Observations on the regulatory effectiveness of Article 25 GDPR.

Nimród Mike LL.M.
Corvinus University of Budapest, Budapest, Hungary
nimrod.mike@uni-corvinus.hu

***Abstract* - Article 25 of GDPR is relatively novel to Europe in terms of being a legal stipulation. This article refers to the well-known data protection principles. Yet, its role should be more than just a reiteration of the sacred chalices provided by Article 5 of GDPR. This applied research paper is meant to discover the role that EU data protection authorities are giving to the concept of data protection by design and by default. The analysis applies machine learning on the forty-nine cases in which this article has been referred to as a reason for a fine. Machine learning is used to find the potential correlation between the severity of infringements and the number of fines within the presence of Article 25. The argument is that while there is not yet a single case in which the monetary fine was issued because of a sole infringement of Article 25, as time progresses, the authorities are developing a more detail-oriented approach in the investigation and fining practices. Hence, the hypothesis is constructed around the statement that data protection by design and by default is for now only a complementary article, where the controller infringed other ones. As we know, "all models are wrong, but some of them are useful." Therefore, the result might be subject to criticism due to the relative data-poor environment in which the model is generated. However, future work is potentially promising with growing case-count that fuels such modelling efforts. This will support researchers to become not only data-rich but also information-smart.**

Keywords - data protection, machine learning, data protection by design.

I. INTRODUCTION

Article 25 of GDPR implements the principles of Privacy by Design and Default (PbD). Among many others, PbD is a design philosophy to improve the overall privacy friendliness of IT systems [1], also a competitive business advantage [2], a set of technical solutions for privacy engineering, and a legal obligation [3]. The latter is where we notice transcendence. Regulatory approach first proposed PbD under the form of guidelines. Later PbD became an express legal obligation.

As a form of legal stipulation, PbD principles have been proposed for computer systems in general but did not supply enough details to be adopted by software engineers when designing and developing applications [4]. The lack of guidance on the 'how' of the PbD was omnipresent in academic discussions. PbD was meant to be technology neutral and therefore its primary goal was to focus on the 'what' and leave the 'how' to the development community. Part of this problem has its source in technicians and designers typically not being fluent in security and privacy [5].

This paper aims to discover the role that European data protection authorities (DPA) are giving to PbD. The argument is that while there is not yet a single case in which the monetary fine was issued because of a sole infringement of Article 25, as time progresses, DPAs are transitioning to a more detail-oriented approach in the investigation and fining practices. Hence, the hypothesis is constructed around the statement that data protection by design and by default is for now only a complementary article, where the controller infringed other ones.

II. RELATED WORK

There is relatively limited literature using data analytics methods to define the root cause or to predict the amount of GDPR fines. A general description of the fines has been provided in [6]. An applied study is the work provided by authors in [7]. Yet another methodology based on linear regression is used in [8].

III. DECISION TREE MODEL

A. Machine Learning Notions

To discover the role DPAs are giving to PbD we deploy a supervised machine learning technique called Decision Tree Model (DTM). DTM algorithms are constructed by implementing particular splitting conditions at each node, breaking down the training data into subsets of output variables of the same class [9]. This process of classification divides datasets into homogeneous subsets. The knowledge learned by a DTM through training is directly formulated into a hierarchical structure. This structure holds and displays the knowledge in such a way that it can easily be understood, even by non-experts [10].

The efficiency of DTMs is evaluated by various splitting indices. For a better understanding of such indices, let us define the notions of entropy, information gain, and gain ratio. As provided in [9], "entropy is the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity". Hence, "entropy is computed between 0 and 1, however, heavily relying on the number of groups or classes present in the data set it can be more than 1 while depicting the same significance, *i.e.*, extreme level of disorder". Therefore, if a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and if all the observations belong to one class, the entropy of that dataset becomes zero [11].

The concept of information gain is used for determining the best variables that render maximum information about a class, while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes [9]. Information gain computes the difference on the entropy before and after the split.

Finally, gain ratio is proposed to "normalize the information gain of an attribute against how much entropy that attribute has" [9]. In other words, gain ratio is information gain divided by entropy.

B. Data collection and preparation

We developed the training dataset using the CMS.Law's GDPR Enforcement Tracker, (www.enforcementtracker.com), by filtering the number of fines to violations that contained Article 25. Currently there are 48 entries. One additional case was discovered from the official communication of Romanian DPA [12]. The data collection and preparation concluded three significant steps.

First, we extracted the list of cases and developed additional attributes to establish numerical and binominal attributes for data analysis. The attribute glossary is described in Table 1.

Second, we defined the parameters for the label. The label is the attribute deciding if a case is rather severe due to Article 25 being referenced. The label is constructed with IF function on the combination of the following conditions:

- The administrative fine issued in the case must be higher than the GDP per capita (calculated in PPP) in the EU country, where the DPA issued the fine. For GDP per capita values, we used the reference date as provided in [13].
- The number of affected individuals has to be greater than the median of the number of affected individuals from all cases. Threshold is calculated on the available data and set to 1062.
- The decision contains any of the Articles 5, 6, 9, 12 to 22 of GDPR since these are defined as higher tier infringements under Article 83 para 5 of GDPR.

Third, we eliminated the attributes used in the label in order to reduce any possible bias that might be induced in the DTM algorithm. Thus, only the training dataset contains the amount of fine, the county GDP, and the number of affected persons.

C. Parameters for DTM

With the dataset composed, we created two different DTMs. The first is using gain ratio as splitting criterion with the parameters described in Table 2. The second is using accuracy for splitting criterion with the identical parameters.

TABLE I. ATTRIBUTE GLOSSARY

Attribute	
Name	**Meaning**
ETid	Permanent ID.
Country	Country in which the fines was given.
Complaints	If complaints received from affected individuals.
Industry	Industry in which controller / processor operates.
Type	Type of GDPR violation.
Art. 32	Article referenced in the decision.
Art. 33	Article referenced in the decision.
Art. 34	Article referenced in the decision.
Art. 35	Article referenced in the decision.
Art. 9	Article referenced in the decision.
Art. 5	Article referenced in the decision.
Art. 6	Article referenced in the decision.
Art. 12-13	Article referenced in the decision.
Art. 15-23	Article referenced in the decision.
Art. 28	Article referenced in the decision.
Private	Controller/processor is from private sector.
Days	Number of days since 25th May, 2018.
Label	Violation is severe or not, given that Article 25 is referenced.

Figure 1. Gain Ratio Country Config

D. Creating the DTM

TABLE II. DTM PARAMETERS

Parameter	
Name	Value
Maximal Depth	10
Apply pruning	Yes
Confidence	0.1
Apply prepruning	Yes
Minimal gain	0.01
Minimal leaf size	1
Minimal size for split	4

The DTMs are yielding the knowledge as illustrated in the figures below. As it is shown in Fig. 1, in order to decide if a violation is severe, the DTM refers to Article 5 of GDPR. If this article is not referenced in the decision, the violation is not severe. If it is referenced, the next split takes place upon Article 35. If this article is referenced in the decision, the violation is severe; otherwise, the algorithm will consider the number of days calculated from the enforcement date of GDPR. Where the number of days are less or equal than 501 (October 8, 2019), the label is pointing towards a severe violation, otherwise the splitting function is checking the country as a splitting criterion. Here we see many different approaches, since Belgium, Romania, Ireland, Iceland, and Hungary apparently do not consider in their decisions having Article 25 that the violation is severe, whereas Finland and Germany do so. Nonetheless, in case of Poland and Italy also industry specific leaves are created. In Poland, the Finance, Insurance and Consulting sector yields a severe violation with higher fines. However, the Media, Telecoms and Broadcasting, as well as the Education and Public Sector hold less severe violations with lower fines. In Italy,

Figure 2. Accuracy No Country Config

we see a rather different approach from the DPA. The sectors in which the violation is not severe are the Real Estate and the Health Care. On the contrary, Industry and Commerce, Transportation and Energy, Media, Telecoms and Broadcasting, Education and the Public Sector are heavily affected in this regard.

As illustrated in Fig. 2, cutting the countries will result in another insightful DTM. Here the main criterion is the type of violation, while we derive that insufficient fulfillment of data subject rights and insufficient fulfilment of information obligations will result in a less severe violation with lower fines. A separation is performed based on the day's attribute in case of insufficient technical and organizational measures to ensure information security. Here, if the violation occurred before 478 days (September 15, 2019), it is classified as severe, otherwise not. In case of an insufficient legal basis for data processing, the class of complaints are used to differentiate. Where data subjects lodged multiple complaints, the violation is severe. In a case where there was a single complaint submitted, if the decision also referenced Article 35, the violation is labeled severe. In case of non-compliance with general data processing principles, the DTM uses the industry to perform further splits. The violation is labeled severe in the Media, Telecoms and Broadcasting sector if the decision referenced Article 5. In the Education and Public Sector, we have a severe violation in case of a single complaint or no complaint. All infringements in Real Estate, Finance, Insurance and Consulting, as well as Industry and Commerce, are labeled as severe. Further, those infringements from Health Care and Unknown sector are appreciated to be less severe.

IV. Discussion

As presented in the earlier section, we gained valuable knowledge from interpreting the decision trees. The first argument that can be made is that instead of providing for a reference to an infringement of Article 25 of GDPR, the decisions are fundamentally based on Article 5 of GDPR. This supplies an insight into the fact that in the fining practices, the DPAs rather see Article 25 as an extension of Article 5, not a stand-alone reason for a fine to be issued.

Further, we see a coupled treatment of Article 25 with Article 35. The argument to be made is that the DPAs are taking a cause-and-effect relationship between these articles. This could be explained by telling that DPAs are looking at Article 25 as a tool utilization obligation and Article 35 as the tool discovery obligation. Thus, the controller must perform a data protection impact assessment (DPIA) in order to find out which tools are

supporting the data processing activity and then implement these, as mandated by Article 25.

Nevertheless, we see different treatments of severity on the country level and industry level. This highlights the inconsistency between the fining practices of DPAs across the EU. The more prominent DPAs to issue higher fines are the Italian Data Protection Authority (Garante per la protezione dei dati personali) and the Spanish Data Protection Authority (Agencia Española de Protección de Datos).

The same inconsistency is shown in cases of complaints lodged by affected data subjects. Certain cases have no complaints at all, the investigations being started due to a notified data breach by the controller itself [14]. Even in such circumstances, the violation is determined to be severe by the DTM, as the fines are significantly higher than the country GDP. Should be noted however that the highest fines are issued in case of multiple complaints [15 - 17]. The main cause here is the still rather insufficient legal basis for processing or non-compliance with general data processing principles.

V. LIMITATIONS

As we know "all models are wrong, but some of them are useful". This is a sentence often given as a response to speculative models presented by a data analyst. It highlights the uncertainty of assumed correlations. It has a similar effect to a disclaimer explaining how past performance cannot be used to reliably predict future performance. Therefore, the results presented in this paper might be subject to criticism due to the relative data-poor environment in which the models are generated. In this regard, it is certainly premature to base some conclusions at least on country level on the effectiveness of Article 25 GDPR as legal obligation. This limitation is especially well-founded when looking at the percentage ratio of reported cases when Article 25 was referenced compared to all reported cases. This ratio is shown in Fig. 3. However, future work is potentially promising with growing case count that fuels such modeling efforts.

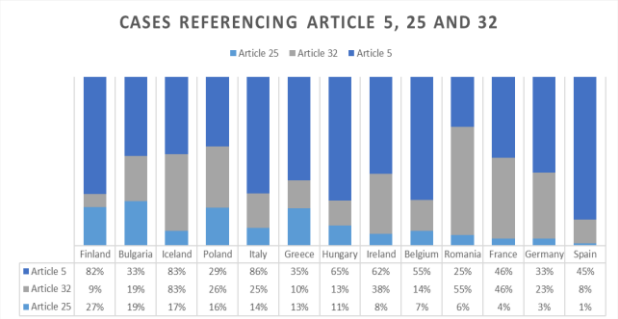

Figure 3. Percentage Ratio

VI. CONCLUSION

The analysis shows that the DPAs are rather taking the position that Article 25 is an aggravating circumstance for cases where a violation of another article is discovered. This is contrary to what Article 83 para. 5 of GDPR mandates: the infringement of this article is a separate reason to issue a fine. The DPAs are currently unable to find this applicable. This lets the practice question what the real content of the examined article is.

ACKNOWLEDGMENT

Special thanks to Quang Anh Nguyen for the support provided during the construction of DTMs in Rapidminer.

REFERENCES

[1] Hoepman, J.-H., "Privacy design strategies: extended abstract." IFIP Advances in Information and Communication Technology 428, pp. 446-459, Jan 2014.

[2] Cavoukian, A., Taylor, S. and Abrams, Martin., "Privacy by design: essential for organizational accountability and strong business practices.", Identity in the Information Society. Vol. 3., pp. 405-413., 2010.

[3] Rachovitsa, A., "Engineering and lawyering privacy by design: understanding online privacy both as a technical and an international human rights issue.", International Journal of Law and Information Technology. Vol. 24., Issue 4, pp. 374-399, 2016.

[4] Perera, C., Mccormick, C., Bandara, A. Price, B. and Nuseibeh, B., "Privacy-by-Design framework for assessing internet of things applications and platforms.", IoT'16: Proceedings of the 6th International Conference on the Internet of Things, pp. 83-92, Nov 2016.

[5] Shapiro, S., "Privacy by design: moving from art to practice.", Communications of ACM, Vol 53., No. 6, pp. 27-29., Jun 2010.

[6] Voigt P., von dem Bussche A., "Enforcement and fines under the GDPR." in The EU General Data Protection Regulation (GDPR)., Springer International Publishing, 2017.

[7] Ruohonen, J. and Hjerppe, K., "Predicting the amount of GDPR fines", Proceedings of the First International Workshop "CAiSE for Legal Documents" (COUrT), Grenoble (online), CEUR-WS, pp. 3-14, 2020.

[8] Mike, N, "Data protection has entered the chat: an analysis of GDPR fines". Masaryk University Journal of Law and Technology., in press.

[9] Tyagi, N., "What is information gain and gini index in decision trees?", Analytics Steps Blog (analyticssteps.com), Mar 2021.

[10] Seif, G., "A guide to decision trees for machine learning and data science", Towards Data Science Blog (towardsdatascience.com), Nov 2018.

[11] Tangirala, S., "Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm.", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 2, 2020.

[12] Description available at:
https://www.dataprotection.ro/index.jsp?page=Amenda_ING_RGPD&lang=ro [01.02.2022].

[13] GDP per capita (PPP) in Europe according to Trading Economics. Link available at: https://tradingeconomics.com/country-list/gdp-per-capita-ppp?continent=europe [01.02.2022].

[14] Decision with the ETid-1024 issued on Jan 27, 2022. Link available at: https://www.enforcementtracker.com/ETid-1024.

[15] Decision with the ETid-1005 issued on Dec 16, 2021. Link available at: https://www.enforcementtracker.com/ETid-1005.

[16] Decision with the ETid-336 issued on Jul 13, 2020. Link available at: https://www.enforcementtracker.com/ETid-336.

[17] Decision with the ETid-438 issued on Nov 12, 2020. Link available at: https://www.enforcementtracker.com/ETid-438