# 2. Context Survey

## 2.1. Legal Background

### 2.1.1. Privacy and the GDPR

### Sources: [1] [17]

* Before first data protection laws in the 1980s, data was simply protected through the way it was stored.
* As technology developed, data became a commodity.
* Data is the new currency of the 21st century, in the sense that an increasing number of businesses are collecting and storing personal information with the aim to make a monetary benefit from this activity [1].
* Essential to have a legal framework which protects people’s privacy and control over their own personal data [1].
* Article 8 EU Charter specifically provides “protection of personal data” [17], and states that data should be processed fairly, for specified purposes, and on the basis of consent or another legitimate basis laid down by the law.
* In the EU, the law clarifying data protection rights since 2018 is the General Data Protection Regulation (GDPR). The aim of this regulation was to give back control to people over their private data [1].
* But the GDPR did not reinvent everything. There was already a legal framework in place in the EU before 2018, called the Data Protection Directive.
* However, because it was a directive and not a regulation, this caused disparate legal regimes across member states, with some countries applying a light touch version of the Directive, and others adopting a much stricter data protection regime.
* Things that have changed with the introduction of the GDPR: data access and portability rights introduced, new territorial scope and increased breach penalties (and more).
* The provision which establishes a legal basis for data processing on different grounds, including consent and legitimate interests, has not changed massively in the GDPR compared to the Directive.

### 2.1.2. Consent

Sources: [1] [2] [20] [21] [7] [8]

* Consent meant to be the default legal basis for processing.
* Once acquired, consent is simple to demonstrate and it allows data controllers to specify all the data use cases to rule out the possibility of unwanted legal issues [1].
* Consent was already given primary status as legal basis by most member states as legal basis for processing. This is because privacy consent has been part of the constitutional doctrine from the development of privacy [2].
* However, in other states, including the UK, where privacy has never been a constitutional right, and thus the value of consent was perceived differently, consent as a ground for processing was only relied upon as a last resort [2].
* A key goal of the GDPR was to provide uniformity and to fix these problems of different implementation in the member states [2].
* But the effectiveness of consent in practice is limited. Good ground for controllers to demonstrate compliance, but raises the question of informed consent, and whether data subjects really know what they are consenting to, or in some cases that they are consenting to something.
* A good example, which will be the focus of this paper, is the reliance on privacy policies as a way to inform data subjects about the use of their personal data, and what they are consenting to.
* According to a 2008 study, it would take the average American 244 hours a year to read all of the privacy policies that they encounter. The cost of this is estimated at US$53.8bn per year [20].
* Chief Justice Roberts - “Providing too much information defeats the purpose of disclosure, since no one reads it” [21].
* To be effectively enforced, we need to make sure that the obtained consent is meaningful.
* However, technology might be used to obtain consent more easily.
* Studies have shown that organisations use UX/UI strategies and dark patterns to nudge users into accepting more privacy-invasive settings.
  + “Users are more likely to interact with a notice shown in the lower left part of the screen” [7].
  + “Given a binary choice, more users are willing to accept tracking compared to mechanisms that require them to allow cookie use for each category or company individually” [7].
  + “Removing the opt-out button from the first page increases consent by 22 percentage points” [8].
  + “Providing more granular controls on the first page decreases consent by 8-20 percentage points” [8].
* All of these limitations suggest that giving people the power to decide by asking for their informed consent might not be the most efficient legal basis for processing.

### 2.1.3. Legitimate Interests

Sources: [14] [12] [13] [2] [10] [11]

* Legitimate interests as a legal ground for processing is only meant to be used in a very limited number of cases where consent is not appropriate (e.g. fraud prevention - it wouldn’t make sense to ask consent to data subjects to detect if they are fraudsters).
* Legitimate interests is one the least discussed grounds for lawful processing, despite being equally binding as other grounds for processing [4]. There is a limited amount of academic papers or official guidance covering the topic, except the Article 29 of the Working Party [12].
* The criteria states that data controllers can process personal data lawfully, without meeting other tight conditions of the law, if the “processing is necessary for the purposes of [their] legitimate interests, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject” [13].
* This provision is formulated broadly enough to address situations where conflicts arise between the private interests of data controllers and the legal rights of data subjects [2].
* This broadness provides flexibility for the development of commercial activities and new technologies that make use of personal identifiable information. However, this business-oriented approach removes a degree of legal certainty and gives rise to potential loopholes in the legal system [2].
* The wording in the provision suggests that a balancing test is necessary to weigh the competing interests and rights of the data controllers and data subjects, leaving the legitimacy of processing to a case-by-case determination [2]. The responsibility to demonstrate the existence of a legitimate interest that justifies processing falls on the data controller, based on the accountability principle. The data controller must also prove that the interest does not override any fundamental rights or freedoms of the data subject.
* There is a lack of a formalised process to determine whether legitimate interest is an appropriate ground. The ICO suggests breaking the Legitimate Interests Assessment (LIA) into a three-part test, formed of a purpose, necessity and balancing test [10]. Kamara et al. suggest a similar balancing act framework, starting by considering the legitimacy of the pursued interest, then measuring the necessity of the processing to meet the purposes of the interest, and finally balancing the opposing interests of the controller and the data subject [4].
* There is still a lack of a formalised process in the law, and different member states have adopted different guidelines to the LIA (see the CNIL’s “Interet Legitime” [11]). Legitimate interests is not a loophole in the GDPR, it is an equally important ground for legitimate interest. However, the lack of guidance on how to assess it can be seen as too lenient, and opening the door to misinterpretation or bad faith in its application [4].

## 2.2. Computer Science Background

### 2.2.1. Standardisation of Privacy Policies

Sources: [14] [23] [24] [15] [25] [17]

* Article 13 and 14 GDPR require all companies operating in the EU, as well as foreign companies that handle personal data of people located in the EU, to have a privacy policy [14] [23]. This is part of the regulations’ goal to ensure that personal information is obtained and processed fairly.
* The ICO Guidance also indicates that a controller must “actively provide privacy information to individuals”. This requirement can be met by putting the information on the controller’s website, as long as the controller makes “individuals aware of it and give them an easy wait to access it” [24].
* The ICO also states that the information provided in the privacy policy must be “concise, transparent and intelligible” [24]. However, the quality and effectiveness of privacy policies are hard to evaluate, which makes it difficult to enforce the law.
* Some attempts have been made at standardising privacy policies. One example of this is P3P (Platform for Privacy Preferences), which is a standard which allows user agents to automatically evaluate privacy policies against users’ preferences by specifying those policies in a machine readable format [15]. P3P provides two mechanisms for each website’s privacy policy on cookies - a full policy and a compact one. The fully policy XML files that represent a website’s privacy policy in detail, while the compact policy summarises the policy using a string of three-character or four-character tokens. Internet Explorer then makes cookie-filtering decisions by comparing a website’s compact policy with the user’s configured privacy preferences.
* In 2002, regulators from several countries agreed that a P3P policy was legal binding and constituted a “representation to consumers on which they can be expected to rely” [25].
* However, a study by Leon et al. from 2010 [15] found that nearly 34% of the total compact policies evaluated had at least one of the following errors: invalid, missing tokens, or conflicting tokens. They also showed that 97% of these errors meant that the compact policy would bypass Internet Explorer’s default privacy filters, therefore clearly misleading users.
* The Nutrition Label approach is another attempt at standardising privacy policies, by using an iterative user-centered design process to develop a more compelling and informative privacy policy format [17]. The aim of the standardisation is not to evaluate the policies like P3P, but improve the impact that privacy policies may have on users.
* Kelley et al. conducted a large online user study to evaluate different variants of privacy policy format, including a standardised-table format, a tabular variant and a full-text natural language policy. They found that policy formats do have a significant impact on users’ ability to quickly and accurately find information, and users’ attitudes regarding the experience of using privacy policies.
* However, they identified room for study and improvement in policy-comparison tasks, which they didn’t manage to achieve with their standardisation techniques.

### 2.2.2. Technical Analysis of Cookie Consent Notices

Sources: [7] [26] [27] [8]

* Technical analysis of privacy online has so far mainly focused on the study of consent, and cookie consent in particular.
* Utz et al. made use of a scraper in their 2019 research which aimed at understanding the impact of the graphical interface (UI) of cookie consent notices on users’ decisions [7]. They investigated the effect of several parameters on users’ interactions with a consent notice, including its position, the number of choices available, the presence of a privacy policy link and whether the notice referred to “cookie use” (technical language) or “data collection” (non-technical). They also studied whether they could nudge websites’ visitors towards giving consent through highlighting and preselection.
* In order to customise the UI and collect useful data for each of the experiments, the researchers partnered with an e-commerce website based on WordPress, which had over 15,000 unique visitors per month. They modified a WordPress plugin, which has now been discontinued, called Ginger - EU Cookie Law [26] to test those different notice variants.
* The findings of the study were significant as they demonstrated that not only the users are more likely to interact with notices depending on their positions on the screen, or to accept tracking if given a binary choice, it also showed that the widespread practice of nudging has a large effect on users’ choices.
* The paper recommended that regulators should not only require consent, they should also provide clear guidance on how this consent should be obtained to ensure that users make free, informed choices.
* Nouwens et al. also study the influence of certain design features on users’ choices when dealing with consent pop-ups, this time using a scraper to collect data. To conduct their research, a scraper was built to collect data from the five Consent Management Platforms (CMPs), which were introduced to help websites conforming with GDPR’s requirements to get users’ consent. The scraper was built using the Python library Scrapy [27] and the JavaScript rendering service Splash [28]. The variables collected by the scraper, on the top 10,000 websites in the UK, included the cookie-consent notification style (banner, barrier, etc.), the type of consent (implicit or explicit) and the specific user actions counted as consent, for example scrolling, closing the pop-up or clicking the page. The data collected also included the existence or not of accept and reject-all buttons, and the minimum number of clicks necessary to make those buttons available.
* Overall, the study has shown that dark patterns and implied consent are ubiquitous, with only 11.8% of the studied websites meeting the minimal requirements based on the GDPR [8]. While the notification style seemed to have no effect on users’ choices, the research has demonstrated that, for instance, removing the opt-out button from the first page increases consent on average by 22 percentage points.
* This research paper also highlights the capacity of automated tools such as scrapers to detect unlawful illegal practices online. The authors even recommend that data protection authorities should make use of automated tools like they have designed to expedite discovery and enforcement.
* Both studies display the power of design and automated tools, not only for users and websites’ developers, but potentially for regulators and law enforcers.

### 2.2.3. Topic Modeling in Privacy Policies

Sources: [16]

* Privacy policies have drawn attention recently for two main reasons. Firstly, they are a key aspect of GDPR compliance, which has recently come into force. Secondly, they are legal binding documents but they are not carefully considered by users, which affect their willingness to share personal data and their expectations about their rights [16].
* Most research in this field has focused on understanding the structure and evolution of these policies, but not many have studied the topics that these policies address.
* There have been recent efforts to use machine learning and text mining methodologies for analysing privacy policies. Several research papers focused on extracting and mapping specific topics addressed by privacy policies [31, 32, 33, 34]. They mostly used supervised learning techniques, relying on tagged repositories and thus important human effort to annotate data.
* For example, Liu et. al [29] suggested using Hidden Markov Models (HMM) for privacy policy segmentation, but found that this performed similarly to the lower half of human evaluators.
* Ramanath et al. [30] actually made an attempt at using unsupervised learning-based techniques for analysing privacy policies, but the paper did not aim at topic modeling per se.
* Sarne et al 2019 paper [16] is the first research paper we have found that attempts to perform topic modelling based on unsupervised learning. The advantages identified to use unsupervised learning over supervised learning techniques is that it enables us to analyse any new corpus with a fraction of the effort required by supervised learning, which is particularly useful to understand specific classes of policies. This also enables comparison of privacy policies over time, which allows reflection on changes in the topic address, compliance with new regulations and evolution in trends.
* Topic modeling is an unsupervised learning method which “attempts to extract the most probable distribution of words into topics through an iterative generative process which terminates upon convergence” [16]. The process does not require prior labeling of the documents.
* A popular method for topic modeling is the Latent Dirichlet Allocation (LDA), in which the algorithm learns and refines different distribution parameters during the iterative training phase.
* Sarne et al used GraphLab (turi.com) to carry out their study, which is a machine learning platform that supports LDA-based topic modeling. They also relied on other machine learning toolkits such as mallet [35] and familia [36].
* The results were partly reviewed manually by a domain expert, and were encouraging, suggesting that topic modeling based on unsupervised learning can be highly effective.

## 2.3. Research Gaps

2.3.1. Limited research on legitimate interests

* Most of the academic papers found on legitimate interests are law papers. They mainly focus on the methodology of the Legitimate Interests Assessment (LIA).
* The methodologies are legal tests, designed for lawyers rather than computer scientists.
* Main focus is always on consent.

2.3.2. Lack of knowledge about the use of legitimate interests in practice

* It is largely assumed that consent is by far the most commonly used legal ground for data processing. However, no study has actually looked at this in practice, and attempted to find out the real importance of legitimate interests as a ground for processing.
* The heavy focus on consent might have caused a lack of focus on legitimate interests. While studying cookie banners, and whether the consent we are given is informed and meaningful, researchers are not considering the facts that many of these privacy policies actually mention relying on legitimate interests.
* To fill this research gap, it will be interesting to find out how many companies rely on legitimate interests based on their privacy policies.

2.3.3. Lack of transparency with regards to Legitimate Interests Assessments

* There is a very limited number of DPIAs and LIAs available online, and companies are reluctant to share these documents even on request. This reflects a lack of transparency with regards to organisations’ justifications for data processing, and what legal ground is relied upon for processing.
* We have found no research that focuses on finding out how companies justify the use of legitimate interests, and which evidence they provide to succeed the LIA.
* A possible way of finding out how organisations explain their legitimate interests for processing is to analyse privacy policies, and look for evidence that an LIA or balance test has been completed.