BIG DATA ANALYTICS, ONLINE TERMS OF SERVICE AND PRIVACY POLICIES[*]

Przemysław Pałka[**] & Marco Lippi[***]

Abstract

In the digital era, consumers continuously use online services and apps for their daily activities. Too often, they do so without having a clear idea about what exactly they have agreed to or how their data is being used by the online platforms and service providers. This is because online terms of use and privacy policies are typically complex documents that are hard for the average citizen to decipher. In a sense, these terms of use and privacy policies are becoming "big data" collections themselves, representing an opportunity for new approaches in artificial intelligence and machine learning. For example, machine learning and artificial intelligence could be used to detect, extract, and categorize relevant information from the terms of use and privacy policies. In this chapter, we survey the research that addresses the recent efforts made to empower consumers via technologies that provide for the automated analysis of terms of service and privacy policies.

INTRODUCTION

Ouch, that hurts! How did you get here, once again?

Two hours ago, when scrolling through Facebook on your phone, you saw that article about jogging. It is really good for you, it seems. You decided to buy running shoes online, but were interrupted by an email from your friend. After responding, you forgot about the shoes and went on Twitter to share an idea you had while talking to your buddy. Then that other article intrigued you. While reading it, you saw an ad—shoes! Right, you were supposed to take up running! You

bought the shoes with a couple of clicks (at this point you no longer know how exactly money comes off your credit card, but it works) and, excited, you decided to go jogging immediately. You downloaded Endomondo, to track your speed, put on some music on Spotify, and jogged off. And suddenly you feel that pain in your ankle (perhaps you should have waited for the new running shoes to arrive). You Googled your symptoms, and it appears that you may have real done some damage to your ankle. So you take an Uber to a doctor to have it checked out. The doctor has excellent ratings online, so you should be fine. On the way to the doctor, you Skype your friend about your injury. Now you are waiting in the doctor's waiting room playing a mobile game, and you are getting increasingly annoyed by an ad for a medical app that keeps popping up. Argh!

Question: How many terms of service[1] (ToS) and privacy policies (PPs) have you agreed to during the last two hours? To whom did you grant permission to gather data about you? And what kind of data? With whom has this data been shared? Who knows that you are about to see a doctor in a minute?

That we live in the "age of big data" seems too obvious to state by now. So does the fact that everything we do online—and many of us are constantly online[2]—leaves a digital footprint. Numerous entities—such as news, social media, and entertainment sites—are gathering information about us, tracking us, creating profiles, and targeting us with personalized commercial communications.[3] Thanks to technological advances like machine learning, it is now possible to generate knowledge and value out of incomprehensibly large data sets. Corporations not only know a lot about us, but are also able to predict and influence our behavior as consumers[4] and political actors (consider the Cambridge Analytica scandal). The imbalance of power between big business and consumers appears to be constantly increasing. In many ways, this is a scary new reality in which we find ourselves. Strictly speaking, we have agreed to all this. We are happy to gain access to all of these services at no cost,[5] and most of us will not take the time to read all the privacy policies and terms of service. "I have read the terms" is said to be the most common lie on the planet. So have we truly consented to our surveillance?[6]

It has been empirically demonstrated that users do not read the privacy policies and terms of service they accept.[7] Even when they do, they often do not understand what these documents mean.[8] The language used in ToS and PPs tends to be vague and the design misleading.[9] Hence, users wishing to actually consult ToS and PPs face cognitive and structural difficulties.[10] There is simply too much to read and understand. It has been estimated that reading the policies of all the

websites visited during a year would take an average user between 80 and 300 hours.[11] From the perspective of a user, ToS and PPs are in themselves big data.

Fortunately, at this point in time this "big data" can be analyzed automatically. Big data analytics can be used to empower individuals and civil society.[12] These technologies no longer have to benefit only corporations and the state. This chapter hopes demonstrate that machine learning can be used to read and analyze ToS and PPs for the benefit of consumers. In this chapter, we take stock of the current state of scholarship regarding this issue, discuss areas where we believe progress can still be made, and examine the technological, legal and market conditions that would allow such technology to be employed on a large scale.

In the next Part we provide basic definitions and context, both from the perspective of computer science (what is big data analytics?) and of the law (what is the legal status of ToS and PPs?). We survey the legal environment in which ToS and PPs operate, paying special attention to differences between the American and European approaches. In Part II we provide an overview of the ways in which different big data analytics technologies—specifically, machine learning—are being employed (or could be employed) to automate the analysis of ToS and PPs. We offer an overview of the research literature and the projects/tools currently available. Throughout this part of the paper, we try to offer explanations that can be understood by a non-technical audience while giving due recognition to the technical sophistication of the projects we analyze. In Part III we present a series of user stories in order to analyze how these techniques could be useful for individual users, academics, regulators, businesses and civil society at large. We then offer an analysis of the preconditions necessary for such applications to be turned from lab-projects into actual tools used in the real life. We close with certain policy recommendations, ultimately arguing that the role of law and policymakers is not only to update regulations so that they better suit the challenges of the big data era, but also to enable the development of technologies that benefit the public and individual consumers.

I. DEFINITIONS AND CONTEXT

Big data analytics, and machine learning in particular, can be used to develop applications that automatically analyze ToS and PPs. Because this technology has the potential to greatly benefit consumers, it is important for the public to understand what exactly is meant by "big data" and "big data analytics." What legal tasks can be automated using these technologies? What is the

law governing the legal status of PPs and ToS? In this section we offer definitions of key terms and discuss the technical and legal context in which PPs, ToS, and big data analytics operate.

A. Computer Science Perspective

The term "big data" typically refers to very large data collections, and also to the set of technologies, platforms, and infrastructures that allow the management of such data collections. For example, all the photos of cats on the internet are "big data." The shopping history of all Amazon users is "big data." From the perspective of a user, all the ToS and PPs he or she has accepted are "big data." The term "big data analytics" is the more accurate term used to describe the technologies that one can employ to make sense of the "big data" itself.

In general, big data are described using the so-called "3Vs," i.e., volume, velocity, and variety.[13] The "3Vs" indicate that nowadays data collections are huge (volume), grow at an extremely fast rate (velocity), and are heterogeneous (variety). Other "Vs" associated with big data are "veracity," which refers to data trustworthiness and integrity, and "value," which indicates that such enormous amounts of information hide precious granules of knowledge.[14] Big data analytics is the process of extracting value from the raw data. In that pursuit, it typically relies on technologies from machine learning, artificial intelligence, data science, computer science, and other disciplines.

More specifically, artificial intelligence and machine learning methodologies provide algorithms for the detection of interesting data patterns and are also used for the classification of data into predetermined categories. Furthermore, algorithms can be used to rank data according to some preference criterion or cluster data with respect to some similarity measure. For example, AI can be employed to teach a computer to recognize if there is a cat in a picture or, importantly for our purposes, to check whether an arbitration agreement is present in any of the terms of service that a given user has accepted. Most of these tasks require the availability of supervised data, which is data that has been manually annotated by experts. This annotation process allows a machine to be trained to produce the desired output from the raw input. Put simply, for a machine to be able to tell if there is a cat in a picture, or an arbitration clause in a contract, if first needs to be shown a significant amount of examples of cats or arbitration clauses. Therefore, a team of humans must first mark the arbitration clauses in many real world contracts or indicate that a picture features a

cat. This annotation process is called "tagging," and when a machine learns from a data set earlier prepared by humans, one speaks of "supervised learning."

"Unsupervised learning" is a different machine learning technique. Instead of relying on supervised data, it typically looks on its own for similarities and patterns in large amounts of data. For example, a machine can be fed thousands of photos, or privacy policies, to get used to how they are structured, what elements occur there in what relation to one another, etc. Supervised and unsupervised approaches can also be combined, so that unsupervised learning can be first exploited to analyze raw data before the machine is trained to perform a certain task using supervised data. In our case, this means that if we first show a computer a really large number of PPs and ToS, and then train it to detect some clauses on a smaller set of supervised data, it will usually fare better than without the unsupervised component. This process will be discussed in more detail in Part II.

B. Legal Perspective

The legal status of ToS and PPs is less clear than one might expect. Even though they seem to be everywhere nowadays, some basic questions regarding their form and content remain unresolved. Are these documents contracts? Are there certain elements that must (or must not) be included in them? And what are the consequences for violating these rules? Answers to these questions have not yet been addressed comprehensively by legislation, regulation or case law. Moreover, the answers differ across jurisdictions, including a quite striking divergence between the United States and the European Union.

ToS are generally treated as contracts of adhesion (or "boilerplate" contracts) by lawyers on both sides of the Atlantic.[15] As long as the user is not acting in his or her professional capacity, these documents are subject to the rules regarding consumer contracts. In the EU, terms of service are not to contain so-called "unfair contractual clauses."[16] The Unfair Contractual Terms Directive states:

> A contractual term which has not been individually negotiated shall be regarded as unfair if, contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer.[17]

This general definition has been concretized by the Annex to the Directive, and by more than thirty judgments by the Court of Justice of the European Union.[18] One should note, however, that this law applies to all consumer contracts, both online and offline. Examples of unfair clauses specific to ToS include: provisions giving service providers a unilateral right to change or terminate the ToS; choice of law and jurisdiction clauses; certain types of limitations of liability; obligatory arbitration clauses, and providers' rights to remove content without reason or notice.[19] If providers choose to insert such clauses into their ToS nevertheless, they do not bind the users. Moreover, various enforcement agencies and civil society organizations have competence to dispute the terms (without an individual consumer's involvement). Through this measure and others, these organizations and agencies can pressure the platforms to change their ToS to be more consumer-friendly. Note that the exact structure of the enforcement systems differs from Member State to Member State.[20] This process is in line with the organic integration of constitutional values into the European private law[21] and the European view that contracts are something to be "regulated," when necessary, through administrative measures.

In the United States, the regulations relating to ToS are much more relaxed. Consumer contracts are typically enforced; however, under the unconscionability doctrine, courts will refuse to enforce a contract if it has been concluded in circumstances that deprived the weaker party of meaningful choice, and if its terms unreasonably favor the other party.[22] The purpose of the unconscionability doctrine resembles that of the European regulations on unfair contractual terms. However, recent Supreme Court case law confirming the validity of arbitration clauses and class action waivers in ToS demonstrates that consumer protection in the U.S. is not currently being strengthened in this regard.[23] In short, there are two important differences between the American and the European systems. In the U.S., the question of what is an "unfair" term is much less clear as a matter of law. Moreover, American consumers and civil society organizations cannot, unlike in the EU, initiate administrative proceedings against the platforms' terms of service.

PPs are also treated differently in the two jurisdictions, and, in fact, here the difference runs much deeper. In the EU, the foundational law governing PPs is the General Data Protection Regulation.[24] The GDPR applies directly throughout the EU (as a type of "federal"[25] law), as well as to data controllers located outside of the EU, but directing their services at the EU's residents. The wide-spanning regulation applies to all data controllers (public and private, across all sectors), with certain exceptions. It is based on several principles: lawfulness, fairness, transparency,

purpose limitations, data minimization, accuracy, storage limitations and security.[26] These principles translate into numerous obligations on the side of data controllers, and are enforced through the combination of administrative actions by supervisory authorities, and private enforcement by data subjects and the civil society.

Within this system, every entity that processes personal information must post on its website a "privacy notice" in plain and intelligible language, conveying certain types of information to data subjects. From the European point of view, having a privacy policy is an administrative requirement. There is a clear standard for assessment of an entity's compliance with the law, and a whole array of regulatory agencies is competent to enforce the law through, for example, the imposition of the (in)famous four percent of yearly revenue fines.[27]

In the U.S., the privacy regulation landscape looks very different. First, as of 2019, there is no general federal regulation for consumer privacy or data protection (although several bills have recently been proposed). State laws differ from one another, with California being the leader in regulation as the first state to require websites to publish PPs.[28] Some federal laws have been created for specific sectors, but their scope of application is limited. The backbone of the American system is the "notice and choice" model, developed and promoted by the Federal Trade Commission.[29] This model favors self-regulation, and is based on the idea that as long as companies enable users to learn what they do with the personal information (notice), users should be able to choose whether or not to use their services (choice). As a result, no federal regulations specifying what exactly should be included in these policies exist. This model is grounded in the market-based logic of fair dealing, as opposed to the European paradigm of fair processing, which is grounded in the logic of human rights.[30]

Moreover, the legal status of PPs is still an open question in the U.S. Solove and Schwartz claim that even though plaintiffs have often argued that PPs should be treated as contracts, currently contract law plays a minimal role in courts' decision-making.[31] On the other hand, research by Bar-Gill et al. indicates that courts seem to agree on the contractual nature of privacy policies.[32] Whether PPs are (or should be) treated as contracts is debatable, but what is clear is that FTC enforces them as promises made by the companies. This is a very different approach from the European one, where PPs are instruments required by law to disclose information about processing, but are not (yet) treated as promises in any sense. Further, arguments have been raised to support the claim that FTC starts to develop extra-contractual standards of fairness applicable

to privacy policies.[33] This could mean that the American landscape might be moving in the direction similar to the standards of the GDPR. However, the actual specification of what exactly these standards are, remains, as of 2019, more an academic project than a regulatory reality.

The difference between the EU and U.S. regulatory environments directly influences the legal-tech projects undertaken on both sides of the Atlantic. Whereas in Europe one can observe attempts to automate *evaluations* of ToS and PPs, in the US the emphasis is on the *understanding and summarization* of these documents (since the idea is that it is up to consumers to decide whether they consider the deal to be fair). In the next section, we provide an overview of various projects currently using big data analytics to process PPs and ToS in the EU and US.

II. STATE OF THE ART

The application of big data analytics to the quantitative and qualitative analysis of PPs and ToS is a recent phenomenon. Most of the existing publications, projects, platforms and tools (either software products or research prototypes) have been developed in the last five years, with a significant increase in the last couple of years. In this section we provide an overview of big data analytics as applied to ToS and PPs. First, we survey the literature, and then we look at the actual tools developed by various research teams.

A. Methods and Tasks

When it comes to applications of big data analytics, the methodology used will depend both on the "real world" task that a researcher has in mind, and the techniques available to realize this task. For example, we might want to create a tool that will tell us whether there are any choices hidden in the terms of service (like an arbitration opt-out) so that the user can take advantage of making these choices. Or we might be interested in summarizing a privacy policy, paying special attention to certain types of information, such as which third party entities will have access to our personal data. Or we might be looking for clauses considered "unfair" in a given jurisdiction, so that a user can (automatically) alert the NGOs combating them with a hope that the ToS will be changed. For a lawyer, these appear to be very different problems to be solved. For an engineer, some of these tasks can be addressed using the same methods. From the point of view of machine learning, it does not matter if we are trying to detect an arbitration clause in order to make a choice about it, or because we just want to know if it is there, or because we want to know if it is

considered unfair according to some metric. A machine just learns to look for something. What action is then undertaken after that discovery is matter of software implementation,[34] not necessarily machine learning. Furthermore, to successfully realize a particular task, various big data analytics techniques might be used at different stages of the project.

Researchers have pursued many different data analytics tasks using various techniques. A common element of all the existing big data analytics approaches to ToS and PPs analysis is the use of sophisticated machine learning and so called "natural language processing" techniques. The latter captures and extracts relevant characteristics of a given text.

1. Text Categorization

The classification of text can be used to address a wide variety of problems, such as the detection of clauses with specific characteristics. For example, text classification can identify potentially unlawful clauses that include problematic statements.[35] Text classification can also be used to check the completeness of a document according to a predetermined standard of assessment.[36]

Another common application is the categorization of paragraphs or clauses into semantic classes. This can in turn be used to summarize the document[37] or to extract text segments related to certain content categories.[38] According to the detail and specificity of annotations, a wide category of problems can be addressed using this approach. Examples include the identification of choices provided in privacy policies[39] or the detection of problematically vague language.[40]

By exploiting text categorization techniques, higher-level tasks can also be realized, for example marking a document with a score that indicates the degree of compliance of the policy.[41] For example, policies describing IoT devices have been evaluated in this sense, yet without machine learning techniques.[42] Clearly, using automatic text categorization would allow to move the analysis to a much larger scale.

2. Knowledge representation and information extraction

Knowledge extraction is another research field in artificial intelligence that can be applied to ToS and PPs. In essence, it involves the automatic extraction of facts, statements, rules (usually referred to as "knowledge") that are represented or encoded (thus the term "representation") into a structured, formal language that can be efficiently searched and updated by a machine (e.g., logic

facts, or ontologies). For example, in the work of Joshi et al., natural language processing and rule-based approaches are used to extract statements of permission and obligation in the form of so called "deontic logic rules."[43] This kind of approach can be used in scenarios such as question answering, where there is need to efficiently retrieve information in order to automatically answer users' questions.[44] Ontologies have also been recently used as a way to model concepts related to privacy legislation and to encode and represent so-called "Privacy Level Agreements."[45] These agreements are typically adopted by cloud service providers to describe their data protection practices.[46]

Put simply, in the text categorization techniques, the machine does not "know" anything about the law or the actual contents of the documents it analyzes—it is simply "taught" to label some parts of the text with categories predetermined by a human. Knowledge representation and information extraction are techniques that go beyond that and actually "teach" the machine something about the matter, so that it can handle more complex tasks.

3. Unsupervised Learning

Most of the aforementioned tasks are supervised. In other words, for machines to learn how to realize certain tasks, a group of humans must manually annotate the documents first. Someone needs to teach the machine to recognize arbitration clauses in ToS by showing it dozens and dozens of contracts with the arbitration clauses highlighted so that it can 'learn' the characteristics of these clauses. Then a human must test whether the teaching was successful. This is clearly a very time-consuming and costly process. A major challenge of machine learning is that of facing also scenarios where these human "supervisions" are rare, or even completely absent. Generally speaking, in "unsupervised learning" projects we have (large) datasets available, but we are not giving a precise task to the machine—so no external supervision or "ground truth" target is made available to be learned by a machine. This can happen both because building supervised data is costly but also because, in some cases, it can be very challenging to formally define (and thus collect) a precise and appropriate target. The typical goal of unsupervised learning is thus that of finding similarities, correlations, and frequent patterns in large data collections.

In the context of PPs and ToS, unsupervised learning is a framework which has recently been gaining attention. For example, in a recent work, an unsupervised learning approach is used to align policies, so that sections regarding the same topics (e.g., statements regarding advertising,

or paragraphs describing children-related data) can be easily retrieved and compared.[47] Another very promising research direction is that of exploiting large collections of unsupervised data to capture characteristics of the language used in privacy policies, so that they can be used as input features for machine learning classifiers.[48] In this case, unsupervised learning is used as a preparatory process for the subsequent supervised task. Put simply, if the computer first gets a chance to "read" several thousand documents without a predetermined task, just to "get used to" their structure, lexicon etc., it can be more successful at the later stage, when learning how to do something on a much smaller set of ToS or PPs annotated by humans.

Finally, it is worth noting that a possible solution for the creation of larger corpora of certain documents is crowdsourcing. Crowdsourcing leverages the power of the crowd in order to reach a high-quality consensus.[49] For example, different NGOs or research teams could enrich a database of annotated ToS or PPs "as they go" or civic-minded consumers could enrich a database using a similar process to Wikipedia's crowd-editing function. With crowdsourcing, there are many challenges to face, particularly regarding the way in which questions should be posed to the public so that useful information can be gathered.

The main limitation of the approaches used so far is that they are based on classical, off-the-shelf methodologies. As tasks become more and more complicated, there will be a need for more sophisticated machine learning techniques that are capable of combining such classifiers with high-level reasoning capabilities.[50]

B. Platforms and Tools

Besides producing very interesting technical reports and publications, the abovementioned research methods have put in motion the development of software products, platforms, and tools for the benefit of end-users. Below we describe such tools and their application to the tasks they attempt to automate. Then, in the next Part, we will illustrate how these newly developed tools can be used by multiple different actors.

1. Usable Privacy

Usable Privacy (https://usableprivacy.org/) is a web platform for the research project bearing the same name. Founded in 2013, the project has produced a large number of publications across many disciplines. The goals of the project are to (1) extract the key features from natural

language PPs, and (2) present these features to users in an easy-to-digest format that enables them to make more informed privacy decisions as they interact with different websites. The overall purpose is thus to enhance the public's understanding of what is contained in a PP (which users typically otherwise do not read or do not understand). The platform offers an online tool that annotates and categorizes several parts of a PP, as well as a very large data set of annotated policies. The whole project builds upon a pioneering application named TAPPA (Toolkit for Automatic Privacy Policy Analysis). TAPPA annotates policies using metadata to allow for a more complex analysis (https://cups.cs.cmu.edu/tappa/).

2. Polisis and Pribot

Polisis (https://pribot.org/polisis) and Pribot (https://pribot.org/) are two web platforms dedicated to the analysis of PPs.[51] In particular, Polisis is a tool that can automatically scan and annotate segments of PPs with a set of labels describing some characteristics of the policy. For example, Polisis can classify text portions according to semantic categories (i.e., third-party sharing, security, data retention, etc.). Pribot is instead a chatbot that can answer questions expressed in natural language regarding one particular privacy policy. Browser extensions are available as well. The overall goal is again that of enhancing public understanding of PPs. As for their methodologies, the two systems use a combination of techniques, including state-of-the-art deep learning approaches for natural language processing.

3. Claudette

The Claudette Project (https://claudette.eui.eu/) builds systems that automatically detect potentially unlawful statements in ToS and PPs.[52] The task is slightly different to that of the aforementioned systems, since in this case the main goal is the evaluation of documents according to some legal standard; however, the output is not supposed to be definitive, but rather indicative. The idea is not to completely replace the human assessment by a machine, but rather to make a human lawyer's work easier and faster by highlighting potentially unlawful clauses Regarding ToS, an online web server is made available to which users can submit the plain text of a contract and receive the predictions (i.e., the annotations) made by Claudette. The system uses a collection of different machine learning systems that rely on lexical information. A similar service for PPs is currently under development.[53] Unlike the previous tools—Usable Privacy, and Polisis and Pribot,

which concentrate on making it easier for users to understand the statements the privacy policies contain, the Claudette Project's goal is to assess the compliance of a given ToS or PP with the EU consumer and data protection legislation.

4. PrivOnto

PrivOnto is a semantic framework that enables formal representation of the content of a privacy policy.[54] By exploiting background knowledge of the topic, encoded in a widely employed formalism in computer science that is named "ontology," PrivOnto has the dual objective of answering privacy questions of interest to users and supporting researchers and regulators in the wide-scale analysis of PPs. PrivOnto's interactive online tool can be used to explore a corpus of pre-annotated documents.

5. PrOnto

PrOnto is an ontology for the representation of legal concepts within the GDPR, including agents, data types, types of processing operations, rights and obligations.[55] The ontology is integrated with deontic logic models in order to support legal reasoning. The framework is designed so as to target practitioners, and it can also be used in combination with natural language processing systems.

6. PrivacyCheck

PrivacyCheck (https://identity.utexas.edu/privacycheck-for-google-chrome) is a browser extension that automatically summarizes privacy policies so that consumers can be given an overview of the data practices of a given service. In particular, the application is oriented toward the prevention of identity theft. Advanced data mining algorithms extract and categorize information according to the level of risk associated to each data practice.

7. ConPolicy

ConPolicy (https://dseanalyser.pguard-tools.de/) is a project that extracts and categorizes information from privacy policies using machine learning and data mining methodologies. A web server is currently available as a prototype. The application focuses on the German language, with a dedicated annotated corpus for this language. Among other categories, ConPolicy annotates

sentences with vague or unclear language and text portions that deal with specific topics personalized advertisement, transfer of data to third parties etc.

8. AppTrans

AppTrans (Transparency for Android Applications) is a research project that develops digital technologies that can enhance the transparency of data practices.[56] The project is focused on mobile applications and compares their declared data practices with their actual practices so that any discrepancies can be detected. AI-based technologies are used to automatically scan and analyze privacy policies and to extract the relevant content, while data-flow analysis tools are used to analyze the application code and detect whether the actual data practice is compliant with the policy.

9. Privacy-Avare

Privacy-Avare (www.privacy-avare.de) is a tool that manages one's privacy preferences across different devices, such as mobile phones, smart homes, and intelligent cars. The key idea of the tool is to set up a privacy profile for the user according to the user's preferences and then check whether the services he/she is using are compliant with the profile. The tool thus searches for violations of the rules set in the user preferences and proposes alternative solutions.

10. AutoPPG

AutoPPG supports the semi-automatic generation of PP for Android applications. Based in part on the source code of a given app, AutoPPG produces a set of human-readable descriptions that can be subsequently modified and turned into a PP. This approach addresses a very interesting and challenging problem: correctly understanding the source code of a given app, and subsequently writing a policy that is compliant with such code. This is certainly not a trivial task, since it requires a deep knowledge of both computer science and the legal domain.

To summarize, many applications are under development, and more research projects and software platforms are likely to become available in the next few years. The most frequently shared goals of these applications are the summarization of documents or the presentation of information in a more user-friendly way. Clearly, there is a huge difference between research projects and

tools/products that actually enter the market. The road has been paved, but there is still a long way to go.

III. FROM THE LAB TO THE (LEGAL) BATTLEFIELD

What all the above-mentioned research projects and (prototypes of) tools have in common is an attempt to address the "too much/too confusing to read" problem through the automation of PP and ToS analysis. As discussed, various tasks can be delegated to machines, from the summarization of PPs and ToS, to the evaluation of whether they meet certain legal standards. In this section we survey possible real-world applications of the technologies discussed in the previous subsection. We look at the technologies from the perspective of different classes of actors: users/consumers, NGOs and enforcers, academics and policymakers, and businesses. We then survey the technological and market conditions that need to be met in order to turn the research projects and prototypes into tools actually used by these different actors on a daily basis.

A. Possible Applications: User Stories

Let us start with everyday users of app and websites like you and me. As already discussed, big data analytics could help such users read and understand ToS and PPs. One could imagine, for example, applications that help users learn which clauses are critical to notice. For example, many ToS include arbitration agreements from which users can opt-out within some specified period after the purchase/acceptance. It is possible to automate the notification to the user and to automate his or her option to opt-out. The same can happen with any other type of right "hidden" in the terms (like a right to withdraw, or to object to profiling). A user could be informed about a clause in a contract he or she just accepted ("the plane ticket you just purchased can be cancelled at no cost to you during the next 24 hours") and given an option to act upon it ("the app you just downloaded has an arbitration clause, tap here if you want to opt-out"). It is also possible to automate certain actions. For example, a user could conceivably activate a setting that automatically opts him or her out of the arbitration clause in any ToS or PP. In jurisdictions (like the EU) where the usage of unfair terms is prohibited, users could be given an option to automatically notify consumer organizations or supervisory authorities. For example, the user could receive a pop-up warning saying that "these (potentially) unfair clauses have been detected;

would you like to send an email to an NGO X/supervisory authority Y?" This action, just like the exercise of rights, could easily be automated (e.g., emails auto-written and sent).

This would open lines of information and communication between users and regulatory watch-dogs such as NGOs. Just like the users, these bodies often enjoy different sets of rights/competences, but lack the capabilities and resources to make use of them. The scenario discussed above could help them aggregate the knowledge about types of clauses used in ToS and PPs. However, one can also imagine organizations using such systems without users' direct involvement. For example, an organization investigating the usage of liability limitation clauses could use a crawler-software to automatically retrieve terms of service of thousands of platforms, and then machine learning techniques to find and annotate these clauses. In this scenario, a human-lawyer, instead of going page-by-page through documents, could receive a table of extracted paragraphs, or pre-annotated documents. The same can happen with obligatory rights. The GDPR, for example, requires that privacy policies inform users about their rights, in clear and intelligible language. If a pre-trained machine fails to detect the required clauses, there is a reason to assume that they are missing, or that they are not communicated clearly. Other steps involved here—like drafting letters to companies, or the creation of legal documents—could be automated as well. Here, again, the exact application will depend on the legal system. In jurisdictions where abstract control exists, this process could be automated, increasing the incentive for companies to comply in the first place. In jurisdictions where NGOs must rely on other means (such as market pressure), the automatic aggregation of knowledge can increase the efficiency and quality of this process.

The discussed techniques can also be employed by academics. Anyone who wishes to study what companies write in their ToS and PPs can rely on big data analytics. What the machine would look for will differ according to the research questions under consideration; but the possible applications are immense. Similarly, policymakers attempting to respond to the actual market practice can rely on such applications in order to better comprehend the current trends.

Finally, such applications could be employed by businesses. Companies, especially startups and small enterprises, wishing to comply with the regulations and/or societally developed standards could use a software to check where their ToS and PPs could be improved. One can imagine that NGOs trying to increase quality of these documents develop tools available free of charge, allowing entities that otherwise could not afford legal services, to be compliant with the law or societal expectations. Otherwise, such systems could be developed and made available by

legal-tech developers, still offering an (automated) legal advice for much smaller amounts of money. While there are many possible benefits for society, companies could try to "game" these applications by modifying their terms to appear 'fair' while actually trying to make the terms as unfriendly as possible to consumers. Further, one could imagine lawyers using them to go after small enterprises and offering to fix the documents for high legal fees, threatening to otherwise denounce them to supervisory authorities. In such a scenario, it would be the vulnerable that get hit strongest, not the big corporations. One should be aware of these risks, and the law will likely need to address them; however, these risks do not appear to outweigh the potential benefits of further research in this area.

B. Preconditions

In the foregoing we have discussed various research projects and the technologies they are developing, as well as the possible ways in which the technologies can be used to empower the individuals and the civil society, and restore the balance of power between big business and consumers. However, there is still a long way to go until such tools are being used on a scale that actually makes a difference. In this section, we survey the technological and financial preconditions that need to be met before this can happen.

1. Creation of Data Sets

To delegate tasks to computers using machine learning, one first needs to feed the computers with data. In the case of the analysis of ToS and PPs, this data will first need to be annotated by humans (i.e., human taggers will need to create a data set). As explained above, this is a costly and time-intensive process. As of today, it is the most significant hurdle to cross before the widespread automation of textual analysis can become a reality.

What does this mean in practice? Imagine one wants to teach a machine to spot arbitration clauses that include the possibility of opting-out. To do so, a human would read many ToS, and mark every sentence containing such a clause. For example, the Terms of Dropbox contain such a clause:

> You and Dropbox agree to resolve any claims relating to these Terms or the Services through final and binding arbitration by a single arbitrator (...) *You can*

> *decline this agreement to arbitrate by clicking here and submitting the opt-out form within 30 days of first registering your account.* (...) The American Arbitration Association (AAA) will administer the arbitration under its Commercial Arbitration Rules and the Supplementary Procedures for Consumer Related Disputes.[57]

When a machine is presented with a series of ToS with a particular type of clause marked in all of them, it will learn to recognize these clauses by studying those features (lexical, syntactic etc.) which are present in these sentences and absent in all the other sentences in that ToS. Since such clauses can be phrased differently, the more examples the machine has, the better. How would such a set be created?

First, a tagging instruction needs to be created. Humans need to specify what they want to teach the machine to look for. This will be based both on the rules and standards, and on the real world examples encountered in the ToS and PPs in use. Second, the documents will need to be tagged. Here, again, human action is necessary. Researchers will have to read dozens, if not hundreds or thousands of documents, and mark them according to the instruction. Usually, at least two people will first mark the same documents, and a comparison will be made (automatically), to detect mistakes and or create further disambiguations in the instruction. The latter is necessary especially if there is a divergence in interpretation of the instructions among human taggers.

Third, once the set is ready, the machine must be trained. Various learning algorithms (such as neural networks, support vector machines, and logistic regression) can be employed, and if necessary combined. Researchers will measure the precision (amount of false positives) and recall (amount of false negatives) to assess whether the performance is satisfactory. One should bear in mind that the machine will never be 100% flawless (but neither are humans). Developers will need to agree on the level, and type, of mistakes they are ready to accept. For example, in some settings, like abstract control (that is NGOs screening ToS and PPs on their own motion, without involvement of a particular consumer) one might prefer more noise to silence (higher recall while sacrificing some precision). In others, like automating emails to businesses, one might prefer higher precision (all clauses that get marked are unfair) to recall (some clauses that are unfair do not get marked).

Once this process is over, the trained algorithm is ready to operate in the lab environment. However, there is still some way it needs to go before it becomes a downloadable app or browser extension.

2. Software Engineering

Performing advanced analysis of large collections of ToS and PPs, and implementing such technology into useful tools for end-users, requires not only application of artificial intelligence and machine learning but also crucial contributions from software engineering. In order to make applications widespread, it will be necessary to empower end-users with accessible and usable software that can be installed without much effort on different devices and platforms: smartphone apps, browser extensions, and the like. User-friendliness becomes a crucial requirement and can make all the difference for the success of an application. Academics and tech companies will likely need to work together to develop applications that are both effective and easy to use.

3. Challenges for Artificial Intelligence

The tasks discussed above use state of the art techniques from machine learning, natural language processing, and artificial intelligence. However, in order to address novel problems and further improve the performance of existing approaches, there will be need to be further advances in AI.

Deep learning has recently brought a revolution to the field of AI, producing stunning results across many different fields that were unthinkable only a few years ago.[58] Nevertheless, AI must continue to develop in order to achieve human-level performance in many domains. Natural language processing is among these domains, as AI still struggles to infer novel knowledge from a given text or perform reasoning operations.

In general, deep learning models (that is, deep artificial neural networks) are often criticized as of being "black-box" models, whose answers, despite being remarkably accurate, are hard to interpret. There is a major need, in the field of AI, to build *explainable* models, i.e., models capable of motivating their choices, that is models whose decision processes can be interpreted by a human. The direction in which the field is moving is that of integrating so-called sub-symbolic (or connectionst) approaches, such as artificial neural networks, with so-called symbolic methodologies, which are built on logic.[59] The former are capable of efficiently and effectively

dealing with uncertainty in data and can easily exploit very large data collections, but lack in interpretability. The latter, on the other hand, are designed to deal with knowledge representation and reasoning, and thus show a high expressivity, a high interpretability, but cannot easily handle noisy information and scale to big data. There is a strong belief within the AI community that the combination of such diverse approaches is a necessary step to fill the performance gap in tasks related to reasoning. Several lines of research have been developed in this direction, such as Statistical Relational Learning[60] and Neural-Symbolic Learning.[61]

Finally, the use of unsupervised data is another major issue for AI. Supervised data is extremely costly and thus difficult to obtain, whereas unsupervised data collections are everywhere, and they are often available for free. As explained above, for the analysis of PPs, a few projects are trying to use unsupervised learning approaches that are capable of capturing specific language characteristics. However, there is still a lot to be done before the use of unsupervised data becomes effective.

4. Market Conditions

All this requires a level of funding which is often only at the disposal of the state and big business. Therefore, funding is necessary to facilitate the interdisciplinary cooperation between computer scientists and lawyers, as well as activists and practitioners. From where could this funding come?

One option is the market itself. If users were willing to pay for these products, one could expect numerous companies to emerge. This might happen for some applications. However, for some types of applications, especially those that empower the NGOs there are not many reasons to be optimistic. These organizations are underfunded in the first place, and the money they spend on tech is money they do not spend on wages for activists. Hence, market forces alone are not the answer.

There is, of course, the possibility of direct intervention by the state. For example, the government could decide to channel public money into civil tech research and development, requiring resulting technology to be open source and/or available to all those who need it free of charge. This approach also has its drawbacks, both political and administrative ones. Yet another option would be to indirectly provide private funding through changes in the law. In jurisdictions where a user can sue a company for using unfair terms, especially in jurisdictions that permit class-

action suits, people can pay for the development of these systems through the fractions of the compensation they will receive

Ultimately, as it usually is the case with civic tech, the funding would come through a complex system of the market, private philanthropy and public spending. Its exact shape will depend on the societal decisions and conditions of different jurisdictions. What one has to bear in mind is that there is much untapped potential in the civic tech field, and resources should be channeled to where they can be best used.

CONCLUSIONS AND THE WAY FORWARD

The comprehension of ToS and PPs—"big data" from the perspective of the users—can be made more effective and efficient through machine learning and other big data analytics techniques. This is the message of this chapter. We have analyzed various ways in which this can be achieved, as well as provided an overview of the current state of the law and computer science.

As we become more and more aware that the "notice and choice" model is ineffective, and as the difference in power between big business and consumers increases, the law might need to change as well. Specifically, the EU approach of prohibiting certain classes of "unfair clauses" is something that could be adopted by other jurisdictions, such as the U.S. This could be paired with a system of abstract control, increasing the role of the FTC. Cooperation with NGOs and civil society at large would also be beneficial.

However, the role for law and policy is not just to constrain the power of giant corporations, but also to enable bottom-up civil society initiatives. We do not necessarily need more regulation. We could achieve the same goals if people become more empowered to make choices regarding their personal privacy. Investing in the development of civic tech is one such possibility. We hope that, whether for the purposes of pushing the scholarly understanding forward, or with the aim of empowering the civil society through novel applications, the argument and resources analyzed in this chapter will serve as a resource for researchers and developers alike.

References

Art. 3 of Council Directive 93/13/EEC on unfair terms in consumer contracts (UCTD) [1993] OJ L95/29.

Austin, Lisa et al. (2018), *Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project* (June 27, 2018), https://ssrn.com/abstract=3203601.

Bakos, Yannis et al. (2014), *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUD. 1, 1-36.

Balkin, Jack (2018), *Fixing Social Media's Grand Bargain*, Yale Law School, Public Law Research Paper No. 652, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1814, at 4 (Oct. 16, 2018).

Bar-Gill, Oren et al. (2017), *Searching for the Common Law: The Quantitative Approach of the Restatement of Consumer Contracts*', 84 U. CHI. L. REV. 7.

CONTISSA, GIUSEPPE ET AL. (2018), *CLAUDETTE MEETS GDPR: AUTOMATING THE EVALUATION OF PRIVACY POLICIES USING ARTIFICIAL INTELLIGENCE*, available at https://www.beuc.eu/publications/beuc-x-2018-066_claudette_meets_gdpr_report.pdf.

Costante, Elisa et al. (2012), *A Machine Learning Solution to Assess Privacy Policy Completeness, in* ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY.

D'Errico, Michela & Siani Pearson (2015), *Towards a Formalised Representation for the Technical Enforcement of Privacy Level Agreements*, *in* IEEE INTERNATIONAL CONFERENCE ON CLOUD ENGINEERING (IC2E).

De Mauro, Andrea et al. (2016), *A Formal Definition of Big Data Based on its Essential Features*, 65 LIBRARY REV. 122.

DINSMORE, JOHN (2014), THE SYMBOLIC AND CONNECTIONIST PARADIGMS: CLOSING THE GAP.

Dropbox Terms of Service, Posted: April 17, 2018, Effective: May 25, 2018, DROPBOX, https://www.dropbox.com/terms (last visited May 9, 2020).

Gandomi, Amir & Haider Murtaza (2015), *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L. J. INFO. MGMT. 137.

GARCEZ, ARTUR S. D'AVILA ET AL. (2015), NEURAL-SYMBOLIC LEARNING SYSTEMS: FOUNDATIONS AND APPLICATIONS.

HANS SCHULTE-NÖLKE, HANS ET AL. (2008), EC CONSUMER LAW COMPENDIUM: THE CONSUMER *ACQUIS* AND ITS TRANSPOSITION IN THE MEMBER STATES.

Hans, G.S. (2012), *Privacy Policies, Terms of Service, and FTC Enforcement: Broadening Unfairness Regulation for a New Era* 19 MICH. TELE. & TECH. L. REV. 163.

Harkous, Hamza et al. (2018), *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning, in* USENIX SECURITY.

Hildebrandt, Mireille (2015), *Dualism is Dead. Long Live Plurality (Instead of Duality)*, *in* THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA (Luciano Floridi ed., 2015).

HILDEBRANDT, MIREILLE (2015), SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY.

Hoffman, David A. (2018), *Relational Contracts of Adhesion*, 85 U. CHI. L. REV. 1395.

INTRODUCTION TO STATISTICAL RELATIONAL LEARNING (Lise Getoor & Ben Taskar eds., 2007).

Jablonowska, Agnieszka et al. (2018), *Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence - Final report of the ARTSY project*, EUI Department of Law Research Paper No. 2018/11 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3228051.

Joshi, Karuna P. et al. (2016), *Semantic Approach to Automating Management of Big Data Privacy Policies*, *in* IEEE Big Data.

Lebanoff, Logan & Fei Liu, *Automatic Detection of Vague Words and Sentences in Privacy Policies*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018).

LeCun, Yann et al. (2015), *Deep Learning*, 521 NATURE 436.

Lippi, Marco et al. (2019), *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 ART. INTELL. L. 117.

Lippi, Marco et al. (2019), *Consumer Protection Requires Artificial Intelligence,* 1 NAT. MACH. INTELL. 168.

Liu, Frederick et al. (2018), *Towards Automatic Classification of Privacy Policy*, *in* CMU-ISR-17-118R CMU-LTI-17-010.

Loos, Marco & Joasia Luzak (2016)*, Wanted: A Bigger Stick. On Unfair Terms in Consumer Contracts with Online Service Providers*, 39 J. CONS. POL. 63.

Mak*,* Chantal (2008), *Fundamental Rights and the European Regulation of iConsumer Contracts,* 31 J. CONS. POL. 425.

MARKOVITS, DANIEL (2012), CONTRACT LAW AND LEGAL METHODS.

McDonald, Leecia M. & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 ISJLP 543 (2008).

Micklitz, Hans-W. & Betül Kas (2014), *Overview of Cases Before the CJEU on European Consumer Contract Law (2008–2013) – Part I*, 10 EUR. REV. OF CONT. L. 1.

Micklitz, Hans-W. & Przemyslaw Palka (2019), *Algorithms in the Service of the Civil Society*, 8 J. OF EUR. CONSUMER & MKT. L. 1.

Micklitz, Hans-W. et al. (2017)*, The Empire Strikes Back: Digital Control of Unfair Terms of Online Services*, 40 J. CONS. POL. 367.

Mik, Eliza (2016), *The Erosion of Autonomy in Online Consumer Transactions*, 8 L. INNOVATION & TECH. 1.

Obar, Jonathan A. & Anne Oeldorf-Hirsch (2018), *The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services*, 21 INFO., COMM. & SOC'Y 1, 1-20.

Oltramari, Alessandro et al. (2017), *PrivOnto: A Semantic Framework for the Analysis of Privacy Policies*, SEMANTIC WEB J., http://www.semantic-web-journal.net/system/files/swj1597.pdf.

Palka, Przemyslaw, *Data Management Law for the 2020s: The Lost Origins and the New Needs,* 68 BUFF. L. REV. (forthcoming 2020).

Palka, Przemyslaw (2018), *Terms of Service Are Not Contracts: Beyond Contract Law in the Regulation of Online Platforms*, *in* EUROPEAN CONTRACT LAW IN THE DIGITAL AGE (Stefan Grundmann ed., 2018).

Palmirani, Monica et al. (2018), *PrOnto: Privacy Ontology for Legal Reasoning*, EGOVIS 139-152.

Paul, Niklas et al. (2018), *Assessing Privacy Policies of Internet of Things Services*, *in* ICT SYSTEMS SECURITY AND PRIVACY PROTECTION.

Ramanath, Rohan et al. (2014), *Unsupervised Alignment of Privacy Policies using Hidden Markov Models*, *in* PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

Reidenberg, Joel R. et al. (2015), *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding*, 30 BERKELEY TECH. L.J. 39.

Sathyendra, Kanthashree M. et al. (2017), *Identifying the Provision of Choices in Privacy Policy*, *in* PROCEEDINGS OF THE 2017 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP).

SOLOVE, DANIEL J. & PAUL M. SCHWARTZ (2015), INFORMATION PRIVACY LAW.

Solove, Daniel J. & Woodrow Hartzog (2014), *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583.

Solove, Daniel J. (2013), *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880.

The California Online Privacy Protection Act [2003] California Business and Professions Code par. 22575-22579.

Tippett, Elizabeth C. & Bridget Schaaff (2018), *How Conception and Italian Colors Affected Terms of Service Contracts in the Gig Economy*, 70 RUTGERS U.L. REV. 459.

Tomuro, Noriko et al. (2016), *Automatic Summarization of Privacy Policies using Ensemble Learning*, *in* CONFERENCE ON DATA AND APPLICATION SECURITY AND PRIVACY (CODASPY).

Waldman, Ari Ezra (2018), *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74.

Wilson, Shomir et al. (2018), *Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations*, *in* ACM TRANSACTIONS ON THE WEB.

Wilson, Shomir et al. (2016), *Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?*, *in* WORLD WIDE WEB CONFERENCE.

Zaeem, Razieh Nokhbeh et al. (2010), *PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining*, *in* ACM TRANSACTIONS OF INTERNET TECHNOLOGY.

[*] The authors would like to thank Giuseppe Contissa, Francesca Lagioia, Hans Micklitz, Giovanni Sartor and Paolo Torroni for their long-term collaboration, as a result of which many of the ideas presented in this Chapter were born. They would also like to thank Roger Ford, Margot Kaminski and Florencia Marotta-Wurgler for valuable comments and suggestions made at the early stage of the project, as well as Daniel Markovits and Roland Vogl for providing comments on the first draft.

[**] Assistant Professor, Future Law Lab at Jagiellonian University in Krakow, Poland; Associate Research Scholar, Yale Law School; Visiting Fellow, Information Society Project at Yale.

[***] Associate Professor of Computer Engineering, Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia.

[1] These documents come under various titles: "terms of service," "terms of use," "terms and conditions," etc. For the sake of consistency and brevity, we refer to them as "terms of service" or "ToS" throughout the chapter. "Privacy Policies" are abbreviated as "PPs."

[2] *See* Mireille Hildebrandt, *Dualism is Dead. Long Live Plurality (Instead of Duality)*, *in* THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA (Luciano Floridi ed., 2015); and MIREILLE HILDEBRANDT, SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY (2015).

[3] *See* Agnieszka Jablonowska et al., *Consumer Law and Artificial Intelligence: Challenges to the EU Consumer Law and Policy Stemming from the Business' Use of Artificial Intelligence - Final report of the ARTSY project*, EUI Department of Law Research Paper No. 2018/11 (2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3228051.

[4] *See* Eliza Mik, *The Erosion of Autonomy in Online Consumer Transactions*, 8 L. INNOVATION & TECH. 1 (2016).

[5] *See* Jack Balkin, *Fixing Social Media's Grand Bargain*, Yale Law School, Public Law Research Paper No. 652, Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper No. 1814, at 4 (Oct. 16, 2018).

[6] Within this chapter we do not engage with the debate on whether consent is the right legal tool to govern online data management. For skeptical arguments, see Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013); Przemyslaw Palka, *Data Management Law for the 2020s: The Lost Origins and the New Needs,* 68 BUFF. L. REV. (forthcoming 2020).

[7] *See* Yannis Bakos et al., *Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts*, 43 J. LEGAL STUD. 1, 1-36 (2014); Jonathan A. Obar & Anne Oeldorf-Hirsch, *The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services*, 21 INFO., COMM. & SOC'Y 1, 1-20 (2018).

[8] Joel R. Reidenberg et al., *Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding*, 30 BERKELEY TECH. L.J. 39 (2015).

[9] *See* Ari Ezra Waldman, *Privacy, Notice, and Design*, 21 STAN. TECH. L. REV. 74 (2018).

[10] *See* Solove, *supra* note 6.

[11] Leecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 ISJLP 543 (2008).

[12] *See* Marco Lippi et al., *Consumer Protection Requires Artificial Intelligence,* 1 NAT. MACH. INTELL. 168 (2019); Hans-W. Micklitz & Przemyslaw Palka, *Algorithms in the Service of the Civil Society*, 8 J. OF EUR. CONSUMER & MKT. L. 1 (2019).
[13] Andrea De Mauro et al., *A Formal Definition of Big Data Based on its Essential Features*, 65 LIBRARY REV. 122 (2016).
[14] Amir Gandomi & Haider Murtaza, *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L. J. INFO. MGMT. 137 (2015).
[15] *See* Marco Loos & Joasia Luzak*, Wanted: A Bigger Stick. On Unfair Terms in Consumer Contracts with Online Service Providers*, 39 J. CONS. POL. 63 (2016); David A. Hoffman, *Relational Contracts of Adhesion*, 85 U. CHI. L. REV. 1395 (2018); Przemyslaw Palka, *Terms of Service Are Not Contracts: Beyond Contract Law in the Regulation of Online Platforms*, *in* EUROPEAN CONTRACT LAW IN THE DIGITAL AGE (Stefan Grundmann ed., 2018).
[16] *See* Hans-W. Micklitz et al.*, The Empire Strikes Back: Digital Control of Unfair Terms of Online Services*, 40 J. CONS. POL. 367 (2017).
[17] *See* art. 3 of Council Directive 93/13/EEC on unfair terms in consumer contracts (UCTD) [1993] OJ L95/29.
[18] *See* Hans-W. Micklitz & Betül Kas, *Overview of Cases Before the CJEU on European Consumer Contract Law (2008–2013) – Part I*, 10 EUR. REV. OF CONT. L. 1 (2014).
[19] *See* Loos & Luzak, *supra* note 15; Marco Lippi et al., *CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service*, 27 ART. INTELL. L. 117 (2019).
[20] *See* HANS SCHULTE-NÖLKE ET AL., EC CONSUMER LAW COMPENDIUM: THE CONSUMER A*CQUIS* AND ITS TRANSPOSITION IN THE MEMBER STATES (2008).
[21] *See* Chantal Mak*, Fundamental Rights and the European Regulation of iConsumer Contracts,* 31 J. CONS. POL. 425 (2008).
[22] *See* DANIEL MARKOVITS, CONTRACT LAW AND LEGAL METHODS (2012).
[23] *See* Elizabeth C. Tippett & Bridget Schaaff, *How Conception and Italian Colors Affected Terms of Service Contracts in the Gig Economy*, 70 RUTGERS U.L. REV. 459 (2018).
[24] Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1. Hereinafter, the "GDPR."
[25] The European Union is not a federation, but a supranational organization. But its Regulations, as opposed to Directives (which must be transposed by the Member State legislation), can be understood as roughly equivalent to US federal law.
[26] GDPR, art. 5.
[27] GDPR, art. 83.
[28] The California Online Privacy Protection Act [2003] California Business and Professions Code par. 22575-22579.
[29] *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).
[30] For a longer discussion of the origins and character of these differences, see Palka, *supra* note 6.
[31] *See* DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW (2015).
[32] Oren Bar-Gill et al., *Searching for the Common Law: The Quantitative Approach of the Restatement of Consumer Contracts*', 84 U. CHI. L. REV. 7 (2017).
[33] *See* G.S. Hans, *Privacy Policies, Terms of Service, and FTC Enforcement: Broadening Unfairness Regulation for a New Era* 19 MICH. TELE. & TECH. L. REV. 163 (2012); Solove & Hartzog *supra* note 29.
[34] *See infra* Part III.
[35] *See* Lippi et al., *supra* note 19.
[36] *See* Elisa Costante et al., *A Machine Learning Solution to Assess Privacy Policy Completeness, in* ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY (2012).
[37] *See* Razieh Nokhbeh Zaeem et al., *PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining*, *in* ACM TRANSACTIONS OF INTERNET TECHNOLOGY (2010); Noriko Tomuro et al., *Automatic Summarization of Privacy Policies using Ensemble Learning*, *in* CONFERENCE ON DATA AND APPLICATION SECURITY AND PRIVACY (CODASPY) (2016).
[38] *See* Frederick Liu et al., *Towards Automatic Classification of Privacy Policy*, CMU-ISR-17-118R CMU-LTI-17-010 (2018).
[39] *See* Kanthashree M. Sathyendra et al., *Identifying the Provision of Choices in Privacy Policy*, *in* PROCEEDINGS OF THE 2017 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP) (2017).
[40] *See* Logan Lebanoff & Fei Liu, *Automatic Detection of Vague Words and Sentences in Privacy Policies*, *in* PROCEEDINGS OF THE 2018 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)

(2018); GIUSEPPE CONTISSA ET AL., *CLAUDETTE MEETS GDPR: AUTOMATING THE EVALUATION OF PRIVACY POLICIES USING ARTIFICIAL INTELLIGENCE* (2018), available at https://www.beuc.eu/publications/beuc-x-2018-066_claudette_meets_gdpr_report.pdf.

[41] *See* Costante et al. *supra* note 36.

[42] *See* Niklas Paul et al., *Assessing Privacy Policies of Internet of Things Services*, *in* ICT SYSTEMS SECURITY AND PRIVACY PROTECTION (2018).

[43] *See* Karuna P. Joshi et al., *Semantic Approach to Automating Management of Big Data Privacy Policies*, *in* IEEE Big Data (2016).

[44] *See* Hamza Harkous et al., *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning, in* USENIX SECURITY (2018).

[45] *See* Alessandro Oltramari et al., *PrivOnto: A Semantic Framework for the Analysis of Privacy Policies*, SEMANTIC WEB JOURNAL (2017), http://www.semantic-web-journal.net/system/files/swj1597.pdf; Monica Palmirani et al., *PrOnto: Privacy Ontology for Legal Reasoning*, EGOVIS 139-152 (2018).

[46] Michela D'Errico & Siani Pearson, *Towards a Formalised Representation for the Technical Enforcement of Privacy Level Agreements*, *in* IEEE INTERNATIONAL CONFERENCE ON CLOUD ENGINEERING (IC2E) (2015).

[47] *See* Rohan Ramanath et al., *Unsupervised Alignment of Privacy Policies using Hidden Markov Models*, *in* PROCEEDINGS OF THE 52ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2014).

[48] *See* Harkous et al., *supra* note 44.

[49] *See* Shomir Wilson et al., *Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?*, *in* WORLD WIDE WEB CONFERENCE (2016); Shomir Wilson et al., *Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations*, *in* ACM TRANSACTIONS ON THE WEB (2018).

[50] *See infra* Part III.

[51] *See* Harkous et al., *supra* note 44.

[52] *See* Lippi et al. *supra* note 12; Contissa et al., *supra* note 40.

[53] *Id.*

[54] *See* Oltramiari et al., *supra* note 45.

[55] *See* Palmirani et al., *supra* note 45.

[56] *See* Lisa Austin et al., *Towards Dynamic Transparency: The AppTrans (Transparency for Android Applications) Project* (June 27, 2018), https://ssrn.com/abstract=3203601.

[57] *Dropbox Terms of Service, Posted: April 17, 2018, Effective: May 25, 2018*, DROPBOX, https://www.dropbox.com/terms (last visited May 9, 2020).

[58] *See* Yann LeCun et al., *Deep Learning*, 521 NATURE 436 (2015).

[59] *See* JOHN DINSMORE, THE SYMBOLIC AND CONNECTIONIST PARADIGMS: CLOSING THE GAP (2014).

[60] INTRODUCTION TO STATISTICAL RELATIONAL LEARNING (Lise Getoor & Ben Taskar eds., 2007).

[61] ARTUR S. D'AVILA GARCEZ ET AL., NEURAL-SYMBOLIC LEARNING SYSTEMS: FOUNDATIONS AND APPLICATIONS (2015).