# Towards dealing with GDPR uncertainty

Sushant Agarwal

Institute for Management Information Systems, Vienna University of Economics and Business, Vienna, Austria
sushant.agarwal@wu.ac.at

## 1 Introduction

Rapid increase in computational power over the last decade enables the processing of large amounts of personal data. IT applications these days deal with 'big data' and rely on data-science and data-driven decision making. On one hand, using data science researchers can predict an epidemic with high accuracy [1–3] , efficiently design traffic system for a city [4, 5] etc. On the other hand, data science is also being used by companies to process personal data of their customers for direct marketing [6, 7]. Constantly reducing prices for storage and computational power lure companies to collect more data than actually required, which is either used for present applications or stored for later exploitation [8]. However, this increasing amount of personal data collection makes companies vulnerable to privacy breaches. In simple words, the increased risk of misuse can be directly correlated with the amount of personal data [9, 10]. Hence, to protect customers' data and privacy, the regulators are adapting and making the regulations stricter [11]. Upcoming EU General Data Protection Regulation (GDPR) is one such example where data protection authorities would now be allowed to fine the companies which do not comply with the regulation [12]. The fines can cost up to 4% of a company’s global annual turnover. Consequently, companies will need to verify the compliance of their existing applications and make required changes in due time. We believe that tools which would help companies in checking and certifying compliance will have a lot of potential in the next 2-3 years, especially to avoid the hefty fines. Even Art 39 of GDPR discusses the need for establishing certification mechanisms for demonstrating compliance. However, GDPR is a legal document which involves subjectivity. Due to the subjectivity, it is difficult to completely automate the compliance checking and hence human intervention is required in this process. Also, a few articles currently lack best practices, for example standardized icons mentioned in Art 12 para 4b, format for data portability in Art 18 para 2 and consent authorization for child related processing mentioned in Art 8 para 1a. Nonetheless, this subjectivity and lack of best practices give rise to a number of challenges and opportunities for compliance certifying tools. For this research, focus is on compliance requirements for the regulation and automated tools for checking it. Requirements are classified as objective and subjective. Automated tools can check objective requirements based on a rule but cannot easily judge the subjective parts. For subjective parts, current best practices are evaluated and gaps between these best practices and the requirements are identified. These current best practices help in judging uncertainty for complying with subjective parts of the regulation. Hence, uncertainty is the approach used for subjective parts.

# 2 Framework to gauge uncertainty

## 2.1 Identifying articles with subjectivity

First articles from regulation are classified into objective and subjective so that the articles difficult for automated tools due to subjectivity can be segregated. Objective category contains paragraphs which can be checked via objective yes/no questions. For example, para 1 of Art 14 lists the information that should be provided to the data subject. Here, the compliance can be checked with a yes/no question like – 'Is the identity and contact details of the controller shared with the data subject?' The subjective category contains paragraphs which are difficult to measure objectively i.e. cannot be simply checked by a yes/no question. For example, para 1 of Art 12 expects that controllers communicate the processing of personal data to data subject in a "*concise, transparent, intelligible and easily accessible form, using clear and plain language*". Currently, as there is no scale to measure transparency or clarity in the language this requirement is subjective. Similarly, this category also contains paragraphs which have some subjectivity making it difficult for a tool to judge in an automated way. For example, art 5 para 1 (c) requires controllers to collect only adequate data ("data minimization"). In this case, a tool can check if all data fields being collected have a purpose defined but then human intervention would be required to judge whether that is legitimate and if that is the minimum possible data required to fulfil a defined purpose. Hence, checking compliance for subjective articles is not only a challenge but also an opportunity to translate them into objective requirements for compliance checking tools.

## 2.2 Analysis for Subjectivity: Challenges and Opportunities in Art 12

Second, based on the classification done in previous section, next we focus on the analysis of subjective parts. For the identified subjective parts in section 2.1, it is difficult to have a yes/no question to gauge compliance. Hence, an uncertainly scale can be introduced for tools to automate the process of compliance checks, which would represent a confidence level with which it estimated the compliance level. For this paper, the focus is specifically on subjectivity in article 12 -*"Transparent information, communication and modalities for exercising the rights of the data subject"*. Through this article, controllers are asked to communicate with the data subjects in a concise, transparent, intelligible and easily accessible form, using clear and plain language. These requirements are highly subjective and difficult to check by automated means without human intervention. Usually, this communication is done with the help of privacy policies to assist data-subjects in making informed privacy decisions [13]. Hence, mainly privacy policies are analysed for compliance with Art12. Art12 para 1 mentions following criteria: 1) concise, 2) transparent, 3) intelligible, 4) easily accessible form, 5) clear and plain language. Easily accessible form can be broken down into usability and accessibility. Similarly the concept of readability relates to both intelligibility and easy and plain language. Transparency with respect to Art 12 has complex interpretation and depends on all other criteria and also includes completeness. In the subsequent part, these measures would be elaborated. Hence, based on the above 5 criteria, we propose following constructs to be used for measuring uncertainty to comply with Art 12 – 1) conciseness, 2) readability, 3) usability, 4) accessibility and 5) completeness.

**Conciseness**

In their book The Elements of style, Strunk and White describe the requirements for conciseness as – *"A sentence should contain no unnecessary words, a paragraph no unnecessary sentences..." [14]*. Oxford dictionary defines concise as *"Giving a lot of information clearly and in a few words"* [15]. Regarding conciseness of privacy policies, to best of our knowledge, there are no benchmarks defined in literature setting guidelines for conciseness of a privacy policy. On the contrary, many studies have shown that as the policies are long and difficult to read, the data subjects ignore them [13, 16]. Aiming for conciseness, there have been research on illustrating privacy policies as short standardized tables (*"nutrition labels"*) [17] which were later proved to be not effective in comparison to a usual text based policies [18]. To show how impractical it is to read privacy policies, a study in US estimated that if Americans were to read online privacy policies word-to-word then the total time lost would have an approx. annual value of $781 billion [19]. Hence, based on these definitions and studies I argue that number of total words in a privacy policy can give and idea of conciseness which can then be used to check compliance for being concise.

**Readability**

Readability is defined as a measure of difficulty experienced by people reading a given text [20]. Flesch refers to readability as comprehension difficulty [21]. It is thus related to requirements of intelligibility and use of clear and plain language. In the literature, there are a lot of statistical formulations defined to understand readability and interpret it in quantifiable terms for instance - Flesch-Kincaid Reading Ease Score [21], Dale–Chall formula [22], Gunning-Fog Score [23], McLaughlin's SMOG formula [20]. For privacy policies, there have been many studies calculating readability for privacy policies [24–27]. Unfortunately, no standards or guidelines are available for the minimum score that companies should attain. However, the Flesch score has been used by Florida's insurance law and companies are expected to achieve a minimum score of 45 for readable text in insurance policies [28]. The Flesch score is also the most widely used formula and one of the most reliable and tested formula for readability [29]. Therefore based on the above reasons, for this paper, the Flesch score is selected to measure readability.

**Usability**

ISO 9241 defines usability as *"Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [30]. Usability in context of Art 12 is about the effectiveness in using a company's website to access the required privacy policy. Usability has been extensively studied in Human-Computer Interaction (HCI) field [31] and many theoretical models have been proposed to measure usability [32–35], especially from a theoretical modelling perspective.. However, most models have subjective constructs like learnability, supportability consistency etc. For measuring uncertainty related to usability, constructs defined in a paper by Lee and Kozar on Understanding of website usability [34] are used. Lee and Kozar extracted usability requirements from the literature and consolidated them into ten major constructs to measure usability.

**Accessibility**
Along with usability, the requirements that information is presented in 'easily accessible form' also include accessibility which is, in context of websites, defined as features using with *"people with disabilities can use the Web ... more specifically [they] can perceive, understand, navigate, and interact with the Web"* [36]. According to W3C, accessibility does not solely focus on social inclusion of people with disabilities but also considering older people, people in rural areas etc. Technological advances, especially the use of mobile based apps, require companies to address changing accessibility issues[37], for instance touch based navigation. Web Content Accessibility Guidelines (WCAG) provide a good set of requirements that can be checked for measuring uncertainty for compliance. Also, on their website of 70+ tools have been listed which can be directly used to rate accessibility [38].

**Completeness**
Completeness refers to the exhaustiveness of the privacy policy and is related to the requirement of transparency in communication. Art 12 mentions a list of things that companies should include in the privacy policy – 1) info referred in Art 14 and 14a and 2) communication related to Art 15 to 20 and Art 32. Hence, for measuring completeness, a checklist can be prepared based on the article. Although, PIA tools have a subsection to check exhaustiveness of the privacy policies, most of them are based on the older directive and need to be updated to include the changes in the regulation [39, 40].

## 3 Conclusions and Future Work

Hence, the paper deals with tackling subjectivity of the new GDPR and identifies measurable constructs to translate the requirements objectively. Based on the constructs, uncertainty in complying with the regulation can be estimated. For quantifiable constructs (conciseness and readability) we take a large sample of privacy policy per industry and use statistics to calculate benchmarks for a particular industry. For the remaining non-quantifiable constructs (usability, accessibility, completeness) we prepare a checklist to check compliance and identify the gaps i.e. requirements that cannot be assessed by automated means. This work thus aims to lay a foundation for tools which could automate the process of compliance checks.

For future work, the proposed measures (conciseness, usability, readability and completeness) would be incorporated in a compliance checking tool and would then be validated with industry for checking compliance of their IT applications. The aim is also to explore advanced methods for more precise estimation of the discussed measures. For example, the use of machine learning and Natural Language Processing (NLP) as a mean to automate compliance checking [41] and to improve the readability formulas [42] would be explored.

## References

1. Valdivia A, Lopez-Alcalde J, Vicente M et al. (2010) Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks-results for 2009-10. Eurosurveillance 15(29): 2–7
2. Schmidt CW (2012) Trending Now: Using Social Media to Predict and Track Disease Outbreaks. Environ Health Perspect 120(1): a30-a33. doi: 10.1289/ehp.120-a30
3. Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using Twitter
4. Collotta M, Lo Bello L, Pau G (2015) A novel approach for dynamic traffic lights management based on Wireless Sensor Networks and multiple fuzzy logic controllers. Expert Systems with Applications 42(13): 5403–5415. doi: 10.1016/j.eswa.2015.02.011
5. WEN W (2008) A dynamic and automatic traffic light control expert system for solving the road congestion problem. Expert Systems with Applications 34(4): 2370–2381. doi: 10.1016/j.eswa.2007.03.007
6. Cheung K, Kwok JT, Law MH et al. (2003) Mining customer product ratings for personalized marketing. Decision Support Systems 35(2): 231–243. doi: 10.1016/S0167-9236(02)00108-2
7. Ling CX, Li C (eds) (1998) Data Mining for Direct Marketing: Problems and Solutions, vol 98
8. Brown B, Chui M, Manyika J (2011) Are you ready for the era of ‘big data’. McKinsey Quarterly 4(2011): 24–35
9. Lebanon G, Scannapieco M, Fouad MR et al. (2006) Beyond k-Anonymity: A Decision Theoretic Framework for Assessing Privacy Risk. In: Domingo-Ferrer J, Franconi L (eds) Privacy in statistical databases: CENEX-SDC project international conference, PSD 2006, Rome, Italy, December 13-15, 2006 : proceedings. Springer, Berlin, New York, pp 217–232
10. Winkler WE (2004) Methods for evaluating and creating data quality. Information Systems 29(7): 531–550. doi: 10.1016/j.is.2003.12.003
11. Kuner C (2012) The European Commission's Proposed Data Protection Regulation: A Copernican Revolution in European Data Protection Law
12. (2015) On the protection of individuals with regard to the processing of personal data and onthe free movement of such data (General Data Protection Regulation)
13. Schaub F, Balebako R, Durity AL et al. (2004) A Design Space for Effective Privacy Notices. In: Proceedings of the Eighteenth Large Installation Systems Administration Conference (LISA XVIII): November 14-19, 2004, Atlanta, GA, USA. USENIX Association, Berkeley, CA, pp 1–17
14. Strunk W, White EB (2000) The elements of style, 4th ed. / with revisions, an introduction, and a chapter on writing by E.B. White. Allyn and Bacon, Boston, London
15. Oxford Dictionaries "concise". Oxford Dictionaries. https://www.oxforddictionaries.com/definition/english/concise

16. Milne GR, Culnan MJ (2004) Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. Journal of Interactive Marketing 18(3): 15–29. doi: 10.1002/dir.20009
17. Kelley PG, Bresee J, Cranor LF et al. (2009) A nutrition label for privacy
18. Kelley PG, Cesca L, Bresee J et al. (2010) Standardizing privacy notices: an online study of the nutrition label approach
19. McDonald AM, Cranor LF (2008) Cost of Reading Privacy Policies, The. ISJLP 4: 543
20. G. Harry Mc Laughlin (1969) SMOG Grading-a New Readability Formula. Journal of Reading 12(8): 639–646
21. Flesch R (1948) A new readability yardstick. Journal of Applied Psychology 32(3): 221. doi: 10.1037/h0057532
22. Chall JS, Dale E (1995) Readability revisited: The new Dale-Chall readability formula. Brookline Books
23. Gunning R (1969) The Fog Index After Twenty Years. Journal of Business Communication 6(2): 3–13. doi: 10.1177/002194366900600202
24. Ermakova T, Fabian B, Babina E (2015) Readability of Privacy Policies of Healthcare Websites. In: Wirtschaftsinformatik, pp 1085–1099
25. Singh RI, Sumeeth M, Miller J (2011) Evaluating the Readability of Privacy Policies in Mobile Environments. International Journal of Mobile Human Computer Interaction 3(1): 55–78. doi: 10.4018/jmhci.2011010104
26. McDonald AM, Reeder RW, Kelley PG et al. (2009) A Comparative Study of Online Privacy Policies and Formats. In: Goldberg I, Atallah MJ (eds) Privacy enhancing technologies: 9th international symposium, PETS 2009, Seattle, WA, USA, August 5-7, 2009 : proceedings. Springer, Berlin, New York, pp 37–55
27. Milne GR, Culnan MJ, Greene H (2006) A Longitudinal Assessment of Online Privacy Notice Readability. Journal of Public Policy & Marketing 25(2): 238–249. doi: 10.1509/jppm.25.2.238
28. (2015) Readable language in insurance policies. Chapter 627, Part II: 627.4145
29. Klare GR (1969) The measurement of readability. Iowa State University Press, Ames
30. Stewart T (1998) Ergonomic requirements for office work with visual display terminals (VDTs): Part 11: Guidance on usability. International Organization for Standardization ISO 9241
31. Shneiderman B (1987) Designing the user interface: Strategies for effective human-computer interaction. Addison-Wesley, Reading, MA
32. Kim J, Lee J, Han K et al. (2002) Businesses as Buildings: Metrics for the Architectural Quality of Internet Businesses. Information Systems Research 13(3): 239–254. doi: 10.1287/isre.13.3.239.79
33. Venkatesh V, Agarwal R (2006) Turning Visitors into Customers: A Usability-Centric Perspective on Purchase Behavior in Electronic Channels. Management Science 52(3): 367–382. doi: 10.1287/mnsc.1050.0442
34. Lee Y, Kozar KA (2012) Understanding of website usability: Specifying and measuring constructs and their relationships. Decision Support Systems 52(2): 450–463. doi: 10.1016/j.dss.2011.10.004

35. Lee Y, Kozar KA (2009) Designing usable online stores: A landscape preference perspective. Information & Management 46(1): 31–41. doi: 10.1016/j.im.2008.11.002

36. Caldwell B, Cooper M, Reid LG et al. (2008) Web content accessibility guidelines (WCAG) 2.0, vol 11. W3C

37. Díaz-Bossini J, Moreno L (2014) Accessibility to Mobile Interfaces for Older People. Procedia Computer Science 27: 57–66. doi: 10.1016/j.procs.2014.02.008

38. W3C (2016) Web Accessibility Evaluation Tools List. https://www.w3.org/WAI/ER/tools/. Accessed 05 Apr 2016

39. Alnemr R, Cayirci E, Corte LD et al. (2015) A Data Protection Impact Assessment Methodology for Cloud. In: Berendt B, Engel T, Ikonomou D et al. (eds) Privacy technologies and policy: Third annual Privacy Forum, APF 2015 Luxembourg, Luxembourg, October 7-8, 2015 revised selected papers / Bettina Berendt, Thomas Engel, Demosthenes Ikonomou, Daniel Le Metayer, Stefan Schiffner (Eds.). Springer International Publishing, pp 60–92

40. Oetzel MC, Spiekermann S (2013) A systematic methodology for privacy impact assessments: A design science approach. Eur J Inf Syst 23(2): 126–150. doi: 10.1057/ejis.2013.18

41. Costante E, Sun Y, Petković M et al. (2012) A machine learning solution to assess privacy policy completeness: (short paper)

42. François T, Miltsakaki E (2012) Do NLP and machine learning improve traditional readability formulas?