CLAUDETTE meets GDPR

Automating the Evaluation of Privacy Policies using Artificial Intelligence

Giuseppe Contissa - *Alma Mater – Università di Bologna, Italy*
Koen Docter - *European University Institute, Florence, Italy*
Francesca Lagioia - *European University Institute, Florence, Italy*
Marco Lippi - *Università di Modena e Reggio Emilia, Italy*
Hans-W. Micklitz - *European University Institute, Florence, Italy*
Przemyslaw Palka - *European University Institute, Florence, Italy*
Giovanni Sartor - *European University Institute, Florence, Italy*
Paolo Torroni - Alma Mater – *Università di Bologna, Italy*

Florence – Bologna – Reggio Emilia, 02/07/2018
http://www.claudette.eu/gdpr/
http://claudette.eui.eu

Study Report,
Funded by The European Consumer Organisation (BEUC)

The authors would like to thank Gerald Spindler and David Martin for their extensive comments, which allowed us to make the report clearer and better thought-through. Obviously, if there are any mistakes, they are fully our responsibility. We would welcome any feedback from the readers.

Executive Summary

This report contains preliminary results of the study aiming at automating legal evaluation of privacy policies, under the GDPR, using artificial intelligence (machine learning), in order to empower the civil society representing the interests of consumers. We outline what requirements a GDPR-compliant privacy policy should meet (comprehensive information, clear language, fair processing), as well as what are the ways in which these documents can be unlawful (if required information is insufficient, language unclear, or potentially unfair processing indicated). Further, we analyse the contents of privacy policies of Google, Facebook (and Instagram), Amazon, Apple, Microsoft, WhatsApp, Twitter, Uber, AirBnB, Booking.com, Skyscanner, Netflix, Steam and Epic Games. The experiments we conducted on these documents, using various machine learning techniques, lead us to the conclusion that this task can be, to a significant degree, realized by computers, if a sufficiently large data set is created. This, given the amount of privacy policies online, is a task worth investing time and effort. Our study indicates that none of the analysed privacy policies meets the requirements of the GDPR. The evaluated corpus, comprising 3658 sentences (80.398 words) contains 401 sentences (11.0%) which we marked as containing unclear language, and 1240 sentences (33.9%) that we marked as potentially unlawful clause, i.e. either a "problematic processing" clause, or an "insufficient information" clause (under articles 13 and 14 of the GDPR). Hence, there is a significant room for improvement on the side of business, as well as for action on the side of consumer organizations and supervisory authorities.

Contents

1. Introduction

During the weeks right before and after the 25th of May 2018, the day since which the GDPR[1] has been applicable, you might have received an email or two, informing you that the online services you use have amended their privacy policies. How many of them have you actually read?

For the purpose of this study, we have read and analysed 14 privacy policies. Specifically, those of Google, Facebook (and Instagram), Amazon, Apple, Microsoft, WhatsApp, Twitter, Uber, AirBnB, Booking.com, Skyscanner, Netflix, Steam and Epic Games. Fourteen. That is less than the number of services European consumers use on regular basis (if you consider that almost *every* website or mobile app has their own privacy policy). These 14 policies, taken together, are about 80.000 words long. That is longer than an average novel. The GDPR has significantly enhanced the legal position of individuals present in the EU with regards to the protection of their personal data. We enjoy new rights. Among them, we have a right to information. But will we have time and power to exercise them? Are we factually able to do what we are legally entitled to do?

Consumers are not alone, though. Next to the Supervisory Authorities, civil society is gearing up to help them. First legal actions, including those of Max Schrems's Noyb[2] and la Quadrature du Net[3] have already taken place. There are more to come. Still, given how many services operate out there, will human lawyers be able to control whether the data controllers fulfil the requirements placed on them by the GDPR?

> *What if the answer to this question is: no, there is just too much to read? Does it mean that our rights will remain just a nice theory on the paper? Not necessarily so. Help, we argue, might come from the technology that many fear will bring more harm than good: artificial intelligence.*

The CLAUDETTE project[4] has been established in order to attempt automating the legal analysis of terms of service and privacy policies of online platforms and services. If machines can detect spam and translate from one language to another, operate driverless cars and trade stocks, maybe they can also assist lawyers trying to pursue the consumer rights? Having established that the answer is: "yes, most probably they can!"[5] regarding the terms of service seen through the lens of the Unfair Contractual Terms Directive[6], we have now embarked on the task of automating the legal analysis of privacy policies seen through the lens of the GDPR.

This is a preliminary study. When we make this report publicly available, the GDPR has been applicable for little more than a month. Most online platforms and services have just amended their privacy policies. We read as many as we could, and tried to train the machine to analyse them as well as possible. We are excited by the results, though there is obviously room for improvement. But what we found is both promising (regarding the possibility to have AI-powered tools assist human lawyers in

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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).
[2] *GDPR: noyb.eu filed four complaints over "forced consent" against Google, Instagram, WhatsApp and Facebook* https://noyb.eu/.
[3] *Écrivons ensemble les plaintes contre les GAFAM* https://www.laquadrature.net/en/plainteGAFAM.
[4] "Automated CLAUse DETeCTER", http://claudette.eui.eu.
[5] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2018). CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. arXiv preprint arXiv:1805.01217.
[6] Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts.

evaluating the privacy policies) and alarming (regarding the content of the privacy policies under study). Hence, we decided to make the results, as preliminary as they are, immediately available to the public.

> *As our analysis of the contents of the privacy policies of the fourteen platforms in question suggests, there is room for improvement on the side of corporations. It seems to us that information provided to data subjects could be more concrete, the language conveying it could be less vague, and the types of processing could be less questionable. There is room for action, on the side of the civil society.*

Important caveat: we are scholars, not activists, nor lawyers working for public authorities. What we do is legal doctrine and legal informatics. We do not claim to have established that any company here infringes the law. On the contrary, we look at these documents with the same mind-set with which a historian might read letters sent by some merchants to a duke of Florence in 16th century. It is our research material. However, since law is a normative discipline, we cannot help to notice that activists, consumer organizations or lawyers working for public authorities might find our preliminary results of importance for their mission. All we say is that, to our best understanding of the legal material in question, there is a chance that the privacy policies under study fall short of fulfilling the GDPR's requirements.

The purpose of this report is three-fold.

First, we want it to serve as a tool for anyone – consumers, lawyers, journalists; but also companies of good will – to assess the lawfulness of privacy policies under the GDPR. The extensive threshold, outlining the types of potentially unlawful clauses, which we present in sections 3 and 4 of this report, has been developed in order to enable human annotators to create data from which a machine can learn how to analyse new policies automatically. However, we reckon, while these documents still must be read by humans, our analysis can help those who have will and time to analyse these documents, in order to better understand what exactly the GDPR requires policies to contain (or not contain).

Second, this report can serve as a source of preliminary data to all the aforementioned stakeholders. Even though our aim was not to provide a comprehensive analysis of the privacy policies under study, but to train a machine how to *assist* humans in doing so, a by-product of our creation of the training set resulted in interesting quantitative and qualitative material. Hence, we pair the normative threshold presented in sections 3 and 4 with numerous examples, as well as provide an overview of the set in section 5. And full tagged policies on our website: http://www.claudette.eu/gdpr/. However, please note that, at this stage, the annotation has been created for the machine, which learns to classify *sentences*, not provisions.

Third, the message we want to convey is that it is possible to largely automate the legal analysis of privacy policies using machine learning. We are probably still far from the day when machines will be able to replace humans in doing so, but we might be closer than one thinks to the day when they can significantly increase the efficiency of human lawyers' work. This will require a larger data set than we managed to create in this couple of weeks; but the results we see now are promising, from the machine learning perspective. However, the more entities join the effort, the sooner we will

get there. This will be especially in order to have the AI-systems work across languages other than English.

The report is structured as follows: in the second section we provide the context for the analysis, including the legal status of privacy policies under the GDPR, a few observations about the market practice, as well as the technological context in which we operate. Then, we provide an extensive overview of all the legal requirements that a properly designed privacy policy should meet. We begin with a short and concise description of the "golden standard", and then specify each and every part of the normative threshold used. Further, we provide an overview of the documents we have analysed, including both quantitative data and qualitative observations. What follows is the methodology and results of the machine learning experiments we have conducted. Finally, we close with some minor observations about the future of the GDPR in the current context, as well as further possibilities in automating the legal oversight and analysis.

The present report comes together with a website: http://www.claudette.eu/gdpr/ where anyone can see the latest versions of the privacy policies of the 14 online platforms and services under analysis, with their clauses already assessed for compliance with the GDPR. On the date of publication of this report such an assessment was done by humans, but if you read this sometime later, who knows – it may even be machine-made already.

2. Context: Law, Market Practice, Technology

2.1. Legal Context: the Legal Status of "Privacy Policies" under the GDPR

The first issue to be clarified and understood in this study is the legal status of "privacy policies", i.e. textual documents made public by online platforms and other controllers, in the light of the GDPR's legal framework. Somehow surprisingly, the GDPR does not use the term "privacy policy" or "privacy notice" in any of its 99 articles, nor does it *directly* oblige data controllers to create those. However, in the light of the totality of the obligations imposed on data controllers, in particular information duties laid down in art. 13 and 14, as well as data subjects' rights, Article 29 Data Protection Working Party (WP29)[7] suggests that every organization maintaining a website makes such a document available in way easy to find for data subjects.[8]

The GDPR applies to "the processing of personal data wholly or partly by automated means (...)" (art. 2.1), by controllers and processors established in the European Union (art. 3.1.) and those established outside of the EU, if they process data of subjects present in the EU, in order to offer goods and services, or monitor the subjects' behaviour (art. 3.2.). In this sense, the rules of the Regulation apply to the *action* of processing, i.e. what data controllers *do* with personal data of the EU residents, and only in some regard to the documents created by data controllers. These documents convey different types of information, and it is imperative to understand against what legal context this information is to be assessed.

The Regulation puts forward a robust normative framework, consisting of principles (Chapter II): lawfulness, fairness, transparency, data minimization, accuracy, storage limitation, integrity and confidentiality and accountability (art. 5); data subjects rights (Chapter III): transparency and information about processing (art. 12-14), right of access (art. 15), right to rectification (art. 16), right to erasure ('right to be forgotten', art. 17), right to restriction of processing (art. 18), right to data portability (art. 20), right to object (art. 21) and a right not to be subject to a decision based solely on automated processing, including profiling (art. 22); numerous rules and obligations applying to controllers and processors, including rules on accountability (Chapter IV), rules on data transfers to third countries and international organization (Chapter V), establishment and competences of Independent Supervisory Authorities (Chapters VI-VII), a whole range of available remedies, accompanied by rules on liability and penalties for infringement by data controllers and processors (chapter VIII), and special provisions relating to specific types of processing (Chapter IX). In short, the GDPR lays down rules specifying *how* personal data should be processed, *what* data subjects must *know*, and *what can happen* if any of these rules are infringed. In this legal context privacy policies come into play.

The central provisions to be considered are articles 12, 13 and 14 of the GDPR. Articles 13 and 14 specify *what* information should be provided to data subjects by data controllers when personal data

[7] Article 29 Working Party has been an advisory body set up by the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Its mandate included issuing opinions and recommendations on data protection matters, serving as a valuable source of insight regarding specific requirements of the Directive, and now the Regulations. As of 25th May it has been replaced by the European Data Protection Board. In this report, we refer to the documents issued by the Working Party using the original title. For the sake of brevity, sometimes the abbreviation "WP29" is used.

[8] Article 29 Working Party, *Guidelines on transparency under Regulation 2016/679*, adopted on 29 November 2017, as last revised and adopted on 11 April 2018, 17/EN WP260 rev.01, hereinafter "Transparency Guidelines".

is (art. 13) and is not (art. 14) collected from the data subject. Art. 12 lays down rules on *how* (in what form, using what language) this information should be provided.

Starting with *what* data controllers are obliged to inform data subjects about, art. 13 lists the following categories of information:

1. the identity and the contact details of the controller and, where applicable, of the controller's representative;
2. the contact details of the data protection officer, where applicable;
3. the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;
4. where the processing is based on point (f) of Article 6(1), the legitimate interests pursued by the controller or by a third party
5. the recipients or categories of recipients of the personal data, if any;
6. where applicable, the fact that the controller intends to transfer personal data to a third country or international organization (...)
7. the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
8. the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject or to object to processing as well as the right to data portability;
9. where the processing is based on point (a) of Article 6(1) or point (a) of Article 9(2), the existence of the right to withdraw consent at any time, without affecting the lawfulness of processing based on consent before its withdrawal;
10. the right to lodge a complaint with a supervisory authority;
11. whether the provision of personal data is a statutory or contractual requirement, or a requirement necessary to enter into a contract, as well as whether the data subject is obliged to provide the personal data and of the possible consequences of failure to provide such data;
12. the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

In addition, art. 14, applying to personal data collected *not* from data subjects, but from third parties (business partners, data brokers etc.) requires that data subjects are informed about:

13. the categories of personal data [obtained];
14. from which source the personal data originate, and if applicable, whether it came from publicly accessible sources.

All this information, according to art. 12 of the GDPR need to be provided "in a concise, transparent, intelligible and easily accessible form, using clear and plain language". This requirement gets strengthened "for any information addressed specifically to a child". Further, regarding the form, "the information shall be provided in writing, or by other means, including, where appropriate, by electronic means". This obligation is not contingent on any request of the data subject (unlike further information regarding actions taken by the controller when data subjects seek to exercise their right) and must be provided to data subjects regardless of whether they ask for it or not.

In the light of all these requirements, as signalled at the beginning of this section, Article 29 Working Party suggests that:

> *Every organization that maintains a website should publish a privacy statement/ notice on the website. A direct link to this privacy statement/ notice should be clearly visible on each page of this website under a commonly used term (such as "Privacy", "Privacy Policy" or "Data Protection Notice"). Positioning or colour schemes that make a text or link less noticeable, or hard to find on a webpage, are not considered easily accessible.*
>
> *For apps, the necessary information should also be made available from an online store prior to download. Once the app is installed, the information still needs to be easily accessible from within the app. One way to meet this requirement is to ensure that the information is never more than "two taps away" (e.g. by including a "Privacy"/ "Data Protection" option in the menu functionality of the app). Additionally, the privacy information in question should be specific to the particular app and should not merely be the generic privacy policy of the company that owns the app or makes it available to the public.*
>
> *WP29 recommends as a best practice that at the point of collection of the personal data in an online context a link to the privacy statement/ notice is provided or that this information is made available on the same page on which the personal data is collected.*[9]

Further, WP29 recommends:

> *the use of layered privacy statements/ notices, which allow website visitors to navigate to particular aspects of the relevant privacy statement/ notice that are of most interest to them (...) However, the entirety of the information addressed to data subjects should also be available to them in one single place or one complete document (whether in a digital or paper format) which can be easily accessed by a data subject should they wish to consult the entirety of the information addressed to them.*

In the light of this legal material, the following initial observations should be made:

First, even though the GDPR does not directly oblige data controllers to create privacy policies, the information duties imposed on controllers, both regarding the substance and the form, indirectly create such a legal obligation. In other words, without a publicly available privacy policy, provider of any online platform or service, will not meet the GDPR information requirements. Hence, every online platform or service must have a privacy policy.

Second, there are several types of information that must be provided to data subjects. Some, regarding data subject's rights, will be similar across the policies. Some, regarding the controller's identity and contact details, will be contingent on who the controller is. The rest, regarding categories of data processed, the purpose of processing, the legal basis of processing, the existence of automated decision making etc. is contingent on *what the controllers do*.

Third, a properly designed privacy policy will be comprehensive regarding content, i.e. convey all the required information, presented in a concrete way, so that data subjects, if they wish, might understand everything that is being done with their data; and comprehensible regarding the

[9] Transparency Guidelines, p. 8.

form. This means that the language used should be understandable to consumers, not only to the specialists. Importantly, data controllers should not use the "what we do is too complicated" defence, and claim that realizing both requirements on the same time is not possible.

Fourth, as a consequence of the previous observation, an improperly designed privacy policy, i.e. a policy that comes short of fulfilling the GDPR's requirements, and as a result infringes the GDPR, might in principle suffer from three types of shortcomings: 1) not including all the required information (if for example a data subject is not informed about his or her rights); 2) being not concrete enough (for example stating that "personal data will be used to improve the service", without specifying what data exactly, and for what type of improvement); 3) being incomprehensible or using unclear language. The exact legal threshold, paired with numerous examples, for all these three types of (potential) unlawfulness of privacy policies will be presented in detail in sections 3 and 4 of the report.

An important observation should be made here. As the WP29 notices:

> *There is an inherent tension in the GDPR between the requirements on the one hand to provide the comprehensive information to data subjects which is required under the GDPR, and on the other hand do so in a form that is concise, transparent, intelligible and easily accessible.*[10]

It is true that sometimes information that is less comprehensive might be more intelligible and vice versa. It is true comprehensive information might come short of being concise. However, this tension is not impossible to reconcile. The WP29 further outlines several strategies on how to deal with this problem, suggesting that the privacy policies are first and foremost layered, allowing consumers to have just an overview of how their data is processed if they wish, but on the same time giving them possibility to acquaint themselves with all the information. What is important is that already the "first layer", attempting to be more concise than comprehensive, should be concrete and not vague.

Fifth, there is difference between unlawful design of a privacy policy, and an in-principle-well-designed privacy policy conveying information about unlawful processing. As stated at the outset of this section, the GDPR regulates first and foremost the act of processing, i.e. what data controllers and processors *do* with personal data, requiring that the processing is lawful, fair and transparent, and only as consequence of these requirements, what data controllers *write* in their privacy policies. Hence, one can imagine that controllers processing data in an unlawful manner, by for example collecting more data than necessary for their purpose, or processing it longer than necessary, or sharing data with third parties without legal basis etc. etc. write about it in their privacy policies. That is why, apart from assessing whether a privacy policy meets the requirements of art. 12, 13 and 14, one should also analyse the provided information in the light of other provisions of the GDPR. Obviously, the fact that no information about unlawful processing is present in a privacy policy does not yet mean that no unlawful processing takes place – there might clearly be a difference between what data controllers claim they do and what they actually do, the latter being much harder to monitor by civil society than the former – but this declaratory level is already an important reference point. The exact legal threshold for this analysis, paired with numerous examples, will be presented in detail in sections 3 and 4.

[10] Transparency Guidelines, p. 19.

2.2. Market Practice: Overwhelming Amount of Services and Privacy Policies

In 2018, an average consumer uses a significant amount of online platforms and services, each of which collecting data about him or her, and each of them having a privacy policy to be (potentially) consulted. Telephone providers, banks, social media, online shops, mobile games, news sites and many others collect data about their users and under GDPR are obliged to inform them about the categories of this data, how they use it, for what purposes, based on what basis etc. However, one cannot reasonable expect that everyone will have time to read all these documents. Of course, first time, second time, fifth time, might be interesting and exciting, but having to go through dozens of these documents, and then file dozens of complains to supervisory authorities, might be a bit overwhelming. Too overwhelming to actually make use of the rights.

Indeed, the days directly preceding the 25th of May 2018, the date since when the GDPR is applicable, might have felt overwhelming for many consumers. Mass emailing, mass privacy changes, mass pop-ups on every website, created the feeling of haste, as if GDPR has not been in force for two years already (as a reminder: the GDPR has been adopted on 27th of April 2016). All this resulted in many individuals expressing their annoyance with having to accept and review the updated privacy policies, and the companies seemed to convey the message "we are really sorry you have to go through this, but we are obliged by law to send you this spam". Suddenly, to astonishment of many privacy activists, social media were full of jokes about the GDPR, instead of people announcing that they now take a few days off to consult all the privacy policies, and make use of their right to object on all the pages that track them. Clearly, after several weeks/months this situation will probably stabilize, and the need to consult privacy policies will be spread over much long time-spans – ideally, whenever a consumer installs a new app, or creates an account on a new service, he or she would consult the document. However, this might be more difficult than it sounds.

One notable example from the market practice should be mentioned here. Washington Post, one of the leading American news sites, in the aftermath of GDPR's applicability, introduced a business model which came as a shock to many observers. Currently, consumers wishing to read the news on this site have three options: free, when they "read a limited number of articles each month & consent to the use of cookies and tracking by us and third parties to provide you with personalized ads", Basic Subscription, already paid, when they get "unlimited access to washingtonpost.com on any device & unlimited access to all Washington Post apps" but still "consent to the use of cookies and tracking by us [Washington Post] and third parties to provide you with personalized ads", or Premium EU Subscription, more expensive of course, when apart from unlimited access they get the privilege of "No on-site advertising or third-party ad tracking"[11]. It seems that not being tracked by the news outlets and third parties is currently a premium option. Whether this business model is lawful or unlawful, whether providing paid-for services and still tracking consumers pursuing legitimate interests is GDPR compliant, remains outside of the scope of this study. Some commentators deemed this approach shocking; others welcomed it as finally starting to use privacy as part of the market competition. What matters for this study is that currently the market practice is such that numerous websites, include news sites, will not only collect data about users' behaviour on their platform; they will also track them on other sites, and help other parties track them on other sites. This information is publicly available, Washington Post provides a list of all third party partners that will track users of their platform[12] (there is 21 of them), but when one scales this up, and realizes that an average consumer might be accessing

[11] Source: https://www.washingtonpost.com/gdpr-consent/, as of June 1st 2018.
[12] https://www.washingtonpost.com/third-party-partners/.

dozens of websites every day, the burden of having to check all this might be just too high. Hence, the rights are granted, but consumers might have no factual ability to exercise them – not only because companies make it too difficult, but also because there is just too many of them.

Here is where the civil society comes into play. Consumer organizations and activists whose job it is to control the data controllers. "Sousveillance" – watching them watching us. A few notable initiatives took place in the aftermath of May 25th, including the legal action by Max Schrems's Noyb[13] and la Quadrature du Net[14]. However, even these actions are directed only against the biggest companies. While there are hundreds, if not thousands, to check.

Our previous research on automation of legal analysis of terms of online services[15] (the other document, next to the privacy policy, which a user of any online platform is obliged to claim to have read, understood and accept) indicates that consumer contracts, despite legislation on unfair terms being in force since 1990s, and despite consumer organizations' and agencies' competence to initiate the abstract control of them, often tend to contain unfair contractual clauses.[16] There might be many reasons behind this undesirable state of affairs, but one of them is the incredible gap between the amount of companies and corporations whose terms of service (and now privacy policies) need to be checked, and the factual capabilities of consumer organizations. There is just too much to check. The same challenge awaits those wishing to analyze the privacy policies. That is where, we claim, the information technology, in particular artificial intelligence, can come in to assist the civil society.

2.3. Technological Context: Artificial Intelligence in the Service of the Civil Society

Artificial intelligence and big data attracted a lot of scholarly attention in the last years, and came to be seen mostly as a source of risks to individual rights, risks that need to be mitigated by regulation. This is, without a doubt, true, and in one of the projects run at the European University Institute we try to map these risks and push the normative and regulatory debate forward.[17] However, we claim that artificial intelligence, just as almost any technology, can be used for good and for bad, and that political efforts should concentrate not only on mitigating the risks that usage of AI by states and companies bring (to state once again: this is important and should obviously be done), but *also* on harvesting the power of AI for the sake of the people. Artificial intelligence can empower the civil society, individual consumers, consumer bodies and organizations.

A couple of words on AI are needed here. After 60 years of its inception, in the last decade, artificial intelligence finally started being commercially applied across different business sectors. As Nello Cristianini points out, this has become possible mostly due to enormous amount of data that lately became available and easily accessible.[18] Now, a lot of this data is personal data, what raises a number of data protection issues, hence the importance of the GDPR. The internet has become a ubiquitous infrastructure for, at the same time, delivery of AI-driven applications and for the collection

[13] *GDPR: noyb.eu filed four complaints over "forced consent" against Google, Instagram, WhatsApp and Facebook* https://noyb.eu/.
[14] *Écrivons ensemble les plaintes contre les GAFAM* https://www.laquadrature.net/en/plainteGAFAM
[15] See Micklitz, H. W., Pałka, P., & Panagis, Y. (2017). The Empire Strikes Back: Digital Control of Unfair Terms of Online Services. Journal of Consumer Policy, 40(3), 367-388; Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Panagis, Y., Sartor G. & Torroni, P. (2017). Automated Detection of Unfair Clauses in Online Consumer Contracts. Legal Knowledge and Information Systems, 145; Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2018). CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. arXiv preprint arXiv:1805.01217.
[16] Ibid.
[17] See the ARTSY Project: *Before the Machines Consume the Consumers*: https://artsy.eui.eu/about/.
[18] Cristianini, N. (2016), The road to artificial intelligence: A case of data over theory, "New Scientist", https://www.newscientist.com/article/mg23230971-200-the-irresistible-rise-of-artificial-intelligence/.

of personal data. To capture this phenomenon, Mireille Hildebrandt popularized the term "onlife world"[19] while describing the environment we currently all inhabit:

> *Our current life world can no longer be described by dichotomizing online and offline, which suggests that we require a new term to more adequately depict our current predicament. 'Onlife' singles out the fact that our 'real' life is neither on- nor offline, but partakes in a new kind of world that we are still discovering.*[20]

This is to say: we obviously live in our physical bodies and physical places; and obviously do a lot of things in what used to be called "cyberspace" – read news, socialize on social media, do shopping, listen to music, watch movies, play games etc. – but the latter, with the introduction of portable smart devices got so neatly blended in our existence that it no longer makes sense to speak of the 'physical' and the 'digital' as if those were two completely separated spheres. As a result, many actions we undertake leave a print, a trail, in the information systems. A print we might know about if we sit down and reflect about it, but a print that we seldom actually notice. This print, in form of data, is stored by different actors.

Ethem Aplaydin, one of the leading experts in machine learning, tells the following story: With the introduction of smartphones, being more computers than phones, computers constantly staying online and being connected to each other (Internet of Things, ubiquitous computing), on which numerous services run, "an increasingly larger part of our lives is recorded and becomes data".[21] This data is a by-product of all the onlife activity, and so throughout the last two decades we moved from the situation in which data needed to be actively collected and constituted a burden, to a current situation when data "generates itself", or "we" generate it, and it can be treated as a resource. Why?

In traditional programming, a software engineer trying to solve a problem would sit down and try to figure out a way for a computer to do something. Some tasks, like mathematical problems, were easy; others, like automatic translation and image recognition, extremely hard. With the introduction of machine learning, this idea is put upside-down. Instead of telling a computer how to realize a particular task, a programmer feeds it with enormous amount of data, both regarding the input and output, and lets the machine figure out by itself how to do this, and sometimes even what the task is. Successful applications of ML are currently being used by numerous consumers – it is enough to mention SPAM filters, machine translation or voice recognition. In order to introduce them, however, huge amounts of data were necessary. And now, in the onlife world, the data is here.

Business also uses this information to pursue their goals. The merger of big data, coming from ubiquitous computing, with machine learning algorithms, profiling and the advances in behavioural sciences, enables traders to recognize patters in consumers activities, and use this knowledge to predict their behaviour and influence it, using, among others, targeted advertising. That is why consumer data is so valuable to business. That is why everyone wants to track you. That is why you have to pay Washington Post more if you do not want to be tracked. An important thing to understand is that this data generation and processing in many ways is necessary for the IT solutions that make consumers' life easier to function. If we want to use social media, a lot of data about us, given the very

[19] The term originally coined by the team lead by Luciano Floridi, see: Floridi, L. (2015). The onlife manifesto. Springer-Verlag GmbH.
[20] Hildebrandt, M. (2015). Smart technologies and the end (s) of law: novel entanglements of law and technology. Edward Elgar Publishing, p. 42.
[21] Alpaydin, E. (2016). Machine Learning: The New AI. MIT Press, p. XI.

architecture of these systems, must be stored by the providers. Nevertheless, business uses the data also for other purposes than delivering us the services. And the real challenge of the GDPR enforcement is that numerous business models, based on processing of personal data which is not in the interest of consumers, are already in place.

However, there is hope. The same technology that businesses use to control consumer behaviour can be used by civil society to control businesses. One of the purposes of this study is to offer a proof for that claim.

Civil control is a lot of work. Given the amount of services to be potentially controlled, it is really a lot of work. But many of the tasks necessary for this work to be conducted can be automated.

For example, imagine you are a lawyer working for a consumer organization and want to know which companies claim they process consumer personal data in third countries. Currently, what you need to do is to check a privacy policy of each platform, one by one, and look for the specific clause. In a long and often non-transparent document. And then again. And again. With machine learning, this can be done by a machine. Reading privacy policies in search for particular types of information can be automated.

In order to do that, large amounts of data are necessary too. This data needs to be generated by researchers and/or civil society. What is this data? Basically, documents tagged in a way that allows the machine to learn what is it supposed to look for. The idea is simple: a lawyer reads a privacy policy, and marks the fragments that he or she finds problematic. For example, an incomprehensible clause. Or a clause using vague language. Or a clause stipulating that personal data will be transferred to third parties, without specifying who they are, what data or for what purpose. The whole list of things we are looking for is presented in section 4, below. After marking a whole document, the annotated policy is fed into the machine. And then another one. And another one. Lawyers show the computers policies with specific elements tagged, and the machine learns how to recognize them.

The more data, the better. The more examples the machine will have to learn from, the higher the chance it will be able to then correctly assess a privacy policy it has not yet seen before. How much is enough? This largely depends on the nature of the task. In our previous study[22], where we attempted to train the machine to analyse the terms of service of online platforms in search of potentially unlawful contractual clauses, we have created a corpus of 50 documents, and using different machine learning techniques, managed to achieve accuracy of more than 93% when detecting potentially unfair clauses about jurisdiction, choice of law, limitation of liability and "contract by using" (a type of clause stating that a user is bound by terms of service by simply accessing the platform), and more than 80% in total[23]. This, given that a machine is able to analyse in a matter of minutes a corpus that would require weeks of work by a human lawyer (about whom one should not assume that their work is flawless either), is a promising result. However, one should bear in mind that the exact number might depend both on the nature of the task, and the heterogeneity of the material to be learned from and analysed.

Two caveats are important here: machine learning powered systems look for potentially unlawful clauses (what we label "problematic"), and their function is to assist lawyers in undertaking

[22] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2018). CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. arXiv preprint arXiv:1805.01217.
[23] A prototype of a user-end tool making use of this technology, where everyone is invited to test for themselves how well the machine works, is available here: https://claudette.eui.eu/use-our-tools/

their tasks, and not to replace them.[24] This is because, as any lawyer knows well, there is much more to establishing unlawfulness of a textual provision than simply analysing how it is formulated. The context matters. Who is the company, what is the nature of the service they provide, what data exactly is being collected, how clear this was to consumers etc. Nevertheless, the difference in workload between having to go through hundreds of pages of unstructured text, and having to analyse limited amount of provisions with an indication of what the problem might be, can be immense. And the larger the object of study, the more significant the efficiency gains.

The ambition of this study has been to test to what extent discovery of potentially unlawful clauses, according to GDPR, as well as checking whether the required clauses are present in the privacy policies, is possible. However, as we will briefly outline in the last section of this document, the potential of automation goes far beyond this task. For now, however, let us better explain the nature of the legal task we attempt to automate at this stage.

[24] Surden, H. (2014). Machine learning and law. Wash. L. Rev., 89, 87.

3. The Methodology for Evaluating Privacy Policies: the "Golden Standard" and the Macro-Categories of Failures

Privacy policies, as noted in the section 2.1. ("Legal context") above, should be comprehensive (regarding the information it provides), comprehensible (regarding the form of expression), and substantively compliant (regarding the types of processing they foresee). Thus, our "golden standard" for privacy policies includes three dimensions:

A. Comprehensiveness of information: the policy should include all the information that is required by articles 13 and 14 of the GDPR;
B. Substantive compliance: the policy should only allow for the types of processing of personal data that are compliant with the GDPR;
C. Clarity of expression: the policy should be framed in an understandable and precise language.

With regard to these three dimensions one can distinguish two levels of achievement:

1. Optimal achievement: In this case the policy clearly meets the GDPR requirements along the dimension at issue.
2. Suboptimal achievement: In this case, the policy fails to clearly meet the GDPR requirement at issue. In some cases we have found it useful to distinguish two levels of suboptimal achievement:
    a. Questionable achievement: it may be *reasonably doubted* that the suboptimal policy reaches the threshold required by the GDPR, along the dimension at issue.
    b. Insufficient achievement or no achievement: the suboptimal policy *clearly fails* to reach the threshold required by the GDPR along the dimension at issue.[25]

Optimality along the three dimensions (comprehensiveness, clarity and substantive compliance) of our golden standard means the policy fully corresponds to that standard.

On the contrary, suboptimal achievement along at least one of these dimensions signals that the policy at issue may fail to comply with the GDPR. Thus, a lawyer, an activist, a journalist, or a consumer should pay particular attention to the clauses which fail to reach the corresponding thresholds.

[25] In our tagging, we shall distinguish between sub-level (2.a) and (2.b), indicating such levels with numbers 2 and 3 respectively, with regard to the dimension C (legality of processing). Both sublevels will be covered by the number 2, with regard to dimensions A and B (comprehensiveness and clarity).

Here is an example of a clause that fails to meet optimality according to *comprehensiveness*.

Example 1.
The controller will use a variety of third party service providers to help us provide services related to Our Platform and the Payment Services. Service providers may be located inside or outside of the European Economic Area ("EEA"). In particular, our service providers are based in Europe, India, Asia Pacific and North and South America.

Rationale
The clause fails along the dimension of comprehensiveness since it does not identify the recipients of the information

The following clause, on the other hand, fails according to *clarity*.

Example 2.
When you as a Guest submit a booking request, certain information about you is shared with the Host (and Co-Host, if applicable), including your full name, the full name of any additional Guests, your cancellation history, and other information you agree to share.

Rationale
The clause fails along the dimension of clarity since it does not specify what information will be transmitted to the Host, in addition to the items expressly mentioned ("certain information (…) including")

Finally, the following on fails according to according to *substantive compliance*.

Example 3.
The controller may provide information to its vendors, consultants, marketing partners, research firms, and other service providers or business partners.

Rationale
The clause fails along the dimension of substantive compliance since it fails to specify under what conditions and compatible purposes the data will be communicated to third parties, or who the third parties are.

As described in section 6, the purpose of our system is to automatically identify clauses that appear to be defective along at least one of the above dimensions, and in this way to support experts by pre-selecting the clauses they should critically examine.

For this purpose the computer system has to be trained to recognize such clauses, being provided with a set of examples. The examples consist of policies where relevant clauses have been appropriately tagged, distinguishing their category and whether they are optimal or defective. A single clause in some cases may fall in different categories and consequently have multiple tags.

Example 4.
We automatically collect log data and device information when you access and use Our Platform, even if you have not created an Account or logged in. That information includes, among other things: details about how you've used Our Platform (including if you clicked on links to third party applications), IP address, access dates and times, hardware and software information, device information, device event information, unique identifiers, crash data, cookie data, and the pages you've viewed or engaged with before or after using Our Platform.

Rationale
The clause fails along both the dimensions of clarity and substantive compliance: on the one hand it vaguely refers to the information being collected through the locution "among other things"; on the other hand it allows for types of processing having no relevant purpose.

4. Analytical Classification of Suboptimal Clauses and Tagging Methodology

In the following sections we will describe our method for tagging the documents in the training set. The method involves distinguishing different ways in which a clause may fail to meet the golden standard, along one the three dimensions, and applying corresponding tags. Please note that this is a method of creating a data set for the machine learning system, the purpose of which is to identify clauses which an expert lawyer should then critically examine. Hence, the detailed instruction presented below is an instruction aimed at human taggers training the machine, and not directly at lawyers assessing privacy policies. This said, we believe that it can be of high value for the latter as well. The most significant difference is that a machine learns how to classify *sentences*, while a human lawyer would sometimes treat different units (only parts of sentences, or sometimes whole paragraphs) as relevant provisions. This difference is predicated by the features of technology used (Support Vector Machines, see section 6), and the function of the system, which is to *assist* a human lawyer by indicating what to pay attention to and why, rather than replace him or her.

4.1. A. Comprehensiveness of Information

The dimension of comprehensiveness of information concerns whether a privacy policy meets all the information requirements of art. 13 and 14 of the GDPR, or fails to do so, either by not providing at all the required item of information, or by providing it insufficiently or imprecisely.

We have identified 12 types of required information clauses. Table 1 below specifies the XML tags used to identify the clauses containing each type of information. For each type of required information, we have classified the corresponding clause either as a *fully informative* (all the required information is present and well specified); or as insufficiently *informative* (information is hinted at, but non-comprehensive).[26] We have not tagged the complete omissions of an item of required information since machine learning methods will be used to detect omissions.

[26] Comprehensiveness of information is particularly relevant with regard to consent. According to the Article 29 Working Party, "[T]here must always be information before there can be consent" (Article 29 Working Party, "Opinion 15/2011 on the definition of consent", l.c.,

Fully informative and insufficiently informative clauses have been distinguished by adding respectively numbers 1 and 2 to their tags. For instance, <id1> indicates that the clause concerning the identity of the controller is fully informative, while <id2> indicates that it is insufficiently informative.

As noted above, a single clause in some cases may fall in different categories and consequently may have multiple taggings. For sake of clarity, each category is discussed in the following showing just the specific tag of the discussed category, even if others might be associated to the same clause.

Type of required information	Symbol:
Identity of the controller and (where applicable) controller's representative	<id>
Contact details of the controller and (where applicable) of controller's representative	<contact>
Contact details of the data protection officer	<dpo>
Purposes of the processing for which the personal data are intended	<purp>
Legal basis for the processing	<basis>
Categories of personal data concerned	<cat>
Recipients or categories of recipients of the personal data	<recep>
Period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period	<ret>
Existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject and to object to processing as well as the right to data portability	<correct>
Right to lodge a complaint with a supervisory authority	<complain>
Source from which personal data originate	<source>
Existence of automated decision-making, including profiling	<auto>

Table 1

4.1.1. Identity of the controller and (where applicable) controller's representative

The GDPR establishes obligations to provide the identity and the contact details of the controller (as defined in art. 4(7)) and, where applicable, of the controller's representative both when personal data are collected from the data subject (art. 13(1)(a)) and when they have not been obtained from the data subject (art. 14(1)(a)). The *ratio* is to enable the exercise of the data subjects rights towards the controller or its representative, also in application of the transparency principle (art. 5(1)).

Moreover, according to WP29 Guidelines for Consent, for the consent to be informed, it is necessary to provide the data subject of certain elements required for obtaining valid consent, among the others "the controller's identity"[27]. See also recital 42 GDPR: "[...]For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended.[...]".

Please note: Under the GDPR, the consent for processing must be "clearly distinguishable from the other matters" (art. 7.2.), meaning that if processing is based on consent, this consent must be

p. 19). Thus For consent to be "informed" under data protection law, the subject must be able to appreciate and understand the facts and implications of his/her action.

[27] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017, as last Revised and Adopted on 10 April 2018, p. 13.

provided separately from accepting the privacy policy and should not be "hidden" there. Moreover, it is possible that personal data is processed based on other legal basis than consent (performance of contract, legitimate purposes etc.); in what case the information requirements form art. 13 and 14 obviously still apply. We refer to the WP29 guidelines on consent in this report, since that document is a valuable source of knowledge on what "informed" means, and allows us to assess whether information requirements from art. 13 and 14 are met or not. In other words, we apply it *mutatis mutandis,* to ground the legal analysis in even more normative material. What we clearly do not claim is that a consent hidden in a privacy policy, if meeting the requirements, would be valid.

We classified the identity clause as *fully informative* (1) when the identity is fully specified; and as *insufficiently informative* (2) when the identity of the controller is vague and the information about the controller is insufficient. For instance, consider the following examples:

Uber Privacy Policy (last updated on 25 May 2018)
<id1>If you use our services in the United States, Uber Technologies, Inc. is the data controller for your information.</id1> <id1>If you use our services in the European Union or elsewhere, Uber B.V. is the data controller.</id1>

Rationale
The above clause is fully informative concerning the identity of the controller.

Netflix Privacy Policy (the version updated on 1 January 2017)[28]
<id2>Your Netflix service provider and data controller depends on your country of membership, and will be listed in your membership or payment confirmation email.</id2>

Rationale
The above clause fails to be fully informative since it fails to provide the data subject with easily accessible information on the data controller: to identify the controller, the data subject is supposed to check in the membership or payment confirmation email, presumably received by the data subject after having already given his or her consent and concluded the contract. As noted in WP29 Guidelines on Transparency,[29] the requirement that the provision of information to, and the communication with, data subjects should be done in a easily accessible manner means that the data subject should not have to seek out the information; it should be immediately apparent to them where and how the information can be accessed, for example by providing it directly to them or by linking them to it. Thus, this clause undoubtedly represents a low standard for the clarity and accessibility of the information.
PLEASE NOTE: this clause comes from Netflix's previous version of privacy policy, dated 1 January 2017. In the most recent version (last updated on 11 May 2018) the identity of the controller is fully specified. We retain it here as an **example** of formulation that the machine classifier should look for.

[28] Please note that the Netflix reported clause comes from a previous version dated on 1 January 2017. In the most recent version (last updated on 11 May 2018) the identity of the controller is fully specified.
[29] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679 p 8

Airbnb Privacy Policy (last updated on 16 April 2018)

<id2>If you change your Country of Residence, the Data Controller and/or Payments Data Controller will be determined by your new Country of Residence as specified above, from the date on which your Country of Residence changes.</id2>

Rationale

This clause, taken in isolation, fails to be fully informative since it requires the data subject to retrieve information and engage in inferences in order to identify the data controller. Thus, even though the criteria to determine the controller's identity are specified (i.e. the country of residence and the date on which it changes) the clause above only reaches a low standard for the clarity and accessibility.[30]

PLEASE NOTE: in the Airbnb privacy policy (last updated on 16 April 2018) the mentioned clause is preceded by a number of clauses that fully specify the controller's identity depending on the data subject country of residence. We present it here, since the machine learns how to classify **sentences**. This is an example of where a particular sentence could be problematic in itself, though not necessarily so if read in conjunction with the rest of the document (see section 5). The reader can see how this looks like, on the user's side, at: http://www.claudette.eu/gdpr/answers/AirBnB.html However, it is important for the machine to learn that such a sentence, as a matter of general rule, should in itself be marked, since a different privacy policy could contain a similar one, without additional information in other parts of the document.

4.1.2. Contact details of the controller and (where applicable) of the controller's representative

The GDPR establishes obligations to provide contact details of the controller (as defined in art. 4(7)) and, where applicable, of the controller's representative both when personal data are collected from the data subject (art. 13(1)(a)) and when they have not been obtained from the data subject (art. 14(1)(a)). As above, the *ratio* is to enable the exercise of the data subject's rights toward the controller or its representative, also in application of the transparency principle (art. 5(1))[31]. In particular, as noted in WP29 Guidelines on Transparency[32] the GDPR required information in relation to the exercise of data subjects' rights and the nature of the information required are designed to meaningfully position data subjects so that they can vindicate their rights and hold data controllers accountable. Besides, the modality provided by a data controller for data subjects to exercise their rights should be appropriate to the context and the nature of the relationship and interactions between the controller and the data subject.

In particular, with regard to the contact details of the controller and (where applicable) of the controller's representative, according to the WP29 guidelines on information requirements[33], this

[30] It should be noted that, in the Airbnb privacy policy (last updated on 16 April 2018) the mentioned clause is preceded by a number of clauses that fully specify the controller's identity depending on the data subject country of residence.

[31] As noted in Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679 (p26), as last revised and adopted on 11 April 2018 , the GDPR required information in relation to the exercise of data subjects' rights and the nature of the information required are designed to meaningfully position data subjects so that they can vindicate their rights and hold data controllers accountable. Besides, the modality provided by a data controller for data subjects to exercise their rights should be appropriate to the context and the nature of the relationship and interactions between the controller and the data subject.

[32] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018, p 26

[33] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018

information should allow for different forms of communication with the data controller. Recital 59 of the GDPR also emphasises that, in order to facilitate the exercise of data subject's rights, data controller should "also provide means for requests to be made electronically, especially where personal data are processed by electronic means". Thus, we classified a clause as *fully informative* (1) when the information allows for different forms of communication with the data controller and (where applicable) of controller's representative, e.g. phone number, email, postal address etc.; and it seems to be adequate with regard to the context, the nature of relationship and interactions between the controller and the data subject; conversely we classified a clause as *insufficiently informative* (2) in any other case then above. Consider the following examples:

Facebook Privacy Policy (last updated on 19 April 2018)
<contact1>The data controller responsible for your information is Facebook Ireland, which you can contact online, or by mail at: Facebook Ireland Ltd. 4 Grand Canal Square Grand Canal Harbour Dublin 2 Ireland</contact1>

Rationale
The above clause is fully informative concerning the contact details of the controller. It provides two forms of communication, i.e. an online form (there is a hyperlink under the words "contact online") and the postal address. Moreover, the online form, that is an electronic means, is an adequate form of communication with regard to the context, the nature of relationship and interactions between the controller and the data subject.

Netflix Privacy Policy (last updated on 11 May 2018)
<contact2>For questions specifically about this Privacy Statement, or our use of your personal information, cookies or similar technologies, please contact our Data Protection Officer/Privacy Office by email at privacy@netflix.com.</contact2>

Rationale
The clause above fails to be fully informative since it provides only one means of communication, i.e. an email address, even though it constitutes an adequate means of communication, considering the context and the nature of the relationship and interactions between the controller and the data subject. As noted above, the provided contact details should allow for different forms of communication with the data controller. Thus, it only reaches a low standard for the clarity and accessibility of the information.

It should be noted that, in the analysed corpus, usually controllers provide maximum two forms of communication, i.e. alternatively postal address and online form; or postal address and email; or postal address and a generic company phone number.

4.1.3. Contact details of the data protection officer

Article 37(7) of the GDPR requires the controller or the processor to publish the contact details of the DPO and to communicate the contact details of the DPO to the relevant supervisory authorities. The objective of these requirements is to ensure that data subjects can easily and directly contact the DPO without having to contact another part of the organisation.

According to WP29[34], the contact details of the DPO should include information allowing data subjects to reach the DPO in an easy way (a postal address, a dedicated telephone number, and/or a dedicated e-mail address). When appropriate, for purposes of communications with the public, other means of communications could also be provided, for example, a dedicated hotline, or a dedicated contact form addressed to the DPO on the organisation's website. Article 37(7) does not require that the published contact details should include the name of the DPO. Whilst it may be a good practice to do so, it is for the controller or the processor and the DPO to decide whether this is necessary or helpful in the particular circumstances.

We classified as *fully informative* (1), clauses including the name of the DPO, a postal address, a dedicated telephone number, and/or a dedicated e-mail address and, when appropriate, for purposes of communications with the public, other means of communications such as a dedicated hotline, or a dedicated contact form addressed to the DPO on the organisation's website. We classified them as *insufficiently informative* (2) in any other case then above. For instance, consider the following examples:

Facebook Privacy Policy (last updated on 19 April 2018)
<dpo2>Contact the Data Protection Officer for Facebook Ireland Ltd.</dpo2>

Rationale
The clause above fails to be fully informative since it generically refers to the possibility of contacting the DPO but does not provide the DPO name and a postal address, only a link to an online form. Thus, it only reaches a low standard for the clarity and accessibility of the information.

Airbnb Privacy Policy (last updated on 16 April 2018)
<dpo2>You may exercise any of the rights described in this section before your applicable Airbnb Data Controller and Payments Data Controller by sending an email to dpo@airbnb.com.</dpo2>

Rationale
The clause above fails to be fully informative since it only provides a dedicated email address, omitting both the name of the DPO and a postal address. Thus, it only reaches a low standard for the clarity and accessibility of the information.

We could not retrieve an example of a *fully informative* clause from the privacy policies we analysed.

[34] Article 29 Data Protection Working Party, WP243rev.01 Guidelines on DPO Adopted on 13 December 2016 As last Revised and Adopted on 5 April 2017.

4.1.4. Purposes of the processing for which the personal data are intended

According to art. 13(1)(c) and 14(1)(c), the controller shall provide information related to the purposes of the processing for which the personal data are intended. Besides, according to WP29 Guidelines for Consent, for the consent to be informed, it is necessary to provide the data subject with certain information, among others "the purpose of each of the processing operations for which consent is sought"[35]. See also recital 42 GDPR, according to which "[...]For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended.[...]".

Moreover, art6(1)(a) confirms that the consent of the data subject must be given in relation to "one or more specific" purposes and that a data subject has a choice in relation to each of them. The requirement that consent must be 'specific' aims to ensure a degree of user control and transparency for the data subject.

The controller must apply purpose specification as a safeguard against function creep, that is a phenomenon that identifies the gradual widening or blurring of purposes for which data is processed, after a data subject has agreed to the initial collection of the data[36].

Thus, we classified as *fully informative* (1), clauses where the purposes of the processing are exhaustive and not vague; and as *insufficiently informative* (2) clauses following outside the cases described above, for example when a clause only provides examples. Consider for instance the following examples:

Airbnb Privacy Policy (last updated on 16 April 2018)

<purp1>If you are a Host, the Payments Data Controller may require identity verification information (such as images of your government issued ID, passport, national ID card, or driving license) or other authentication information, your date of birth, your address, email address, phone number and other information in order to verify your identity, provide the Payment Services to you, and to comply with applicable law.</purp1>

Rationale

The above clause is fully informative since it clearly specifies the purposes of the processing with regard to specific data or categories of data and it is clear how the collected data help to verify the data subject identity and to provide the payment services.

[35] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018, p. 13.

[36] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018, p. 12.

WhatsApp Privacy Policy (last updated on 24 April 2018)
<purp2>WhatsApp must receive or collect some information to operate, provide, improve, understand, customize, support, and market our Services, including when you install, access, or use our Services.</purp2>

Rationale
The clause above fails to be fully informative since it generically refers to "some information" and it is vague as to the purposes of the processing. In particular, it is unclear (i) what type of data will be processed; (ii) what is the correlation between the collected information and the specific purposes, since the clause only lists a number of different purposes one after the other (iii) what "improve" "customize", "understand", "support" or "market" entails, for example what kind of customization this refers to; (iii) the type of analysis which the controller is going to undertake as well as what the consequences of the processing entails

4.1.5. Legal basis for the processing

According to art. 13(1)(c) and 14(1)(c), the controller shall provide information related to the legal basis for the processing, where the relevant legal basis must be specified according to art. 6(1) for personal data and to art. 9 for special categories of personal data. According to WP29 Guidelines on transparency[37], where legitimate interests (Article 6.1(f)) is the legal basis for the processing, the specific interest in question must be identified. As a matter of best practice, the data controller should also provide the data subject with the information from the balancing test, which should have been carried out by the data controller to allow reliance on Article 6.1(f) as a lawful basis for processing, in advance of any collection of data subjects' personal data.

We classified as *fully informative* (1) clauses where at least one of the conditions laid out in Art.6 and for special categories of personal data, Art.9 is specified; and as *insufficiently informative* (2) the clauses following outside the cases described above, for instance when a clause provides only examples of legitimate interests, or the interest at stake does not appear to be legitimate. For instance, consider the following examples:

Airbnb Privacy Policy (last updated on 16 April 2018)
<basis1>The Payments Data Controller needs to collect the following information, as it is necessary for the adequate performance of the contract with you and to comply with applicable law (such as anti-money laundering regulations).</basis1>

Rationale
The above clause is fully informative since it clearly describes the legal basis for the processing, i.e. the adequate performance of the contract.

[37] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018.

Facebook Privacy Policy (last updated on 19 April 2018)

<basis1>We access, preserve and share your information with regulators, law enforcement or others: In response to a legal request, if we have a good-faith belief that the law requires us to do so.</basis1>

Rationale

The above clause is fully informative since it clearly describes the legal basis for the processing, i.e. the compliance with a legal obligation to which the controller is subject.

Uber Privacy Policy (last updated on 25 May)

Uber must collect and use certain information in order to provide its services. This includes:[…] <basis2>Background check Information necessary to enable drivers to provide transportation services through the Uber app.</basis2>[…]

Rationale

The clause above fails to be fully informative. In particular it claims that background check Information is necessary to enable drivers to provide transportation services. However, under the terms of GDPR (art. 10), background screening can only occur under very specific conditions, namely (i) that the processing is under the control of an official authority; and (ii) that the organisation is authorised by EU Member State's law for providing appropriate safeguards. Moreover, it should be noted that according to art. 6(4)(c) of the GDPR "Where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia: […] the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10 […]. On the basis of the above-mentioned rules, to safeguard the rights and freedoms of the data subject, the privacy policy should at least mention (or provide a link to) a document concerning the authorisation and the official authority in charge of controlling the activity.

Booking Privacy Policy (last updated on 9 May, 2018)

<basis2>We may use your information for our legitimate interests, such as to provide you with the best suitable content of the website, emails and newsletters, to improve and promote our products and services and the content on our website, and for administrative, fraud detection and legal purposes.</basis2>

Rationale

The clause above fails to be fully informative since it only provides some examples of legitimate interests, which seems to be extensively applied to marketing practices and mixed with other very different purposes such as legal purposes. As a best practice and in application of the transparency principle (article 12 GDPR), legitimate interests should be clearly differentiated from other types of basis for the processing.

4.1.6. Categories of personal data concerned

According to article 14(1)(d), where personal data have not been obtained from the data subject, the controller shall provide the data subject with the certain information, among others, the categories of personal data concerned. This information is required in an article 14 scenario because the personal data has not been obtained from the data subject, who therefore lacks the awareness of which categories of personal data the data controller has obtained.

It should be also noted that, as reported by the Guidelines on Transparency[38], WP29's position is that "there is no difference between the status of the information to be provided under sub-article 1 and 2 of Articles 13 and 14 respectively. All of the information across these sub-articles is of equal importance and must be provided to the data subject."

Besides, whenever the data subject consent constitutes the legal basis for the processing, under article 6 and 9 of the GDPR, the categories of personal data concerned should be provided to obtain a valid consent. In particular, article 4(11) of the GDPR defines consent as: "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her." According to WP29 Guidelines on Consent, for consent to be informed, it is necessary to provide the data subject of certain elements required for obtaining valid consent, among the others "what type of data will be collected and used"[39]. See also recital 42 GDPR: "[...]For consent to be informed, the data subject should be aware at least of the identity of the controller and the purposes of the processing for which the personal data are intended.[...]"[40].

We classified as *fully informative* (1) clauses where the categories of personal data are comprehensively specified and not vague; and as *insufficiently informative* (2) clauses falling outside the cases described above, such as when a clause only provides examples. For instance, consider the following examples:

Google Privacy Policy (effective on 25 may 2018)

```
<cat1>We collect information about your location when you use our services,
which helps us offer features like driving directions for your weekend getaway
or showtimes for movies playing near you.</cat1>
```

Rationale

```
The above clause is fully informative, since it clearly specifies the category
of personal data collected, i.e. the data subject location information.
```

[38] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018, p 14.
[39] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018, p. 13.
[40] See also WP29 Opinion 15/2011 on the definition of consent (WP 187) pp.19-20.

Steam Privacy Policy (last updated 25 may 2018)
<cat2>Personal Data we collect may include, but is not limited to, browser and device information, data collected through automated electronic interactions and application usage data.</cat2>

Rationale
The clause above fails to be fully informative since it only provides a summary of the collected data and the list is not exhaustive, as it is clear from the wording of the clause. Besides it is indubitably vague, unclear and not inferable in any way what types of data will be collected through automated electronic interactions and application usage data.

4.1.7. Recipients or categories of recipients of the personal data

According to art.13(1)(e) GDPR, the controller shall provide, among other information, "the recipients or categories of recipients of the personal data, if any". The term "recipient" is defined in Article 4.9 GDPR (and referenced in Recital 31), which clarifies that a recipient to whom personal data are disclosed does not have to be a third party. Therefore, a recipient may be a data controller, joint controller or processor.

According to the WP29 Guidelines on Transparency[41], the actual (named) recipients of the personal data, or the categories of recipients, must be provided. In accordance with the principle of fairness, controllers must provide information on the recipients that is most meaningful for data subjects. In practice, this will generally be the named recipients, so that data subjects know exactly who has their personal data. If controllers opt to provide the categories of recipients, the information should be as specific as possible by indicating the type of recipient (i.e. by reference to the activities it carries out), the industry, sector and sub-sector and the location of the recipients.

Thus, we classified as *fully informative* (1) clauses where (i) the data controller provides information on the actual (named) recipients of the personal data, or (ii) where a data controller opts only to provide the categories of recipients, by indicating (or at least by providing a link to a document listing) the type of recipient (i.e. by reference to the activities it carries out); the industry, sector and sub-sector and the location of the recipients. In any other case then above we classified the clauses as *insufficiently informative* (2). For instance, consider the following examples:

Skyscanner Privacy Policy (last updated on 11 May 2018)
<recep1>We share information relating to our users with selected third parties who provide us with a variety of different services that support the delivery of our services (let's call them "Third Party Processors"). These Third Party Processors range from providers of technical infrastructure to customer service and authentication tools. We require any Third Party Processor which handles information on our behalf to do so pursuant to contractual terms which require that the information is kept secure, is processed in accordance with applicable data protection laws, and used only as we have instructed and not for that Third Party Processor's own purposes (unless you have explicitly consented to them doing so).</recep1>

[41] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018.

Rationale
The above clause is fully informative, since it specifies the categories of recipients, by indicating the type and the sector, i.e. providers of technical infrastructure, customer service and authentication.

Apple Privacy Policy (last updated on 22 May 2018)
<recep2>At times Apple may make certain personal information available to strategic partners that work with Apple to provide products and services, or that help Apple market to customers.</recep2>
Rationale
The above clause fails to be fully informative, since it is unclear the type, the industry, the sector and the location of the mentioned strategic partners.

4.1.8. Period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period

According to Article 13.2(a) and Article 14.2(a) of the GDPR, the controller shall provide the data subject with information on "the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period". This is linked to the data minimisation requirement in Article 5.1(c) and storage limitation requirement in Article 5.1(e).

According to the WP29 Guidelines on transparency, the storage period (or criteria to determine it) may be dictated by factors such as statutory requirements or industry guidelines but should be phrased in a way that allows the data subject to assess, on the basis of his or her own situation, what the retention period will be for specific data/ purposes. "It is not sufficient for the data controller to generically state that personal data will be kept as long as necessary for the legitimate purposes of the processing. Where relevant, the different storage periods should be stipulated for different categories of personal data and/or different processing purposes, including where appropriate, archiving periods"[42].

Thus, we classified as *fully informative* (1) clauses stating that the processor will not retain personal data, or specifying the period for which the personal data will be stored, or – if that is not possible – the criteria used to determine that period, and as *insufficiently informative* (2), clauses generically stating that personal data will be kept as long as necessary for the legitimate purposes of the processing, or phrased in a way that does not allow the data subject to assess, on the basis of his or her own situation, what the retention period will be.

For instance, consider the following examples:

WhatsApp Privacy Policy (last updated on 24 April 2018)
<ret1>We do not retain your messages in the ordinary course of providing our Services to you. Once your messages (including your chats, photos, videos, voice messages, files, and share location information) are delivered, they are deleted from our servers.</ret1>

Rationale

[42] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018 , pp. 33-34.

The above clause is fully informative since it clearly specifies that the data controller does not retain messages, which are deleted right after the delivery.

Facebook Privacy Policy (last updated on 19 April 2018)
<ret2>We store data until it is no longer necessary to provide our services and Facebook Products, or until your account is deleted - whichever comes first. This is a case-by-case determination that depends on things like the nature of the data, why it is collected and processed, and relevant legal or operational retention needs.</ret2>

Rationale
The above clause fails to be fully informative since it does not allow the data subject to assess, at least on the basis of his or her situation, what the retention period will be for specific data/purposes, only mentioning as a general criterion to determine this period the necessity of data for providing services and products.

4.1.9. Existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject and to object to processing as well as the right to data portability

According to Article 13.2(b) and Article 14.2(c) of the GDPR, the controller shall provide the data subject with information on the rights to: access; rectification; erasure; restriction on processing; objection to processing and data portability. According to the WP29 Guidelines on transparency, this information should be specific to the processing scenario and include a summary of what the right involves and how the data subject can take steps to exercise it and any limitations on the right. In particular, "the right to object to processing must be explicitly brought to the data subject's attention at the latest at the time of first communication with the data subject and must be presented clearly and separately from any other information[43].

Thus, we classified as *fully informative* (1), clauses that include a summary of what the rights involve and how the data subject can take steps to exercise them and as *insufficiently informative* (2) clauses falling outside the cases described above. For instance, consider the following examples:

Netflix Privacy Policy (last updated on 11 May 2018)
<correct1>You can request access to your personal information, or correct or update out-of-date or inaccurate personal information we hold about you. You can most easily do this by visiting the "Account" portion of our website, where you have the ability to access and update a broad range of information about your account, including your contact information, your Netflix payment information, and various related information about your account (such as the content you have viewed and rated, and your reviews). You must be signed in to access the "Account" section. You may also request that we delete personal information that we hold about you. To make requests, or if you have any other question regarding our privacy practices, please contact our Data Protection Officer/Privacy Office at privacy@netflix.com.
We respond to all requests we receive from individuals wishing to exercise their data protection rights in accordance with applicable data protection laws.

[43] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018, p. 39. See also Article 21.4 and Recital 70 (which applies in the case of direct marketing), and Guidelines on the right to data portability, WP 242 rev.01, last revised and adopted on 5 April 2017.

Please also see the "Your Choices" section of this Privacy Statement for additional choices regarding your information.</correct1>

Rationale
The above clauses are fully informative since they clearly provides the data subject with information on the rights to access, update, correct and delete personal information hold by the data controller, as well as on the required steps to exercise them.

Airbnb Privacy Policy (last updated on 16 April 2018)
<correct2>You may access and update some of your information through your Account settings. If you have chosen to connect your Airbnb Account to a third-party application, like Facebook or Google, you can change your settings and remove permission for the app by changing your Account settings. You are responsible for keeping your personal information up-to-date.[…]You have the right to ask us to correct inaccurate or incomplete personal information concerning you (and which you cannot update yourself within your Airbnb Account).[…] In some jurisdictions, applicable law may entitle you to request copies of your personal information held by us. You may also be entitled to request copies of personal information that you have provided to us in a structured, commonly used, and machine-readable format and/or request us to transmit this information to another service provider (where technically feasible).</correct2>

Rationale
The above clauses fail to be fully informative since, even though they provide a summary of what the rights involve, given the wording it is unclear under what condition data subjects can exercise their rights and how they can take steps to exercise them.

4.1.10. Right to lodge a complaint with a supervisory authority

According to Article 13.2(b) and Article 14.2(c) of the GDPR, the controller shall provide the data subject with information on the right to lodge a complaint with a supervisory authority. According to the WP29 Guidelines on transparency, this information should explain that, in accordance with Article 77, a data subject has the right to lodge a complaint with a supervisory authority, in particular in the Member State of his or her habitual residence, place of work or place of an alleged infringement of the GDPR[44].

Thus, we classified as *fully informative* (1), clauses where the information explains that the data subject has the right to lodge a complaint with a supervisory authority in the Member State of his or her habitual residence, place of work or place of an alleged infringement of the GDPR, and as *insufficiently informative* (2), clauses that do not specify the information above or state a different criteria to identify the Supervisory authority.

[44] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018 , pp. 39

For instance, consider the following examples:

WhatsApp Privacy Policy (last updated on 24 April 2018)
<complain1>You have the right to lodge a complaint with WhatsApp Ireland's lead supervisory authority, The Irish Data Protection Commissioner, or your local supervisory authority.</complain1>

Rationale
The above clause is fully informative since it provides the data subject with information on the right to lodge a complaint with a supervisory authority, complying with the criteria established by the GDPR.

Twitter Privacy Policy (effective on 25 May 2018)
<complain1>If you wish to raise a concern about our use of your information (and without prejudice to any other rights you may have), you have the right to do so with your local supervisory authority or Twitter International Company's lead supervisory authority, the Irish Data Protection Commission. You can find their contact details here.</complain1>

Rationale
The above clause is fully informative since it provides the data subject with information on the right to lodge a complaint with a supervisory authority, complying with the criteria established by the GDPR.

Steam Privacy Policy (last updated on 25 May 2018)
<complain2>You also have the right to lodge a complaint at a supervisory authority.</complain2>

Rationale
The above clause fails to be fully informative since it does not specify that the mentioned supervisory authority is the one located in the Member State of the data subject habitual residence, place of work or place of an alleged infringement of the GDPR or of the company's main establishment.

Uber Privacy Policy (last updated on 25 May 2018)
<complain2>Users in the EU also have the right to file a complaint relating to Uber's handling of your personal information with the Autoriteit Persoonsgegevens, the Dutch Data Protection Authority.</complain2>

Rationale
The above clause fails to be fully informative since it seems that the only possibility is to complain to the Dutch DPA (Uber main establishment) and it does not mention the possibility to complain to the local DPA of the data subject.

4.1.11. Source from which personal data originate

According to art. 14(2)(f) of the GDPR, if personal data are not coming directly from the data subject the controller shall provide the data subject with information "from which source the personal data originate, and if applicable, whether it came from publicly accessible sources".

However, according to art. 14(5)(b), such obligation shall not apply when "the provision of such information proves impossible or would involve a disproportionate effort". According to the WP29 Guidelines on transparency, "the lifting of the requirement to provide data subjects with information on the source of their personal data applies only where this is not possible because different pieces of personal data relating to the same data subject cannot be attributed to a particular source"[45].

Besides, Recital 61 of the GDPR states that "where the origin of the personal data cannot be provided to the data subject because various sources have been used, general information should be provided". Thus, even if a database has been compiled by a data controller using multiple sources, it is not enough to lift this requirement whenever is possible to identify the source from which the personal data of individual data subjects derived.

Thus, we classified as *fully informative* (1) clauses where the information includes the nature of the sources (i.e. publicly/ privately held sources; the types of organisation/ industry/ sector; and where the information was held (EU or non-EU) etc.), and as insufficiently informative (2) clauses falling outside the cases described above. For instance, consider the following examples:

Airbnb Privacy Policy (last updated on 16 April 2018)
<source1>For Members outside of the United States, to the extent permitted by applicable laws and with your consent where required, Airbnb and Airbnb Payments may obtain the local version of police, background or registered sex offender checks.</source1>

Rationale
The above clause is fully informative since it clearly provides information about the nature of source and from which data originate.

Facebook Privacy Policy (last updated on 19 April 2018)
<source2>Advertisers, app developers, and publishers can send us information through Facebook Business Tools they use, including our social plug-ins (such as the Like button), Facebook Login, our APIs and SDKs, or the Facebook pixel. These partners provide information about your activities off Facebook—including information about your device, websites you visit, purchases you make, the ads you see, and how you use their services—whether or not you have a Facebook account or are logged into Facebook.</source2>

Rationale
The above clause fails to be fully informative since it is unclear the nature of sources from which personal data originate.

Airbnb Privacy Policy (last updated on 16 April 2018)
<source2>Airbnb and Airbnb Payments may collect information, including personal information, that others provide about you when they use the Airbnb Platform and the Payment Services, or obtain information from other sources and combine that with information we collect through the Airbnb Platform and the Payment Services.</source2>

Rationale

[45] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018 p. 29

The above clause fails to be fully informative since the nature of some of the mentioned sources remains unclear, i.e. what the "other sources" entails, as well as how their nature could be identified.

4.1.12. Existence of automated decision-making, including profiling

According to Article 13.2(f) and Article 14.2(g) of the GDPR, the controller shall provide the data subject with information on "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject".

WP29 has produced guidelines on automated individual decision-making and profiling[46] which should be referred to for further guidance on how transparency should be given effect in the particular circumstances of profiling. It should be noted that, aside from the specific transparency requirements applicable to automated decision-making under Articles 13.2(f) and 14.2(g), the comments in these guidelines relating to the importance of informing data subjects as to the consequences of processing of their personal data, and the general principle that data subjects should not be taken by surprise by the processing of their personal data, equally apply to profiling generally (not just profiling which is captured by Article 22[47]), as a type of processing[48].

Thus, we classified clauses as a *fully informative* (1), clauses that provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject; and they inform the data subject about the right not to be subject to a decision based solely on automated decision making, including profiling; and that the decisions referred are not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place. We classified as *insufficiently informative* (2) clauses falling outside the cases described above. For instance, consider the following examples:

Uber Privacy Policy (last updated on 25 May 2018)
<auto2>In certain cases such incidents may lead to deactivation by means of an automated decision making process.</auto2>
<auto1>Users in the EU have the right to object to this type of processing; see Section II.I.1.d for more information</auto1>

Rationale
The first clause above fails to be fully informative since it does not provide neither meaningful information about the logic involved in the automated decision making process, nor the significance and the envisaged consequences of such processing for the data subject.
Conversely, the second clause above is fully informative since it clearly specifies that provides the data subject has the right to object to this type of processing, and it is immediately apparent where and how this information can be accessed, i.e. the specified number of section in the privacy policy.

[46] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP 251
[47] This applies to decision-making based solely on automated processing, including profiling, which produces legal effects concerning the data subject or similarly significantly affects him or her.
[48] Recital 60, which is relevant here, states that "Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling".

Apple Privacy Policy (last updated on 22 May 2018)
<auto2>Apple does not take any decisions involving the use of algorithms or profiling that significantly affects you.</auto2>

Rationale
The above clause fails to be fully informative since it is unclear whether or not there is an automated decision making. The clause could be interpreted to imply that there is an automated decision making that however does not significantly affect the data subject. It should be also noted that, under this interpretation, also the significance and the envisaged consequences of the processing for the data subject are unclear. Such evaluation seems to remain under the sole discretional assessment of the data controller, since the criteria on which it is based are not provided.

4.2. B. Substantive Compliance

The dimension of substantive compliance concerns whether the types of processing stipulated in a privacy policy are themselves GDPR compliant. Many of the business practices currently in place have been created without taking the European personal data protection rules into account. Hence, many privacy policies currently contain "problematic" clauses, signalling that the data controller behaves (might behave) in a way that is not GDPR compliant (based on, among others, art. 5, 6 and 9 of the GDPR).

In analysing privacy policies, we identified 10 categories of clauses and for each category, we defined a corresponding XML tag as shown in table 1. We also assumed that each category could be classified either as a *fair processing* clause; a *problematic processing* clause; and as an *unfair processing* clause. To this end, we appended a numeric value to each XML tag, with 1 meaning fair; 2 problematic; and 3 unfair.

As noted above, a single clause in some cases may fall in different categories and consequently may have multiple taggings. For sake of clarity, each category is discussed in the following showing just the specific tag of the discussed category, even if others might be associated to the same clause.

Type of clause	Symbol
Processing of special categories of personal data (health, sex life, political opinions etc.)	<sens>
Consent by using	<cuse>
Take or leave it approach	<tol>
Third party data transfers	<tp>
Policy Change	<pch>
Transfer of data to third countries	<cross>
Processing of children's data	<child>
Licensing data	<lic>
Advertising	<ad>
Any other type of consent	<c>
Any other type of clause we find outstandingly problematic	<out>

Table 2

4.2.1. Processing of special categories of personal data

According to art 9 of the GDPR "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited." However, article 9(2) also provides several exceptions to the general prohibition to process special categories of data. In this section we analyse different ways in which a clause stating that special categories of data can be processed under the privacy policy may fail to meet the golden standard.

In particular, we classified as fair processing clauses (1) the clauses that are compliant at least with one of the conditions specified in art. 9 GDPR. For example, clauses stating that the explicit consent of the data subject is required for one or more specific purposes, where the purposes are completely specified. We classified as problematic processing clauses (2) the clauses claiming to fall under one of the exceptions ex art. 9(2) of the GDPR, but the information is incomplete and not exhaustive. For instance, clauses stating that the explicit consent of the data subject is required, but the specific purpose is not specified. We classified as *unfair processing* clauses (3) the clauses falling outside the cases described above. An example of this type of clause is the following:

Facebook Privacy Policy (last updated on 19 April 2018)

<sens2>To create personalized Products that are unique and relevant to you, we use your connections, preferences, interests and activities based on the data we collect and learn from you and others (including any data with special protections you choose to provide where you have given your explicit consent); how you use and interact with our Products; and the people, places, or things you're connected to and interested in on and off our Products.</sens2>

Rationale

The clause above is a ***problematic processing*** clause, even though the processing of sensitive categories seems to require the explicit consent of the data subject. In particular, the clause does not provide sufficient and exhaustive information about the purpose of the processing since it is unclear what kind of personalisation the clause refers to; the type of analysis the controller is going to undertake as well as what consequences such processing entails.

4.2.2. Consent by using

According to the GDPR, consent should be given "by a statement or by a clear affirmative action" (art 4(11)); "This could include ticking a box when visiting an internet website" (Recital 32). The "consent by using" clauses states that the user is bound by the privacy policy of a specific service, simply by using the service. Thus, we classified the clauses either as *fair processing* clauses (1), when the consent is explicitly required, or as unfair processing clauses (3), clauses stating that by simply using the service, the user consents to the terms of the privacy policy. For instance, consider the following examples:

Airbnb Privacy Policy (last updated on 16 April 2018)
<cuse1>This additional information will be processed based on your consent.</cuse1>

Rationale
The clause above is a fair processing clause since the consent of the data subject is explicitly required.

Epic games Privacy Policy (last updated on 24 May 2018)
<cuse3> when you use our websites, games, game engines, and applications, you agree to our collection, use, disclosure, and transfer of information as described in this policy, so please review it carefully.</cuse3>

Rationale
The clause above is an unfair processing clause since it states that the data subject consents to the collection, use, disclosure and transfer of his/her information, and thus s/he is bound by the privacy policy, simply by using the Epic Games web-sites, games, game engines and applications.

4.2.3. Take or leave it approach

According to the GDPR, "When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract." (art 7(4)). The Article 29 Working Party Guidelines on consent[49] also state that "the situation of "bundling" consent with acceptance of terms or conditions, or "tying" the provision of a contract or a service to a request for consent to process personal data that are not necessary for the performance of that contract or service, is considered highly undesirable." (section 3.1.2, page 8).

To state once again: this normative threshold is used to explain why the "take it or leave or" approach should be assessed negatively according to the GDPR. This does not change the fact that, based on art 4(11) and 7(2), any consent "hidden" in the privacy policy will be *per se* invalid.

We always considered this category as *problematic processing* clauses (2). For instance, consider the following examples:

Booking Privacy Policy (last updated on 9 May 2018)
<tol2>Sad but necessary bit: If you disagree with this Privacy Statement, you should discontinue using our services.</tol2>

Rationale
The clause above is a problematic processing clause since it clearly states that in case of disagreement the data subject should discontinue using the service. Since the clause forces the data subject to accept the privacy policy, it would seem that all processing operations are based on forced consent.

[49] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018.

Netflix Privacy Policy (last updated on 11 May 2018)
<tol2>If you do not wish to acknowledge or accept any updates to this Privacy Statement, you may cancel your use of the Netflix service.</tol2>

Rationale
The clause above is a problematic processing clause since it clearly states that in case of disagreement the data subject should discontinue using the service. Since the clause forces the data subject to accept the privacy policy, it would seem that all processing operations are based on forced consent.

4.2.4. Third party data transfers

In practice, privacy policies will often include clauses stipulating that more parties than just the data subject and data controller (or processor) will be involved in the processing relation(s). We used this tag predominantly when transfers of data by data controllers to third parties are foreseen, but also for situations when data about the data subject is collected from third parties; or data about third parties collected from the data subject (for example contacts from smartphones). The legal requirements for lawfulness of third-party data transfers are robust, and their detailed representation exceeds the possibilities of a single tagging system. Hence we treat this category as (to a certain extent) an umbrella one, bearing in mind that the purpose of research is to teach the machine to highlight clauses which a human lawyer must pay attention to, and not yet to fully assess their legality. Given this, certain degree of simplification, was necessary for the technical reasons.

We classified this type of clause as a *fair processing* clause (1) when all the following conditions are met: the third parties are specified; the scope of the use is specified; the consent is optional or the transfer is necessary for the performance of the contract or it is required by law, or it is based on a legitimate interest and the specific interest in specified.

We classified it as *problematic processing* clause (2) when one of the following applies: the scope is not specified, and the consent is optional; or the scope is specified but the consent is not optional or necessary to access the service; or the third parties are not specified.

We classified it as *unfair processing* clause (3) when all the following applies: the object of transfer and the scope are not specified, the consent is necessary to access the service (i.e. not optional), and the transfer is not necessary for the performance of the contract. For instance, consider the following examples:

Apple Privacy Policy (last updated on 22 May 2018)
<tp1>It may be necessary – by law, legal process, litigation, and/or requests from public and governmental authorities within or outside your country of residence – for Apple to disclose your personal information.tp1>

Rationale
The clause above is a fair processing clause since the scope as well as the third parties are specified (i.e. public and governmental authorities) or at least easily inferable case by case from the fully mentioned purposes (i.e. necessary by law, legal process, litigation).

Amazon Privacy Policy (last updated on 22 May 2018)
<tp2>You can tell when a third party is involved in your transactions, and we share customer information related to those transactions with that third party.</tp2>

Rationale
The clause above is a problematic processing clause since, even though it specifies the object of the transfer, i.e. customer information, it is not clearly specified the scope of transfer or at least whether it is necessary for the performance of the contract, or based on the legitimate interest of the data controller and/or of the third parties and in this case what is the specific interest.

Apple Privacy Policy (last updated on 22 May 2018)
<tp3>We may process your personal information for the purposes described in this Privacy Policy with your consent, for compliance with a legal obligation to which Apple is subject or when we have assessed it is necessary for the purposes of the legitimate interests pursued by Apple or a third party to whom it may be necessary to disclose information.</tp3>

Rationale
The clause above is an unfair processing clause since it does not provide any useful information to identify neither the third parties, nor the object of transfer, and the consent is not optional, the specific legitimate interest is not identified.

4.2.5. Policy Change

According to WP29 Guidelines on transparency, being accountable as regards transparency pertains to the entire processing life cycle, including the change of contents of existing privacy statements/ notices. Thus, the controller should adhere to the same principles when communicating both the initial privacy statement/ notice and any subsequent substantive or material changes to this statement/ notice.

Factors which controllers should consider in assessing what is a substantive or material change include the impact on data subjects, and how unexpected/ surprising the change would be to data subjects. Changes to a privacy statement/ notice that should always be communicated to data subjects include *inter alia*: a change in processing purpose; a change to the identity of the controller; or a change as to how data subjects can exercise their rights in relation to the processing.

The controller should take all measures necessary to ensure that these changes are communicated in such a way that ensures that most recipients will actually notice them. Appropriate modalities of notifications include email, hard copy letter, pop-up on a webpage or other modalities which will effectively bring the changes to the attention of the data subject specifically devoted to those changes. References in the privacy statement/ notice to the effect that the data subject should regularly check the privacy statement/notice for changes or updates are considered not only insufficient but also unfair in the context of Article 5.1(a) GDPR[50]. Moreover, we identified two possible ways in order to ensure that most recipients notice these substantive or material changes, namely a

[50] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018, p.16 ff

new consent request, whenever it constitutes the legal basis for the processing, and/or a reading confirmation.

Thus, we classified this type of clause as a *fair processing* clause (1) when it states that a new notice is provided and a new consent (whenever it constitutes the legal basis for the processing), or at least a confirmation of reading is requested; as a *problematic processing* clause (2) when a new notice is provided but a new consent or a confirmation of reading is not requested; and as an *unfair processing* clause (3) when a new notice is not necessarily provided and a new consent or a confirmation of reading is not requested, for example, clauses stating that it is a responsibility of the data subject to check for the last updated version of the privacy policy. We only marked found two instances of what could be, in itself, marked as fair policy change clause; however, since they were problematic for other reasons (take it or leave it approach), we do not list them here. For others, consider the following examples:

Twitter Privacy Policy (effective on 25 May 2018)

<pch2>We may revise this Privacy Policy from time to time. The most current version of the policy will govern our processing of your personal data and will always be at https://twitter.com/privacy. If we make a change to this policy that, in our sole discretion, is material, we will notify you via an @Twitter update or email to the email address associated with your account.</pch2>

Rationale

The clause above is a problematic processing clause since even though a new notice will be provided a new consent or at least a reading confirmation will not be requested.

Booking Privacy Policy (last updated on 9 May 2018)

<pch3>We might amend the Privacy Statement from time to time. If you care about your privacy, visit this page regularly and you'll know exactly where you stand.</pch3>

Rationale

The clause above is an unfair processing clause since it states that the data subject is required to regularly visit the service web-page, implying that a new notice will not be provided, and a new consent or at least a reading confirmation is not required.

4.2.6. Transfer to a Third Country

Article 13(1)(f) and article 14(1)(f) require the controller to inform the data subject about (i) whether he/she intends to transfer personal data to third countries and (ii) the existence or absence of adequacy decision by the Commission or appropriate safeguards.

Besides, Chapter V (Articles 44 through 49) of the GDPR governs transfer of personal data to third countries. Article 45 states the conditions for transfers with an adequacy decision; Article 46 sets forth the conditions for transfers by way of appropriate safeguards in the absence of an adequacy decision; Article 47 sets the conditions for transfers by way of binding corporate rules; Article 48 addresses situations in which a foreign tribunal or administrative body has ordered transfer not otherwise permitted by the GDPR; and Article 49 states the conditions for derogations for specific situations in the absence of an adequacy decision or appropriate safeguards.

We classified this type of clause as a fair processing clause (1) when it provides complete information (i.e. explicitly referring to a code of conduct, or approved certification mechanism) about the transfer under at least one of the following transfer mechanism: (i) Adequacy Decision: Transfer to a country (or sector within a country) on the EC's approved list ; (ii) Binding Corporate Rules (BCRs); (iii)Model Contractual Clauses (Model Clauses), including EC-approved Model Clauses in contracts for the transfer of data; (iv) EU-US and Swiss-US Privacy Shield; (v) Explicit Consent from individual; (vi) Derogations for specific situations as specified ex art. 49 GDPR.

We classified it as a problematic processing clause (2) when it only provides the list of the transfer mechanism above and does not provide any specific information allowing the data subject to be effectively informed. We classified it as an *unfair processing* clause (3) when the transfer mechanism-requirements are not provided or specified.

For instance, consider the following examples:

Steam Privacy Policy (last updated on 25 May 2018)
<cross1>Valve complies with the EU-U.S. Privacy Shield Framework and the Swiss-U.S. Privacy Shield Framework as set forth by the U.S. Department of Commerce regarding the collection, use, and retention of personal information transferred from the European Union and Switzerland to the United States. Valve has certified to the Department of Commerce that it adheres to the Privacy Shield Principles. If there is any conflict between the terms in this privacy policy and the Privacy Shield Principles, the Privacy Shield Principles shall govern. To learn more about the Privacy Shield program, and to view our certification, please visit http://www.privacyshield.gov </cross1>

Rationale
The clause above is a fair processing clause since it clarifies that the personal data transfer to third countries is grounded on the EU-U.S. Privacy Shield Framework and it also provides a link to get more information about the Privacy shield program and the data controller certification.

Apple Privacy Policy (last updated on 22 May 2018)
<cross2>Apple uses approved Model Contractual Clauses for the international transfer of personal information collected in the European Economic Area and Switzerland.</cross2>

Rationale
The clause above is a problematic processing clause since it only provides the transfer mechanism but does not allow the data subject to be effectively informed, for example linking to the approved sets of standard contractual clauses for data transfers.

Twitter Privacy Policy (effective on 25 May 2018)
<cross3>Where the laws of your country allow you to do so, you authorize us to transfer, store, and use your data in the United States, Ireland, and any other country where we operate.</cross3> <cross3>In some of the countries to which we transfer personal data, the privacy and data protection laws and rules regarding when government authorities may access data may vary from those of your country.</cross3>

Rationale
The clauses above are *unfair processing* clauses since they do not provide any type of transfer mechanism and do not guarantee any adequate standard with regard to the privacy of the data subject and lawfulness of the processing.

4.2.7. Processing of children's data

According to art. 8(1) of GDPR (Conditions applicable to child's consent in relation to information society services), "the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child". Moreover, according to art. 8(1) Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.

Regarding the authorization, since the GDPR does not specify practical ways, the WP29 recommends the adoption of a proportionate approach in line with Article 8(2) GDPR and Article 5(1)(c) GDPR (data minimisation)[51]. For example, a proportionate approach may be to focus on obtaining a limited amount of information, such as contact details of a parent or guardian.

What is reasonable may depend up on the risk inherent on the processing, as well as the available technology. In low-risk cases, verification of parental responsibility via email may be sufficient. Conversely, in high-risk cases, it may be appropriate to ask for more proof, so that the controller is able to verify and retain the information pursuant to Article 7(1) GDPR. It should be also noted that the GDPR does not provide a definition of child. In this context we intend as a child every human being below the age of 18.

We classified as *fair processing* clauses (1), clauses stating that the processing concerns (i) a child which is at least 16 years old or 13 years old if a national legislation lowers the threshold down to 13 years and that has given consent to the processing for one or more specific purposes, or (ii) a child which is below the age of 16 years and where consent is given or authorised by the holder of parental responsibility; and where (iii) the controller declares that reasonable efforts will be made to verify in such cases that consent is given or authorised by the holder of parental responsibility over the child, taking into consideration available technology.

We classified as *problematic processing* clauses (2) clauses where the first two conditions above are met but the types of efforts to verify that the consent is given or authorised by the holder of parental responsibility over the child are not mentioned and specified. We classified as *unfair processing* clauses (3) any other case than above. For instance, consider the following examples:

Apple Privacy Policy (last updated on 22 May 2018)
<child1>Children under the age of 13, or equivalent minimum age in the relevant jurisdiction, are not permitted to create their own Apple IDs, unless their parent provided verifiable consent or as part of the child account creation process in Family Sharing or they have obtained a Managed Apple ID account (where available) through their school.</child1> <child1>For example, a parent must review the Apple ID and Family Sharing Disclosure and agree to the Consent

[51] Article 29 Working Party Guidelines on consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018.

to Apple's Collection, Use and Disclosure of Your Child's Information; and the iTunes Store Terms and Conditions, before they can begin the Apple ID account creation process for their child.</child1>

Rationale
The clause above is a fair processing clause since it clearly specifies that the processing does not concerns children that are under the age of 13 years or equivalent minimum age in the relevant jurisdiction, unless their parent provide a verifiable consent or authorisation through the means mentioned above.

Twitter Privacy Policy (effective on 25 May 2018)
<child2>Our services are not directed to children, and you may not use our services if you are under the age of 13. You must also be old enough to consent to the processing of your personal data in your country (in some countries we may allow your parent or guardian to do so on your behalf). You must be at least 16 years of age to use Periscope.</child2>

Rationale
The clause above is a problematic processing clause since it generically refers to "some countries" and to the mere possibility that the processing will be authorised and consent given by the holder of parental responsibility for child under the age required by the law. Besides, it does not specify whether or not reasonable efforts will be made to verify in such cases that consent is given or authorised by the holder of parental responsibility over the child.

Epic Games Privacy Policy (last updated on 24 May 2018)
<child3>If you are under the age of 18 and have an account for our websites, mobile apps, game engines, games, or other online services, you may request that we remove certain content that you provided, such as deleting or editing comments you have posted.</child3>

Rationale
The clause above is an *unfair processing* clause since it is possible to infer that the data controller processes data of child and it does not specify any requirement with regard to the minimum age for the processing, whether the parental consent and authorisation is required, and whether the data controller makes any efforts in order to verify the requirements above.

4.2.8. Advertising

This category concerns clauses stating that personal data will be used for advertising communications (including direct marketing and behavioural/targeted advertising).

Please note: Advertising and direct marketing are closely related, in fact direct marketing is a form of advertising, but not the same, nor subject to exactly the same legal requirements. Direct marketing by email for example will typically require consent on the basis of the ePrivacy Directive, while the GDPR generally considers direct marketing to be a legitimate interest.

Regarding online behavioural advertising (OBA), though industry has pushed for grounding this practice on legitimate interest and the right to object (i.e. opt-out approach), the WP29 in its guidelines on OBA has already stated that consent is the right legal basis for this practice (i.e. opt-in approach). Also, the ePrivacy Directive would typically require consent for the placing of the cookies

that are used for OBA. Moreover the balancing exercise that the use of legitimate interest under the GDPR would require could not favor OBA.

Art. 21(2), art 21(3) GDPR, and Recital 70 of the GDPR deal with the right to object when personal data are processed for direct marketing purposes. Besides, also The Article 29 Working Party Guidelines on Automated individual decision-making and Profiling[52] deal with the matter in relation with Art 5(1) (b) of the GDPR (Further processing and purpose limitation), identifying scenarios where profiling for marketing purposes can involve the use of personal data that was originally collected for something else. Whether this additional processing is compatible for the original purposes for which the data were collected will depend up a range of factors - highlighted in the Article 29 Data Protection Working Party Opinion 03/2013 on purpose limitation[53] - including what information the controller initially provided to the data subject (art. 6(4) GDPR).

Default settings with regard to behavioural profiling and advertising (essentially "opt-out") remain problematic, as noted above. According to the Article 29 Working Party on Online Behavioural Advertising[54], consent is required for behavioural advertising and such consent cannot be inferred from the data subject's inaction, thus inactivity cannot be considered consent.

Having said this, for the purposes of this particular study – which is to train the machine to automatically highlight clauses that a human lawyer should pay special attention to – we treated both as one category, without explicitly distinguishing between advertising and direct marketing for the tagging purposes. Being aware of the fact that this might, at the later stage of research, be used for further refining – already quite detailed – instruction, we believe that at this point such an approach is a sound one, particularly given the need to enable the classifier to mark the clauses as *potentially* problematic.

On these grounds, we classified as fair processing (1) clauses where the consent is requested in line with GDPR requirements, and the opt-out is possible, or where it is stated that data collected are not used for marketing purposes. We classified as problematic (2) clauses where consent is not required, but the opt-out is possible. We classified as unfair processing clauses (3) clauses where the consent is not requested, and the opt-out is not possible. For instance, consider the following examples:

Steam Privacy Policy (last updated on 25 may 2018)
<ad1>Subject to your separate consent or where explicitly permitted under applicable laws on email marketing, Valve may send you marketing messages about products and services offered by Valve to your email address.</ad1>

Rationale
The clause above is a ***fair processing*** clause. It pertains direct marketing by email and clearly states that the data subject separate consent is required except where direct marketing by email is explicitly permitted under applicable laws.

[52] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01) Adopted on 3 October 2017 As last Revised and Adopted on 6 February 2018
[53] Article 29 Data Protection Working Party Opinion 03/2013 on purpose limitation, 2 April 2013.
[54] See Article 29 Data Protection Working Party Opinion 2/2010 on Online Behavioural Advertising, 24 June 2010

Airbnb Privacy Policy (last updated on 16 April 2018)
<ad2>You can opt-out of receiving marketing communications from us by following the unsubscribe instructions included in our marketing communications or changing your notification settings within your Airbnb Account.</ad2>

Rationale
The clauses above is a ***problematic processing*** clause since it only allows the data subject to opt out, specifying the needed steps.

Facebook Privacy Policy (last updated on 19 April 2018)
<ad3>Ads and other sponsored content: We use the information we have about you-including information about your interests, actions and connections-to select and personalize ads, offers and other sponsored content that we show you.</ad3>

Rationale
The clauses above is an ***unfair processing*** clause since it refers to targeting advertising on online platform and it clarifies that the data controller processes data for targeted advertising but the data subject's consent is not requested and the opt-out is not possible.

4.2.9. Any other type of consent

Since consent cannot be given through the general acceptance of a privacy policy or terms of use (art. 4(11) and 7(2) of the GDPR), we used this generic category to highlight consents "hidden" in those documents. Given the very generic type of category, we classified as *problematic processing* clauses (2) all clauses of this type. For instance, consider the following examples:

Twitter Privacy Policy (effective on 25 May 2018)
<c2>If you provide us with your phone number, you agree to receive text messages from Twitter to that number as your country's laws allow.</c2>

Airbnb Privacy Policy (last updated on 16 April 2018)
<c2>In jurisdictions where Airbnb facilitates the Collection and Remittance of Occupancy Taxes as described in the "Taxes" section of the Terms of Service, Hosts and Guests, where legally permissible according to applicable law, expressly grant us permission, without further notice, to disclose Hosts' and Guests' data and other information relating to them or to their transactions, bookings, Accommodations and Occupancy Taxes to the relevant tax authority, including, but not limited to, the Host's or Guest's name, Listing addresses, transaction dates and amounts, tax identification number(s), the amount of taxes received (or due) by Hosts from Guests, and contact information.</c2>

4.2.10. Any other type of clause we find "outstandingly problematic"

This category deals with clauses stating anything else that does not fall into the scope of the abovementioned clauses, and yet can be considered as problematic (2) for different reasons (including the general lawyerly intuition). For example, clauses that are obscure in their meaning, or that claim the responsibility of the data subject regarding the processing of third parties' data by the data controller. For instance, consider the following examples:

Apple Privacy Policy (last updated on 22 May 2018)

<out2>You are not required to provide the personal information that we have requested, but, if you chose not to do so, in many cases we will not be able to provide you with our products or services or respond to any queries you may have.[…] Because this information is important to your interaction with Apple, you may not opt out of receiving these communications.</out2>

Rationale

The clause above is a ***problematic processing*** clause since it is obscure in its meaning. In particular, the clause states that if the data subject choses to not provide the personal data requested by the data controller, in many cases the controller will not be able to provide products or services or respond to queries, but does not specify what are the information or at least the category of information that are necessary to provide the service. At the same time, the clause also states that because this information is important the data subject cannot opt-out.

Booking Privacy Policy (last updated on 9 May 2018)

<out2>However, at this point we have to point out that it's your responsibility to ensure that the person or people you have provided personal data about are aware that you've done so, and have understood and accepted how Booking.com uses their information (as described in this Privacy Statement).</out2>

Rationale

The clauses above is a ***problematic processing*** clause since it states that the data subject is responsible for ensuring that person or people he/she has provided personal data about have understood and accepted how Booking.com uses their information. The ways in which a consumer would be supposed to do that remain unclear.

4.3. C. Clarity of Expression

The dimension of clarity of expression concerns whether a privacy policy is framed in an understandable and precise language.

Art. 5(1)(a) of the GDPR requires that personal data must be processed lawfully, fairly and in a transparent manner. Further, art. 12(1) of the GDPR requires that information provided to the data subjects must be provided in concise, transparent intelligible and easily accessible form, using clear and plain language. Read jointly, these two provisions impose an obligation on data controllers to provide (a) all the information about processing (transparency: types of data, legal basis, purpose etc.), (b) in a way that enables one to fully comprehend this information. A *contrario*, presenting information that is unclear, vague or non-transparent constitutes a breach of the GDPR.

Besides, Recital 42 of the GDPR states that "In accordance with Council Directive 93/13/EEC[55] a declaration of consent pre-formulated by the controller should be provided in an intelligible and easily accessible form, using clear and plain language and it should not contain unfair terms".

According to WP29 Guidelines on transparency[56], the requirement for clear and plain language means that information should be provided in as simple a manner as possible, avoiding complex

[55] Article 5 of Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts

[56] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018 pp. 9-10

sentence and language structures. The information should not be phrased in abstract or ambivalent terms or leave room for different interpretations. With regard to the plain and intelligible language requirement, we identified two types of clauses, namely clauses expressed in a clear language (not tagged) and clauses expressed in an unclear language, for which we defined the following XML tag: <vag>.

WP29 Guidelines on transparency also lists the following examples as poor practices[57]:

"We may use your personal data to develop new services" (as it is unclear what the "services" are or how the data will help develop them); "We may use your personal data for research purposes" (as it is unclear what kind of "research" this refers to); and "We may use your personal data to offer personalized services" (as it is unclear what the "personalization" entails). It also suggests that language qualifiers such as "may", "might", "some", "often", and "possible" should be avoided.

Besides, with regard to vagueness, we considered as vague language qualifiers also "including" and "such as" when they are present within a list, for example of the category of data collected. In this case, it is possible to infer that the categories are not comprehensively specified. This language qualifiers can be present in clauses that fail to reach the golden standard also along the dimensions of both comprehensiveness and substantive compliance. We have tagged as possibly vague all those clauses where it is specified that a certain process will possibly take place, using qualifiers such as may, might, can, possibly, etc. This may appear questionable, but we have preferred to be on the safe side, considering that in general it should be possible for the controller to state under what conditions the processes will or will not take place. For instance, consider the following examples:

Apple Privacy Policy (last updated on 22 May 2018)

<vag>Apple and its affiliates may share this personal information with each other and use it consistent with this Privacy Policy. They may also combine it with other information to provide and improve our products, services, content, and advertising.</vag>

Rationale

The clause fails along the dimension of clarity since it is unclear whether the data controller and its affiliates (i) share with each other the data subject personal information and in case of affirmative answer what information and under what conditions; (ii) combine such information with other unspecified information and under what conditions.

Amazon Privacy Policy (last updated on 22 May 2018)

<vag> As a result of those actions, you might supply us with such information as: your name; address and phone number; payment information; your age; your location information; people to whom purchases have been dispatched or people listed in 1-Click settings (including addresses and phone numbers); e-mail addresses of your friends and other people; content of reviews and e-mails to us; personal description and photograph in Your Profile; voice recordings when you speak to Alexa; images and video stored in connection with Amazon Services, information and documents regarding identity and standing; corporate and financial information; credit history information; VAT numbers; and device log

[57] Article 29 Data Protection Working Party, WP260 Guidelines on transparency under Regulation 2016/679, as last revised and adopted on 11 April 2018, p. 9

files and configurations, including Wi-Fi credentials, if you choose to automatically synchronise them with your other Amazon devices.</vag>

Rationale
The clause fails along the dimension of clarity since it is unclear (i) whether the data subject, as a result of the mentioned actions, provides the specified information and, in case of affirmative answer, (ii) under what conditions.

Twitter Privacy Policy (effective on 25 May 2018)
<vag>Subject to your settings, we may collect, use, and store additional information about your location - such as your current precise position or places where you've previously used Twitter - to operate or personalize our services including with more relevant content like local trends, stories, ads, and suggestions for people to follow.</vag>

Rationale
The clause fails along the dimension of clarity since it is unclear (i) whether the data controller collects use and store the information mentioned above and in case of affirmative answer, (ii) under what conditions; and (iii)what are the purposes as well as the legal basis for the processing.

5. Object of Inquiry

The normative threshold presented in the previous section has been used to analyse and tag the privacy polices of fourteen data controllers (online platforms and services): Google, Facebook (and Instagram, sharing the same policy), Amazon, Apple, Microsoft, WhatsApp, Twitter, Uber, AirBnB, Booking, Skyscanner, Netflix, Steam, Epic Games (the developer of Fortnite). These services have been chosen as they are some of the most used services in different areas: the online "giants": "GAFAM"; social networks and messaging (Facebook, Instagram, Twitter, WhatsApp); travel (Booking.com and Skyscanner); "collaborative economy" transport (Uber) and accommodation services (AirBnB); and entertainment: Netflix for films and series, Steam as the most popular game distribution platform, and Epic Games, the developer of one of the most popular games at the moment (Fortnite). All these platforms collect enormous amounts of data about users, and as the big players, with essentially unlimited resources for legal counsel, should set a good example on how a proper privacy policy should look like. This is important both for the sake of the data they process themselves, and for the larger societal context, when we can easily assume that other, smaller players will look up to their privacy policies in order to see "how this is done".

In this section we provide an overview of the results of the analysis, both quantitative and qualitative. Regarding qualitative analysis, we did not aim here at providing a comprehensive legal analysis of each and every privacy policy of each and every service – i.e. we are not moving "down" from the potentially problematic level – since the scope of such an exercise would vastly exceed the size of this study. The main purpose of the data generated is to train the machine to automatically evaluate privacy policies for the presence of required information clauses, presence of unclear language clauses, and potentially "problematic" processing clauses.

Given this, apart from quantitative data, we simply provide some examples and/or reflections on some aspects of each policy. Readers can review full privacy policies, annotated in accordance with the instruction presented in the previous section, here: http://www.claudette.eu/gdpr/.

The analysed corpus consisted of 14 documents, comprising 3658 sentences (80.398 words) in total. This is a size of regular monograph, comparable to a small PhD thesis. Out of this material, 401 sentences (11.0%) were marked as containing unclear language, and 1240 (33.9%) as containing a "problematic clause", i.e. either a "problematic" processing clause, or an insufficient information clause (under articles 13 and 14), that is a clause that a lawyer should pay attention to while analysing the content of privacy policies.

None of the analysed documents contained all the required information clauses. None of the analysed documents was free of potentially problematic processing clauses. None of the analysed documents could be told to completely fulfil the requirements of the GDPR.

Important caveat: a perfectly designed privacy policy would not necessarily need to have 0% unclear language clauses and or insufficient information clauses. On the contrary, one can easily imagine a nicely designed layered privacy policy, where the highest (shortest and clearest) level includes sentences speaking of examples only, or containing expressions with language that by itself could be judged unclear, in order to facilitate understanding, while at the same time provide all the comprehensive information about types of data, purposes of processing, specifically named third parties, etc. on the lower layers of the document. Such a policy would result in a report generated by a

machine that would alert the human lawyer, but at the same time make him or her realize that each section is followed by a comprehensive counterpart. Unfortunately, none of the analysed policies had such a quality.

Regarding policy-specific observations:

AirBnB: The 'legitimate interest' legal basis is used 11 times in the AirBnB privacy policy. The way the concept of 'legitimate interests' is used here is very blurry and unclear; to the lay reader it might seem like a basis that can be used for anything, for example: "These processing activities are based on our legitimate interest in undertaking marketing activities to offer you products or services that may be of your interest."

AirBnB also mentions 'some jurisdictions' three times in its privacy policy, without specifying which jurisdictions: "Additionally, in some jurisdictions, applicable law may give you the right to limit the ways in which we use your personal information [...]".

The privacy statement of AirBnB includes a problematic clause about sharing personal data with social media platforms: "The social media platforms with which we may share your personal data are not controlled or supervised by Airbnb. Therefore, any questions regarding how your social media platform service provider processes your personal data should be directed to such provider." This clause is problematic for two reasons. First, AirBnB does not specificy with which social media platforms they share the personal data. In an earlier paragraph it mentions "social media platforms, such as Facebook or Google", but this is explicitly non-exhaustive. Second, instead of directly clarifying the use of these data, AirBnB states that users should contact these unspecified social media platforms with any questions. How can users do that if they do not even know which platforms they are referring to?

Amazon: At the beginning of its privacy notice, Amazon underlines that customers can decide not to disclose certain personal information. Yet, it follows up on this with a 'vague threat': "You can choose not to provide certain information, but then you might not be able to take advantage of many of our Amazon Features". This 'threat' is vague, because it is unclear which features are affected in which way.

Further, Amazon states: "Our business changes constantly and our Privacy Notice will change also. You should check our website frequently to see recent changes." This statement, on the one hand, shifts the burden of figuring out whether the information is still valid on the consumer; on the other does not mention the right to object/withdraw consent if certain circumstances occur, suggesting that a consumer might be "taken by surprise" when a change occurs, and not even know that his or her data is now being processed differently.

Apple: Apple seems discontent with the inclusive definition of personal data employed by the GDPR, and decided to authoritatively state what data it shall consider non-personal. The company states, rather broadly: "The following are some examples of non-personal information that we collect and how we may use it: We may collect information such as occupation, language, zip code, area code, unique device identifier, referrer URL, location, and the time zone where an Apple product is used so that we

can better understand customer behavior and improve our products, services, and advertising." One could assume that if location, occupation and unique device identifier are combined, these could actually be directly associated with an individual.

The privacy statement of Apple contains a particularly unclear segment about the use of datasets containing images and voices: "For research and development purposes, we may use datasets such as those that contain images, voices or other data that could be associated with an identifiable person. When acquiring such datasets, we do so in accordance with applicable law in the jurisdiction in which the dataset is hosted. When using such datasets for research and development, we do not attempt to re-identify individuals who may appear therein." This fragment is unclear in many regards. It does not specify where the datasets come from, what 'other data' they contain or what the research and development purposes entail.

Booking: Booking tries to make its privacy statement accessible by using a video and popular language: "This document describes how we use and process your personal data, hopefully provided in a readable and transparent manner so you can get where we're coming from without getting bored senseless." The following 'take it or leave it' clause also exemplifies this popular language: "Sad but necessary bit: If you disagree with this Privacy Statement, you should discontinue using our services. If you agree with our Privacy Statement, then you're all set to book your next Trip through us. Let the good times roll!"

Booking presents its advertising as a service to the customers, instead of admitting it is mostly in the interest of Booking itself: "If you have not finalized a Trip Reservation online, we can contact you with a reminder to continue with your reservation. We believe that this additional service benefits you as it allows you to carry on with a Trip Reservation without having to search for Trip Provider or fill in your reservation details again."

The privacy statement also mentions 'cross-device tracking': Booking tracks user behaviour across multiple devices and combines data collected from different devices.

Epic Games: Just like Amazon, Epic Games proposes in its privacy statement that users should change their settings elsewhere to limit the data it gathers, rather than offering this opt-out option on its own website (or, as they actually should, give users ability to opt-in). In this case the statement is about the transfer of data from third parties: "You can change your privacy settings on other parties' websites, such as social networks, which will stop or limit our receipt of information from those other websites." Thus, instead of asking the users for their consent, the privacy policy proposes them to change their settings on third-party websites. In addition, the words 'stop *or limit*' are vague and suggest that personal data can still be transferred from these third parties' websites.

When it comes to advertising, Epic Games even goes a step further. It refers to other websites through which users can opt out: "You can limit interest-based advertising by opting out at www.aboutads.info/choices/ or www.networkadvertising.org/choices/. If you are located in Europe, more information is available at www.youronlinechoices.eu/."

Facebook: The privacy statement of Facebook suggests awareness about the GDPR's provisions, but gives rather the impression of the company using the legal terms and buzzwords and catch-phrases, than attempting at constructing a truly user-centric, GDPR complain policy. For example, concerning the sensitive data, the company mentions is as a separate category, but it does not give any information about how this data is used: "You can choose to provide information in your Facebook profile fields or Life Events about your religious views, political views, who you are "interested in," or your health. This and other information (such as racial or ethnic origin, philosophical beliefs or trade union membership) is subject to special protections under EU law." Further, in its privacy policy, Facebook lists all legal bases for processing, almost copy-pasted from the regulation, without specifying what data will be processed in what way based on what purpose.

Moreover, Facebook states: "We also collect information about how you use features like our camera." Yet it remains unclear how this is done and what is being analysed. For example, does this include studying the objects of the photographs?

Facebook also gathers various types of data from the devices: "Device attributes: information such as the operating system, hardware and software versions, battery level, signal strength, available storage space, browser type, app and file names and types, and plugins." It is unclear why Facebook needs all of these data. Especially file names and types could be very problematic, as it suggests Facebook can gather information about all content on the devices of its users.

Further, Facebook receives data from third party partners, even about people who do not have a Facebook account: "these partners provide information about your activities off Facebook—including information about your device, websites you visit, purchases you make, the ads you see, and how you use their services—whether or not you have a Facebook account or are logged into Facebook." If people do not even have a Facebook account, it means that by definition they have not given consent for this. This problem has been, as a matter of fact, a subject of legal proceedings already before GDPR's applicability[58].

With Facebook it is particularly unclear where personal data end up and which third parties get access to it. Facebook makes sure its purposes sound noble, but it is doubtful to what extent these activities are actually part of their business: "We use the information we have (including from research partners we collaborate with) to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being. For example, we analyze information we have about migration patterns during crises to aid relief efforts. [...]We also provide information and content to research partners and academics to conduct research that advances scholarship and innovation that support our business or mission, and enhances discovery and innovation on topics of general social welfare, technological advancement, public interest, health and well-being."

Google: Just like Booking, Google tries to make its privacy policy accessible through the use of videos and relatively popular language. However, this popular language can actually also be rather unclear. Google uses the words 'things like' four times, for example: "We collect information to provide better services to all our users – from figuring out basic stuff such as which language you speak, to more

[58] See, for example: Samuel Gibbs, *Facebook ordered to stop collecting user data by Belgian court,* 16th February 2018, "The Guardian": https://www.theguardian.com/technology/2018/feb/16/facebook-ordered-stop-collecting-user-data-fines-belgian-court

complex things like which ads you'll find most useful, the people who matter most to you online or which YouTube videos you might like."

Just like Facebook, Google essentially copy-pastes the list of legal basis for processing into its policy. Particularly regarding the legitimate interests, the clause is astonishingly broad: "We process your information for our legitimate interests and those of third parties while applying appropriate safeguards that protect your privacy. This means that we process your information for things like:

- Providing, maintaining, and improving our services to meet the needs of our users
- Developing new products and features that are useful for our users
- Understanding how people use our services to ensure and improve the performance of our services
- Customizing our services to provide you with a better user experience
- Marketing to inform users about our services
- Providing advertising to make many of our services freely available for users
- Detecting, preventing, or otherwise addressing fraud, abuse, security, or technical issues with our services
- Protecting against harm to the rights, property or safety of Google, our users, or the public as required or permitted by law
- Performing research that improves our services for our users and benefits the public
- Fulfilling obligations to our partners like developers and rights holders
- Enforcing legal claims, including investigation of potential violations of applicable Terms of Service"

It is not specified what data will be used for what purpose exactly; moreover, most of them "customising our services to provide better experience", "performing research that (...) benefits the public" are essentially meaningless, and convey no information whatsoever.

Microsoft: Microsoft does not only collect personal data directly from the users or from third parties, it also 'infers or derives' information from other data, without specifying how: "Data about your interests and favourites, such as the sport teams you follow, the programming languages you prefer, the stocks you track or cities you add to track things like weather or traffic updates. In addition to those you explicitly provide, your interests and favourites can also be inferred or derived from other data we collect."

Microsoft mentions that users have the right to complain with a Data Protection Authority, but does not directly provide the names of these authorities. Instead, users should contact Microsoft to be directed: "If you have a privacy concern, complaint or a question for Microsoft's Chief Privacy Officer or Data Protection Officer, please contact us by using our Web form. We will respond to questions or concerns within 30 days. You can also raise a concern or lodge a complaint with a data protection

authority or other official with jurisdiction." There is no indication on how to do it, or how to determine who is the DPA "with jurisdiction".

Steam: In the section about the right to access and correct personal data, Steam includes the following clause: "If the request affects the rights and freedoms of others or is manifestly unfounded or excessive, we reserve the right to charge a reasonable fee (taking into account the administrative costs of providing the information or communication or taking the action requested) or refuse to act on the request." It is unclear when a request would be sufficiently unfounded or excessive for Steam to charge a fee or refuse to act on it. Steam basically copies the general provision of the GDPR (art. 12(5)). However, one could imagine that a controller, knowing the character of the service provided, and the costs of delivering potential requests, to be slightly more specific regarding what circumstances they will consider as excessive, or how much they plan to charge for that,

Moreover, the privacy policy states: "the minimum age to create a Steam User Account is 13. Valve will not knowingly collect Personal Data from children under this age." Yet, Steam offers numerous games aimed specifically at children. It seems that in such a case, when a product is directly aimed at users possibly younger than 13, more robust provisions and safeguards could be conceived of.

Uber: According to its privacy statement, Uber can collect location information of riders even when the app is running in the background: "In certain regions, Uber may also collect this information when the Uber app is running in the background of your device if this collection is enabled through your app settings or device permissions." It does not specify in which regions this is the case, nor whether a user should opt-in or opt-out from such a processing.

Uber also mentions a noteworthy feature, the so-called "Real-Time ID Check feature, which prompts driver partners to share a selfie before going online. This helps ensure that the driver partner using the app matches the Uber account we have on file, preventing fraud and helping to protect other users." Strikingly, the privacy statement does not include any further information about what exactly happens with these selfies, such as how long they are retained and who has access to them.

Uber's privacy statement contains another remarkable clause, in the section 'How we use your information': "Uber may also use the information to inform you about elections, ballots, referenda and other political and policy processes that relate to our services."

Additionally, regarding consent for processing expressed by a child, Amazon, Apple, Epic Games and Steam all mention a minimum age of 13, while the GDPR specifies 16 as the age of consent as a general rule, only allowing Member states to make that age lower, but no lower than 13. Regarding changing the policy: According to the privacy statement of Epic Games, users should check their privacy policy every time they use the website our play a game, to stay informed about potential updates: "This policy may be updated periodically to reflect changes in our personal information practices or relevant laws. We will indicate at the top of this policy when this policy was last updated. Please review this policy every time you access or use our websites, game engines, games, or applications to make sure that you

have reviewed the most recent version." Amazon, Microsoft, Booking and Uber include similar clauses and also encourage users to read the privacy statement regularly to stay updated. Yet in addition they mention that users will be informed directly about 'significant' or 'material' changes. Finally, some companies explicitly mention limitations to exercise of their rights by users (see Steam above). Apple for example states: "We may decline to process requests that are frivolous/vexatious, jeopardize the privacy of others, are extremely impractical, or for which access is not otherwise required by local law."

Summing up: even though all the companies under consideration have updated their privacy policies almost right before the GDPR started being applicable, every single privacy policy under consideration leaves room for significant improvements. The extensive usage of unclear language, significantly higher number of clauses conveying *insufficient information* rather than *full information*, and the alarming amount of clauses expressing will to engage in *problematic processing* is something that civil society and supervisory authorities should pay close attention to.

6. Automated Analysis

We developed a web crawler that monitors the privacy policies of a list of online services. The data retrieved by the crawler is then processed using supervised machine learning technology. In particular, we implemented a Support Vector Machine-based classifier trained on the data set annotated by experts following the guidelines discussed in the report. Such a data set contains over 3500 sentences taken from 14 privacy policies. The accuracy of the classifier was evaluated using a standard leave-one-document-out procedure, showing encouraging precision/recall in several sub-tasks. Our analysis indicates that the task of identifying problematic clauses in this kind of documents is in principle automatable. An extended data set is under construction, whose purpose is to improve the accuracy of the classification results. The expert annotations can be visualized using a standard browser at the CLAUDETTE GDPR web site, http://www.claudette.eu/gdpr/

6.1. Methods and Experiments

Our approach for the automated analysis of privacy policies is based on supervised machine learning methodologies.[59] Within this setting, our system is fed with a set of labelled documents, where relevant sentences are tagged according to the procedure defined in the previous section, and machine learning algorithms are trained to learn the dependencies between the text (input) and the tagging (desired output). Ideally, one would like to submit a new document to our trained system, and be returned the set of vague and problematic clauses that are present, as well as the set of required clauses that are present and adequate, or present but insufficient.

Given the complexity of the problem, there are several ways in which the system could be developed. For example, regarding problematic clauses, one could first detect all problematic clauses (in general) and then distinguish each category (data regarding children, advertising, etc.) or the other way round, so first detect all sentences regarding one certain category (e.g. advertising) and then decide whether they are problematic or nor. The same holds for required clauses. Vague clauses instead do not have sub-categories.

As for the adopted algorithms, in our current approach we started with the technologies that have been successfully employed in the detection of potentially unfair clauses in online Terms of Service.[60] Therefore, we tested several, different machine learning algorithms, including Support Vector Machines[61] (also in their structured variant, which takes into account also the sequentiality of sentences in the document), as well as Deep Learning models such as Convolutional Neural Networks and Long Short-Term Memory Networks.[62] Support Vector Machines are the best performing method, thus the results reported in the following section will regard that approach, if not otherwise stated.

In some cases, as it will be shown in the next section, a solution based on manually defined rules and patterns could also be used to detect some categories of problematic or required clauses, as often done in data mining applications[63]

[59] Bishop, C. M. (2006). Pattern recognition and machine learning, Springer.
[60] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2018). CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. arXiv preprint arXiv:1805.01217
[61] Bishop (2006).
[62] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning, Nature.
[63] Han, J., Kamber, M., Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Haryana: Morgan Kaufmann.

Regarding our software architecture, we developed a very simple web crawler that every night checks for updates in the list of monitored services. If any of these services has been updated (i.e., its text appears to be different from the day before), then the machine learning system is automatically called to process the new document, and results are updated on our server (http://www.claudette.eu/gdpr/).

6.2. Results of the Experiments

For the evaluation and comparison of the different systems, we employed a standard leave-one-document-out (LOO) procedure: given a set of N documents, the training procedure is repeated N different times, where a single document of the corpus is considered as the test set (for evaluation), whereas the remaining N-1 documents form the training set. We measure performance with metrics classically adopted in natural language processing and information retrieval applications:

(1) precision, as the percentage of predicted positive sentences that are indeed positive in the corpus;

(2) recall, as the percentage of positive sentences that are indeed classified as positive;

(3) f-measure, as the harmonic mean between precision and recall.

Here, the definition of positive clause depends on the task: it could be the set of problematic clauses, as well as the set of required or vague clauses.[64]

Unclear language clauses

As a first experiment, we considered unclear language clauses only. Here, a simple grammar that detects whether some keyword (or combination of keywords) is present in each sentence is capable of recognizing 89% of vague clauses, yet with a low precision of 25% (one sentence out of four is actually tagged as vague).

A machine learning classifier based on Support Vector Machines and bag-of-words, instead, detects 72% of vague clauses, yet with a low precision of 30%, which means that only one out of three sentences was tagged as unclear. A combination of grammar and machine learning achieves 81%/32% recall/precision.

Actually, a detailed analysis of these false positives (sentences detected as unclear, which actually were not tagged as such) shows that most of them are indeed problematic clauses. This observation made us argue that probably a machine learning classifier could take advantage of observing the combination of both problematic and unclear clauses.

[64] Sebastiani, F. (2002). Machine learning in automated text categorization, ACM computing surveys (CSUR).

Problematic and unclear clauses

We repeated the same experiments, this time considering the positive class (to be detected) as the union of problematic and vague clauses.

Following the same approach described above, a hand-crafted grammar correctly detects 92% of positive clauses, yet with 31% precision. A pure machine learning classifier achieves instead 70% recall and 50% precision. A combination of the two approaches reaches a 75% recall with 47% precision, with an overall 57% f-measure.

Although these numbers could seem unimpressive to a lay observer, we shall remark that as preliminary results they are not bad at all. Indeed, they are comparable to the results obtained with the analysis of Terms of Service with a corpus of 20 documents, where we had initially obtained a 72%/62% recall/precision[65], which further increased to 80%/83% when the corpus was extended to include 50 documents.[66] We could imagine a similar trend also for privacy policies, where our current corpus consists in 14 documents only. It is well known that by expanding the dataset size with consistent data, the error committed by the classifier typically lowers.[67] Moreover, while we employed rather standard techniques in machine learning and natural language processing, more sophisticated and dedicated methods could be developed, in order to exploit the context given by the whole document, which might give a significant contribution.

Once problematic clauses have been detected, performing an automatic categorization of such sentences into unlawfulness categories is a much simpler task. In this setting, Support Vector Machines are capable of identifying the correct category with a recall that is on average above 80%/75%.

Required information clauses

Some required information clauses can also be easily detected with manual grammars and regular expressions. For example, the category of automatic decision making can be identified with 95% precision and 83% recall, and the required information about complaints can be identified with 94% precision and 91% recall. Similarly, data protection officer clause can be detected with 78% precision and 85% recall. Other tags are much more heterogeneous, and thus difficult to detect with hand-crafted rules: this is the case, for example, of the purposes and the legal basis of data processing. In these cases, machine learning achieves performance that are comparable to those obtained for the detection of problematic clauses.

In summary, our analysis indicated that the task is in principle automatable, but a larger and more sophisticated training data set is necessary. We are assembling such a data set as you are reading this report.

[65] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Panagis, Y., Sartor G. & Torroni, P. (2017). Automated Detection of Unfair Clauses in Online Consumer Contracts. Legal Knowledge and Information Systems, 145.
[66] Lippi et al. (2018).
[67] Bishop (2006); LeCun et al. (2016).

7. A Few Remarks Regarding the Future of GDPR and Law-Automation

Everything written up to this point has been concerned with where we are right now. The purpose of this study has been to define the standard for a correctly designed, in form and in content, privacy policy, under the currently existing standards put forward by the GDPR; to analyse the privacy policies of 14 relevant online platforms and services in accordance with that standard; and to verify to what extent, given the amount of data we were able to generate in the aftermath of the mass privacy policy changing in the last days and weeks, such an analysis can be automated. In this last section we offer a brief reflection about the future. Firstly, we include some thoughts on the direction in which the law might go one day; secondly we draw a picture of how the future of automation could look like, if the efforts continue.

7.1. Future of GDPR

The GDPR has just started being applicable, and given how long it took to have it adopted, one should not expect that its contents shall change in the upcoming years. Nevertheless, pondering on how the European data protection regime might be updated, especially now when so much is being written about it, is always a worthwhile exercise – especially in the light of the legal actions have been initiated by civil society, which might result in a GDPR-based ruling of the Court of Justice of the European Union. Here we include several observations on what the European lawmaker could take into account.

1. Privacy policies are the main point of reference for civil society and individual consumers when it comes to controlling how personal data is being processed by the data controllers. The fact that these documents' legal status is not directly defined should be assessed negatively. Clear guidelines on the form and contents of these documents could significantly enhance levels of compliance by data controllers, who currently must rely on dispersed information, presented in various documents issued by the Article 29 Working Party. This could be one of the variables explaining the current, unsatisfactory, state of affairs when content and form of analysed privacy policies is concerned. Clearly, not the only one – it might just as well be that data controllers know how to design the policies and choose not do so, especially regarding the big players analysed in this report. Nevertheless, a clearer stipulation could on one hand be of help to good-will controllers with smaller resources, and on the other a more direct point of reference when criticising the non-complaint market players.
2. Regarding the tension between comprehensiveness and comprehensibility, there seems to be a blind spot between information requirements put in place for the benefit of consumers, and robust accountability mechanisms put in place in order to facilitate controls by supervisory authorities. In particular, the role of civil society – consumer organizations and activists – seems to be underappreciated. These entities, given standing by the GDPR, are capable of processing more technical and detailed information, which should be made publicly available by the data controllers. Currently, however, this particular obligation is only indirectly expressed by the GDPR.
3. The overall assumption of the Regulation seems to be the same as of the Directive 95/46, that is that personal data is being actively collected by some entity. However, in the information society we currently live in, data is often a by-product of consumers using online platforms and services. A lot of information processed by the data controllers must be processed for the services to function properly; it is the *other* uses that can be problematic. This distinction, even though present in the Regulation, could be made more explicit, and reflected in the information requirements for the data controllers.

7.2. Future of Civil Control and Law Automation

One of the purposes of this study has been to provide a proof of concept for the claim that the analysis of privacy policies can be automated using artificial intelligence. The results of our experiments show that in principle the task could benefit from a machine learning approach, although more data is needed in order to obtain higher quality results. Once the performance is good enough – and we will proceed with increasing the dataset volume – the developed technology could leverage applications such as:

1. Scanning tools for supervisory authorities and consumer organizations. Entities created to control the behaviour of data controllers have, obviously, limited factual capacities. The work of the humans they employ could be made significantly more efficient, if the most "automatic" (already on the level of natural language) tasks they have to engage in could be performed by machines.
2. Consumer-end tools. One could also envision creation of apps/browser extensions to be used by consumers. Instead of having to go through significantly long texts of privacy policies, consumer could be using an application allowing them to quickly see what quality of data processing they are agreeing to. Apart from giving consumers themselves more knowledge, and increasing their autonomy, these tools could also have functionalities that allow them to alert consumer organizations and supervisory authorities (send an email, with source and results of analysis, to a locally competent entity), or other consumers, by for example making the results available on social media.
3. Web-crawlers. The next step would be not only the automation of privacy policies' evaluation, but also retrieval. One could imagine a situation in which a web-crawler automatically traverses the web in search for privacy policies, scans them and communicates the results. On one hand, such a crawler could send information to the company ("I just read your privacy policy, and it seems suboptimal, please find a pdf with procedures on how to fix it attached. And, by the way, I also let the supervisory authority and local NGOs know about you"). On the other, as said in the hypothetical email, inform the civil society, the supervisory authorities, the press (all already equipped with their own AI-powered tools) about its finding.

These applications all concern AI-powered analysis of the privacy policies, i.e. textual documents. However, we may be not so far from the day when AI-based tools automatically check also data processing related *activities* such as tracking, cookies, etc. Algorithms controlling the algorithms. Automatically exercising the right to object. Automatically notifying civil society and supervisory authorities not only about potentially unlawful content of privacy policies, but also potentially unlawful data processing activities.

8. Conclusions and Takeaways

The purpose of this study has been: first, to clearly formulate the legal standards that privacy policies of online platforms and services should meet; second, to analyse the privacy policies of 14 chosen online service providers; third, to determine the extent to which this type of analysis can be automated using machine learning. The first objective has been realized *in abstracto* based on the GDPR's provisions, the second and third has been realized to the extent possible in the limited amount of time that has passed since GDPR's applicability date (given that essentially all online platforms and services under study amended their privacy polices around that date). Although we are still in the early days, we already obtained results that we consider significant and worthy of being made accessible to the public.

In summary, our study suggests that the current privacy policies of online platforms and services still have a significant margin for improvement. None of the 14 analysed privacy policies gets close to meeting the standards put forward by the GDPR. Unsatisfactory treatment of the information requirements; large amounts of sentences employing vague langue; and an alarming number of "problematic" clauses cannot be deemed satisfactory. All this leaves us with several takeaways.

Takeaways for civil society: we would like this report to be treated as a source of valuable – though preliminary – data on the levels of compliance when it comes to the privacy policies of the biggest online service providers. We make the policies analysed by us freely available online (http://www.claudette.eu/gdpr/), together with the precise legal threshold we have employed, presented here in the report. Moreover, civil society, in cooperation with machine learning experts, can develop tools that could significantly increase the efficiency of their work. But most of all: there is room for action.

Takeaways for academia: The results we deliver are promising. Clearly, more research is needed, especially when data creation is concerned, but our study suggests that the fields of legal informatics and applied machine learning can be significantly pushed forward. Additionally, work such as this one should be conducted in multiple languages. Since machine learning tools for natural language processing are language specific, a separate set should be created for each of the EU's official languages.

Takeaways for companies: there is much improvement to be made. Companies should take GDPR's requirements seriously, especially given the possibility of fines. Hopefully, they would start taking a more user-centric approach towards these documents, instead of treating them simply as a box to be checked. Moreover, if this study is treated as an inspiration to others, civil society might be soon equipped with artificial intelligence tools for the automated analysis of privacy policies. When this is the case, they will leave no stone untouched, no policy unread, no infringement unnoticed.

Bibliography

Literature:

1. Alpaydin, E. (2016). Machine Learning: The New AI. MIT Press
2. Bishop, C. M. (2006). Pattern recognition and machine learning, Springer
3. Cristianini, N. (2016). The road to artificial intelligence: A case of data over theory, "New Scientist", https://www.newscientist.com/article/mg23230971-200-the-irresistible-rise-of-artificial-intelligence/
4. Floridi, L. (2015). The onlife manifesto. Springer-Verlag GmbH.
5. Han, J., Kamber, M., Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Haryana: Morgan Kaufmann
6. Hildebrandt, M. (2015). Smart technologies and the end (s) of law: novel entanglements of law and technology. Edward Elgar Publishing
7. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning, Nature
8. Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. (2018). CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. arXiv preprint arXiv:1805.01217
9. Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Panagis, Y., Sartor G. & Torroni, P. (2017). Automated Detection of Unfair Clauses in Online Consumer Contracts. Legal Knowledge and Information Systems, 145
10. Micklitz, H. W., Pałka, P., & Panagis, Y. (2017). The Empire Strikes Back: Digital Control of Unfair Terms of Online Services. Journal of Consumer Policy, 40(3), 367-388
11. Sebastiani, F. (2002). Machine learning in automated text categorization, ACM computing surveys (CSUR)
12. Surden, H. (2014). Machine learning and law. Wash. L. Rev., 89, 87

Legal acts and official documents:

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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)
2. Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts
3. Article 29 Working Party, Guidelines on transparency under Regulation 2016/679, adopted on 29 November 2017, as last revised and adopted on 11 April 2018, 17/EN WP260 rev.01,
4. Article 29 Data Protection Working Party, WP243rev.01 Guidelines on DPO Adopted on 13 December 2016 As last Revised and Adopted on 5 April 2017
5. Article 29 Working Party Guidelines on Consent under Regulation 2016/679 (WP259 rev.01) Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018
6. WP29 Opinion 15/2011 on the definition of consent (WP 187)
7. Guidelines on the right to data portability, WP 242 rev.01, last revised and adopted on 5 April 2017.
8. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01) Adopted on 3 October 2017 As last Revised and Adopted on 6 February 2018
9. Article 29 Data Protection Working Party Opinion 2/2010 on Online Behavioural Advertising, 24 June 2010

This publication is part of an activity which has received funding under an operating grant from the European Union's Consumer Programme (2014-2020)."