0.187	0.199
Adj. R^2	0.143	0.156

Note: Pooled OLS regression with GDPR interaction, including the dummyvariable *GDPR* to take into account the effects of the GDPR, numbers in parentheses are robust standard errors. The variables are defined in Table 1. $^{+}p<0.1$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 13 Composition and descriptive statistics of *Data Index* and *Transparency Index* pre-GDPR

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Data index							
Name	0.678	0.468	0.000	0.000	1.000	1.000	1.000
Gender	0.116	0.321	0.000	0.000	0.000	0.000	1.000
Title	0.036	0.187	0.000	0.000	0.000	0.000	1.000
Language	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Identifier	0.098	0.298	0.000	0.000	0.000	0.000	1.000
Password	0.145	0.353	0.000	0.000	0.000	0.000	1.000
Age	0.326	0.470	0.000	0.000	0.000	1.000	1.000
Place of birth	0.080	0.271	0.000	0.000	0.000	0.000	1.000
Address	0.572	0.496	0.000	0.000	1.000	1.000	1.000
E-mail address	0.612	0.488	0.000	0.000	1.000	1.000	1.000
Phone number	0.322	0.468	0.000	0.000	0.000	1.000	1.000
Residence city	0.029	0.168	0.000	0.000	0.000	0.000	1.000
Residence country	0.051	0.220	0.000	0.000	0.000	0.000	1.000
Marital status	0.040	0.196	0.000	0.000	0.000	0.000	1.000
Occupation	0.054	0.227	0.000	0.000	0.000	0.000	1.000
Bank	0.250	0.434	0.000	0.000	0.000	0.200	1.000
PIN	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Income	0.040	0.196	0.000	0.000	0.000	0.000	1.000
Tax residency	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Social security number	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Tax ident number	0.040	0.196	0.000	0.000	0.000	0.000	1.000
Driving license	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Passport, registration	0.069	0.254	0.000	0.000	0.000	0.000	1.000
Graduation, qualification	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Insurance	0.033	0.178	0.000	0.000	0.000	0.000	1.000
IP-address	0.141	0.349	0.000	0.000	0.000	0.000	1.000
GPS, location	0.029	0.168	0.000	0.000	0.000	0.000	1.000
Personal data published	0.149	0.356	0.000	0.000	0.000	0.000	1.000
Personal data transfer	0.158	0.365	0.000	0.000	0.000	0.000	1.000
Social Plugins, third party	0.525	0.500	0.000	0.000	1.000	1.000	1.000
Behavior, usage, movement	0.967	0.178	0.000	1.000	1.000	1.000	1.000
Google Analytics	0.826	0.380	0.000	1.000	1.000	1.000	1.000
Health	0.014	0.120	0.000	0.000	0.000	0.000	1.000
Religion	0.004	0.060	0.000	0.000	0.000	0.000	1.000
Nationality	0.083	0.277	0.000	0.000	0.000	0.000	1.000
Picture	0.072	0.260	0.000	0.000	0.000	0.000	1.000
Conversation record	0.004	0.060	0.000	0.000	0.000	0.000	1.000
Signature	0.014	0.120	0.000	0.000	0.000	0.000	1.000
Data Index	0.206	0.103	0.000	0.125	0.200	0.275	0.575
Transparency index							
Data	0.395	0.490	0.000	0.000	0.000	1.000	1.000
Purpose	0.859	0.349	0.000	1.000	1.000	1.000	1.000
Storage	0.489	0.501	0.000	0.000	0.000	1.000	1.000
Avoid	0.033	0.178	0.000	0.000	0.000	0.000	1.000
Opt-in	0.029	0.168	0.000	0.000	0.000	0.000	1.000

Table 13 continued

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Pseudo	0.014	0.120	0.000	0.000	0.000	0.000	1.000
Third	0.113	0.317	0.000	0.000	0.000	0.000	1.000
Third data	0.498	0.501	0.000	0.000	0.000	1.000	1.000
Transparency Index	0.303	0.175	0.000	0.125	0.375	0.375	0.875

Note: Composition and descriptive statistics of *Data Index* and *Transparency Index* before the GDPR became binding. N=276. The variables are defined in Table 1

Table 14 Composition and descriptive statistics of *Data Index* and *Transparency Index* post-GDPR

Statistic	Mean	SD	Min	Q1	Median	Q3	Max
Data index							
Name	0.768	0.423	0.000	1.000	1.000	1.000	1.000
Gender	0.192	0.395	0.000	0.000	0.000	0.000	1.000
Title	0.054	0.227	0.000	0.000	0.000	0.000	1.000
Language	0.014	0.120	0.000	0.000	0.000	0.000	1.000
Identifier	0.105	0.307	0.000	0.000	0.000	0.000	1.000
Password	0.199	0.400	0.000	0.000	0.000	0.000	1.000
Age	0.330	0.471	0.000	0.000	0.000	1.000	1.000
Place of birth	0.123	0.329	0.000	0.000	0.000	0.000	1.000
Address	0.580	0.495	0.000	0.000	1.000	1.000	1.000
E-mail address	0.790	0.408	0.000	1.000	1.000	1.000	1.000
Phone number	0.486	0.501	0.000	0.000	0.000	1.000	1.000
Residence city	0.025	0.158	0.000	0.000	0.000	0.000	1.000
Residence country	0.040	0.196	0.000	0.000	0.000	0.000	1.000
Marital status	0.043	0.204	0.000	0.000	0.000	0.000	1.000
Occupation	0.065	0.247	0.000	0.000	0.000	0.000	1.000
Bank	0.301	0.459	0.000	0.000	0.000	1.000	1.000
PIN	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Income	0.033	0.178	0.000	0.000	0.000	0.000	1.000
Tax residency	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Social security number	0.004	0.060	0.000	0.000	0.000	0.000	1.000
Tax ident number	0.054	0.227	0.000	0.000	0.000	0.000	1.000
Driving license	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Passport, registration	0.116	0.321	0.000	0.000	0.000	0.000	1.000
Graduation, qualification	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Insurance	0.018	0.134	0.000	0.000	0.000	0.000	1.000
IP-address	0.366	0.483	0.000	0.000	0.000	1.000	1.000
GPS, location	0.025	0.158	0.000	0.000	0.000	0.000	1.000
Personal data published	0.185	0.389	0.000	0.000	0.000	0.000	1.000
Personal data transfer	0.163	0.370	0.000	0.000	0.000	0.000	1.000
Social Plugins, third party	0.638	0.482	0.000	0.000	1.000	1.000	1.000
Behavior, usage, movement	0.949	0.220	0.000	1.000	1.000	1.000	1.000
Google Analytics	0.808	0.395	0.000	1.000	1.000	1.000	1.000
Health	0.014	0.120	0.000	0.000	0.000	0.000	1.000
Religion	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Nationality	0.101	0.302	0.000	0.000	0.000	0.000	1.000

Table 14 continued

Statistic	Mean	SD	Min	Q1	Median	Q3	Max
Picture	0.087	0.282	0.000	0.000	0.000	0.000	1.000
Conversation record	0.011	0.104	0.000	0.000	0.000	0.000	1.000
Signature	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Data Index	0.237	0.098	0.000	0.169	0.225	0.300	0.550
Transparency index							
Data	0.279	0.449	0.000	0.000	0.000	1.000	1.000
Purpose	0.920	0.271	0.000	1.000	1.000	1.000	1.000
Storage	0.406	0.492	0.000	0.000	0.000	1.000	1.000
Avoid	0.007	0.085	0.000	0.000	0.000	0.000	1.000
Opt-in	0.051	0.220	0.000	0.000	0.000	0.000	1.000
Pseudo	0.051	0.220	0.000	0.000	0.000	0.000	1.000
Third	0.076	0.266	0.000	0.000	0.000	0.000	1.000
Third data	0.569	0.496	0.000	0.000	1.000	1.000	1.000
Transparency Index	0.295	0.158	0.000	0.125	0.250	0.375	0.875

Note: Composition and descriptive statistics of *Data Index* and *Transparency Index* after the GDPR became binding. N=276. The variables are defined in Table 1

References

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM Conference on Electronic Commerce* (EC '04, pp. 21–29). New York: Association for Computing Machinery.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science, 347*(6221), 509–514.

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature, 54*(2), 442–492.

Agarwal, S., Steyskal, S., Antunovic, F., & Kirrane, S. (2018). Legislative compliance assessment: Framework, model and GDPR instantiation. In M. Medina, A. Mitrakas, K. Rannenberg, E. Schweighofer, & N. Tsouroulas (Eds.), *Privacy Technologies and Policy* (pp. 131–149). Cham: Springer International Publishing.

Ahlers, G. K. C., Cumming, D., Günther, C., & Schweizer, D. (2015). Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice, 39*(4), 955–980.

Aridor, G., Che, Y.-K., & Salz, T. (2020). *The economic consequences of data privacy regulation: Empirical evidence from GDPR.* Working Paper 26900, National Bureau of Economic Research.

Arora, C., Sabetzadeh, M., Briand, L. C., & Zimmer, F. (2014). Requirement boilerplates: Transition from manually-enforced to automatically-verifiable natural language patterns. *2014 IEEE 4th International Workshop on Requirements Patterns (RePa)* (pp. 1–8).

Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2014). Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies, 43*(1), 1–35.

Bamberger, R., & Vanecek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitsstufen von Texten in deutscher Sprache*. Vienna: Jugend und Volk Verlagsgesellschaft.

Becher, S. I., & Benoliel, U. (2021). Law in books and law in action: The readability of privacy policies and the GDPR. In K. Mathis & T. Avishalom (Eds.), *Consumer Law & Economics: Economic Analysis of Law in European Legal Scholarship* (Vol. 9, pp. 179–204). New York: Springer.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845–2897.

Bernstein, S., Korteweg, A., & Laws, K. (2017). Attracting early-stage investors: Evidence from a randomized field experiment. *The Journal of Finance, 72*(2), 509–538.

Betzing, J. H., Tietz, M., vom Brocke, J., & Becker, J. (2020). The impact of transparency on mobile privacy decision making. *Electronic Markets, 30*, 607–625. https://doi.org/10.1007/s12525-019-00332-3.

Biasiotti, M., Francesconi, E., Palmirani, M., Sartor, G., & Vitali, F. (2008). *Legal informatics and management of legislative documents.* Working Paper 2, Global Center for ICT in Parliament.

Brown, S. V., & Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research, 49*(2), 309–346.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *The Journal of Finance, 75*(3), 1371–1415.

Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.-W., Palka, P., Sartor, G., & Torroni, P. (2018). *Claudette meets GDPR: Automating the evaluation of privacy policies using artificial intelligence.* Working Paper 3208596, Social Science Research Network.

Cudd, M., Davis, H. E., & Eduardo, M. (2006). Mimicking behavior in repurchase decisions. *Journal of Behavioral Finance, 7*(4), 222–229.

Cumming, D., Meoli, M., & Vismara, S. (2019). Investors' choices between cash and voting rights: Evidence from dual-class equity crowdfunding. *Research Policy, 48*(8), 103740.

Cumming, D. J., Leboeuf, G., & Schwienbacher, A. (2020). Crowdfunding models: Keep-it-all vs. all-or-nothing. *Financial Management, 49*(2), 331–360.

Cumming, D. J., & Schwienbacher, A. (2018). Fintech venture capital. *Corporate Governance: An International Review, 26*(5), 374–389.

De Clercq, D., & Dimov, D. (2008). Internal knowledge development and external knowledge access in venture capital investment performance. *Journal of Management Studies, 45*(3), 585–612.

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2019). We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *26th Annual Network and Distributed System Security Symposium, NDSS 2019*. San Diego: The Internet Society.

Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research, 17*(1), 61–80.

Dorfleitner, G., & Hornuf, L. (2019). *FinTech and Data Privacy in Germany: An Empirical Analysis with Policy Recommendations*. Cham: Springer International Publishing.

Dorfleitner, G., Hornuf, L., Schmitt, M., & Weber, M. (2017). *FinTech in Germany*. Cham: Springer International Publishing.

Drasch, B. J., Schweizer, A., & Urbach, N. (2018). Integrating the 'troublemakers': A taxonomy for cooperation between banks and fintechs. *Journal of Economics and Business, 100*, 26–42.

Duchesneau, D. A., & Gartner, W. B. (1990). A profile of new venture success and failure in an emerging industry. *Journal of Business Venturing, 5*(5), 297–312.

Earp, J. B., Anton, A. I., Aiman-Smith, L., & Stufflebeam, W. H. (2005). Examining internet privacy policies within the context of user privacy values. *IEEE Transactions on Engineering Management, 52*(2), 227–237.

Engert, A., & Hornuf, L. (2018). Market standards in financial contracting: The euro's effect on debt securities. *Journal of International Money and Finance, 85*, 145–162.

Ermakova, T., Baumann, A., Fabian, B., & Krasnova, H. (2014). Privacy policies and users' trust: Does readability matter? *Americas Conference on Information Systems*. Savannah.

Fabian, B., Ermakova, T., & Lentz, T. (2017). Large-scale readability analysis of privacy policies. *Proceedings of the International Conference on Web Intelligence*. (WI '17, pp. 18–25). New York: Association for Computing Machinery.

Fernback, J., & Papacharissi, Z. (2007). Online privacy as legal safeguard: the relationship among consumer, online portal, and privacy policies. *New Media & Society, 9*(5), 715–734.

Firtel, K. B. (1999). Plain English: A reappraisal of the intended audience of disclosure under the securities act of 1933. *Southern California Law Review, 72*, 851–898.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221.

Gai, K., Qiu, M., Sun, X., & Zhao, H. (2017). Security and privacy issues: A survey on fintech. In M. Qiu (Ed.), *Smart Computing and Communication* (pp. 236–247). Cham: Springer International Publishing.

Gazel, M., & Schwienbacher, A. (2021). Entrepreneurial fintech clusters. *Small Business Economics, 57*, 883–903.

Goldberg, S. G., Johnson, G. A., & Shriver, S. K. (2021). *Regulating privacy online: An economic evaluation of the GDPR.* Working Paper 3421731, Social Science Research Network.

Gunning, R. (1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

Hajduk, P. (2021). The powers of the supervisory body in GDPR as a basis for shaping the practices of personal data processing. *Review of European and Comparative Law, 45*(2), 57–75.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., & Aberer, K. (2018). Polisis: Automated analysis and presentation of privacy policies using deep learning. In *USENIX Security Symposium* (pp. 531–548).

Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing value from big data - a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management, 36*(10), 1382–1406.

Hillebrand, K., Hornuf, L., Müller, B., & Vrankar, D. (2023). The social dilemma of big data: Donating personal data to promote social welfare. *Information and Organization, 33*(1), 100452.

Hornuf, L., Kloehn, L., & Schilling, T. (2018). Financial contracting in crowdinvesting: Lessons from the German market. *German Law Journal, 19*(3), 509–578.

Hornuf, L., Klus, M. F., Lohwasser, T. S., & Schwienbacher, A. (2021). How do banks interact with fintech startups? *Small Business Economics, 57*, 1505–1526.

Hornuf, L., Schilling, T., & Schwienbacher, A. (2021b). The relevance of investor rights in crowdinvesting. *Journal of Corporate Finance* (pp. 101927).

Hornuf, L., Schmitt, M., & Stenzhorn, E. (2018). Equity crowdfunding in Germany and the United Kingdom: Follow-up funding and firm failure. *Corporate Governance: An International Review, 26*(5), 331–354.

Hsu, D. H. (2006). Venture capitalists and cooperative start-up commercialization strategy. *Management Science, 52*(2), 204–219.

Ingram Bogusz, C. (2018). Digital traces, ethics, and insight: Data-driven services in FinTech. In R. Teigland, S. Siri, A. Larsson, A. M. Puertas, & C. Ingram Bogusz (Eds.), *The Rise and Development of Fintech: Accounts of Disruption from Sweden and Beyond* (pp. 207–222). London: Routledge.

Kahan, M., & Klausner, M. (1997). Standardization and innovation in corporate contracting (or the economics of boilerplate). *Virginia Law Review, 83*(4), 713–770.

Kaur, J., Dara, R. A., Obimbo, C., Song, F., & Menard, K. (2018). A comprehensive keyword analysis of online privacy policies. *Information Security Journal: A Global Perspective, 27*(5–6), 260–275.

Kondra, A. Z., & Hinings, C. R. (1998). Organizational diversity and change in institutional theory. *Organization Studies, 19*(5), 743–767.

Kubick, T. R., Lynch, D. P., Mayberry, M. A., & Omer, T. C. (2015). Product market power and tax avoidance: Market leaders, mimicking strategies, and stock returns. *The Accounting Review, 90*(2), 675–702.

Laursen, K., & Salter, A. J. (2014). The paradox of openness: Appropriability, external search and collaboration. *Research Policy, 43*(4), 867–878.

Lewis, S. D., Colvard, R. G., & Adams, C. N. (2008). A comparison of the readability of privacy statements of banks, credit counseling companies, and check cashing companies. *Journal of Organizational Culture, Communications and Conflict, 12*(2), 87–93.

Li, H., Yu, L., & He, W. (2019). The impact of GDPR on global technology development. *Journal of Global Information Technology Management, 22*(1), 1–6.

Linden, T., Khandelwal, R., Harkous, H., & Fawaz, K. (2020). The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* (pp. 47–64).

Lindgreen, E. R. (2018). Privacy from an economic perspective. *The Handbook of Privacy Studies: An Interdisciplinary Introduction* (pp. 181–208). Amsterdam: Amsterdam University Press.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance, 69*(4), 1643–1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54*(4), 1187–1230.

Mac an Bhaird, C., & Lucey, B. (2010). Determinants of capital structure in Irish SMEs. *Small Business Economics, 35*, 357–375.

Marotta-Wurgler, F. (2008). Competition and the quality of standard form contracts: The case of software license agreements. *Journal of Empirical Legal Studies, 5*(3), 447–475.

Marotta-Wurgler, F. and Chen, D. L. (2012). Does contract disclosure matter? *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft, 168*(1) 94–123.

Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. *Journal of Marketing, 81*(1), 36–58.

Martin, N., Matt, C., Niebel, C., & Blind, K. (2019). How data protection regulation affects startup innovation. *Information System Frontiers, 21*, 1307–1324.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading, 12*(8), 639–646.

Miller, A. R., & Tucker, C. (2009). Privacy protection and technology diffusion: The case of electronic medical records. *Management Science, 55*(7), 1077–1093.

Mohan, J., Wasserman, M., & Chidambaram, V. (2019). Analyzing GDPR compliance through the lens of privacy policy. In V. Gadepally, T. Mattson, M. Stonebraker, F. Wang, G. Luo, Y. Laing & A. Dubovitskaya (Eds.), *Heterogeneous Data Management, Polystores, and Analytics for Healthcare* (Lecture Notes in Computer Science, pp. 82–95. DMAH 2019, Poly 2019). Cham.

Mulder, T., & Tudorica, M. (2019). Privacy policies, cross-border health data and the GDPR. *Information & Communications Technology Law, 28*(3), 261–274.

Müller, N. M., Kowatsch, D., Debus, P., Mirdita, D., & Böttinger, K. (2019). On GDPR compliance of companies' privacy policies. In K. Ekstein (Ed), *Text, speech, and dialogue. TSD 2019. Lecture notes in computer science 11697*. (pp. 151–159). Cham: Springer.

O'Donoghue, T., & Rabin, M. (2000). The economics of immediate gratification. *Journal of Behavioral Decision Making, 13*(2), 233–250.

Omri, B.-S., & Schneider, C. E. (2014). *More than you wanted to know: the failure of mandated disclosure*. Princeton: Princeton University Press.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619–632.

Peacock, C., Milewicz, K., & Snidal, D. (2019). Boilerplate in international trade agreements. *International Studies Quarterly, 63*(4), 923–937.

Peterson, K., Schmardebeck, R., & Wilks, T. J. (2015). The earnings quality and information processing effects of accounting consistency. *The Accounting Review, 90*(6), 2483–2514.

Porter, M. E. (1998). Clusters and the new economics of competition. *Harvard Business Review*, *76*(6), 77–90.

Posner, R. A. (1981). The economics of privacy. *The American Economic Review, 71*(2), 405–409.

Ramadorai, T., Uettwiller, A., & Walther, A. (2021). *The market for data privacy.* Working Paper 3352175, Social Science Research Network.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM, 18*(11), 613–620.

Singh, R. I., Sumeeth, M., & Miller, J. (2011). A user-centric evaluation of the readability of privacy policies in popular web sites. *Information Systems Frontiers, 13*, 501–514.

Stewart, H., & Jürjens, J. (2018). Data security and consumer trust in fintech innovation in Germany. *Information and Computer Security, 26*(1), 109–128.

Strahilevitz, L. J., & Kugler, M. B. (2016). Is privacy policy language irrelevant to consumers? *The Journal of Legal Studies, 45*(S2), 69–95.

Sunyaev, A., Dehling, T., Taylor, P. L., & Mandl, K. D. (2015). Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association, 22*(1), 28–33.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., & Serna, J. (2018). PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In IWSPA '18 (Ed.), *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics* (pp. 15–21). New York: ACM.

Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research, 22*(2), 254–268.

Wachter, S. (2018). The GDPR and the internet of things: a three-step transparency model. *Law, Innovation and Technology, 10*(2), 266–294.

Weesie, J. (1999). Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Technical Bulletin, 9*(52), 231–248.

Wild, F. (2007). An LSA package for R. In F. Wild, M. Kalz, J. van Bruggen & R. Koper (Eds.), *Mini-Proceedings of the 1st European Workshop on Latent Semantic Analysis in Technology-Enhanced Learning* (pp. 11–12). Heerlen.

Wolff, J., & Atallah, N. (2021). Early GDPR penalties: Analysis of implementation and fines through May 2020. *Journal of Information Policy, 11*(3748837), 63–103.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Working Party on the Protection of Individuals with Regard to the Processing of Personal Data (2018). *Guidelines on Transparency under Regulation 2016/679*. WP260 rev.01.

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. Washington, D.C.: World Bank.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*(298), 348–368.

Zhang, Y., Wang, T., & Hsu, C. (2020). The effects of voluntary GDPR adoption and the readability of privacy statements on customers' information disclosure intention and trust. *Journal of Intellectual Capital, 21*(2), 145–163.