Articles Studia Iuridica Lublinensia vol. 32, 2, 2023
DOI: 10.17951/sil.2023.32.2.249-266

Maciej Siwicki
Nicolaus Copernicus University in Toruń, Poland
ORCID: 0000-0002-3120-0211
msiwicki@umk.pl

Big Data Profiling and Predictive Analytics from the Perspective of GDPR

Profilowanie i analiza predykcyjna z wykorzystaniem zbiorów big data z perspektywy RODO

ABSTRACT

The text analyses the normative regulations adopted by the 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 (GDPR) in order to answer the question whether the said regulations properly balance the interests of both entities that use predictive analytics and profiling in their economic activity, and of persons whose data they process. As this type of processing is based on big data, the proper analysis of this issue had to begin with determining which types of data processed in such sets can be considered personal information and in what conditions they can be treated as such. Based on these findings, the study analyzed the duties imposed by the GDPR on entities processing personal data in situations when such information has been obtained from big data. This in turn made it possible to assess the adopted normative regulations as well as point to the possible solutions and development paths.

Keywords: GDPR; personal data; profiling; big data; predictive analytics

CORRESPONDENCE ADDRESS: Maciej Siwicki, Dr. Habil., Associate Professor, Professor of the University, Nicolaus Copernicus University in Toruń, Faculty of Political Science and Security Studies, Batorego 39L, 87-100 Torun, Poland.

INTRODUCTION

Due to the development of predictive analytics,[1] based mostly on studying big data,[2] it has become easier to foretell human behavior. It has also been simpler to identify persons or groups with desired personal characteristics and to find their addresses, e-mails, location data as well as their names and surnames. Having access to such information, businesses can better plan their activity and adapt it to dynamically changing conditions as predictive analytics and profiling[3] enable systematic studying of relations between risks, customers' demands, and company's invested resources, thus allows optimization of processes and strategies in a given firm.[4]

While these technologies bring significant benefits to entrepreneurs, we must not forget that their use often significantly limits the rights of persons who are the object of such analysis. Their right of privacy is frequently breached without consent, in particular the ability to decide whether they want to remain anonymous in the Internet space or whether they want to be recognized, and if so, by what features. In this context, a need arises to adopt solutions that, on the one hand, will include legal instruments of preventive character that will allow an individual the right to control different aspects of processing their personal data, and on the other hand, will not negatively impact further technological development and competitiveness by, e.g., putting on entrepreneurs an excessive and limiting burden of unnecessary duties. Simultaneously, the adopted solutions must stigmatize all activities solely aimed at illegal obtaining and trading personal data.

---

[1] Predictive analytics is a part of statistics that studies and interprets data in order to determine patterns and trends that serve as basis for realistic prognoses.

[2] Big data is a loosely defined term used to describe data sets too large and complex for standard statistic software to cope with. The term has been used since the 1990s. Some consider J. Mashey as the person who contributed the most to making it popular. See S. Lohr, *The Origins of 'Big Data': An Etymological Detective Story*, 1.2.2013, https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story (access: 15.3.2023). More on this concept, see W.R. Wiewiórowski, *Założenia wstępne dla zrównoważonego przetwarzania informacji ze źródeł publicznych w czasach big data*, [in:] *Jawność i jej ograniczenia*, vol. 12: *Model regulacji*, ed. T. Bąkowski, Warszawa 2016, pp. 1–4.

[3] "Profiling" or "profile building" means a technique of automatic data processing that involves assigning to a specific person a so-called profile, based on data related to them, in order to make decisions concerning this person or to analyze/predict their preferences, behaviors and attitudes. For more, see X. Konarski, *Profilowanie danych osobowych na podstawie ogólnego rozporządzenia o ochronie danych osobowych – dotychczasowy i przyszły stan prawny w UE oraz w Polsce*, [in:] *Polska i europejska reforma ochrony danych osobowych*, eds. E. Bielak-Jomaa, D. Lubasz, Warszawa 2016, pp. 273–294. See also Article 29 Data Protection Working Party, Opinion 2/2010 on Online Behavioural Advertising, adopted on 22 June 2010, 00909/10/EN WP 171, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf (access: 15.3.2023).

[4] Accenture, *We, the Post-Digital People. Can Your Enterprise Survive the "Tech-Clash"?*, 2020, https://www.accenture.com/pl-en/insights/technology/technology-trends-2020 (access: 15.3.2023).

In the past, an analysis of personal factors, behaviors, interests, and socio-economic situation of a natural person, prepared in order to make predictions or a specific decision concerning such a person, required specific knowledge and qualifications, depended on access to relevant information and was often expensive. Today the progress in technology has made such a predictive model much easier and cheaper. There are widely available programs that not only facilitate running complex calculations and analyses but also enable appropriate collection and processing of data; furthermore, they can create advanced models presented in simple graphs.[5] Appropriate algorithms produce highly probable simulations of human behaviors and reactions to specific stimuli or scenarios. This can be done not only due to the option that allows building a credible customer's profile based on cross-analysis of the collected information regarding the said customer but also because the customer's attitudes towards various message forms and criteria of communication as well as the customer's reactions to specific products, services or brands have been determined. Big data analysis also facilitates personalization of offers, which are much better received by consumers because the former, e.g., better match the latter's needs or facilitate deepening the relations between the seller and the customer.[6] What contributed to the significant progress made in predictive analytics was the growing speed of data processing with many of the related processes.

Predictive analytics is widely used, e.g., in banking to assess the credit capacity of a customer based on their income, home budget, number of dependent persons, etc. The purpose of such analysis is not only to make the right credit-related decision(s) but also to predict the customer's preferences as well as their future behaviors and attitudes, which can translate into expanding the offer to include hitherto withheld banking products or services. In marketing, the analysis of the history of customer's behaviors in the Internet space (monitoring the pages visited and ads watched by the user, analyzing the likes on Facebook and Google search queries, etc.) is used to determine the customer's shopping preferences and the chance that a specific advertisement will be well received. Thus, such analysis contributes to more proactive and effective advertising strategies. Moreover, predictive analytics also helps in influencing people's voting preferences by predicting their behavior after they are presented with specific content;[7] it is also used in the support system for making medical decisions (to determine which patients are prone to, e.g.,

[5] For example, see IBM, *What Is Predictive Analytics?*, https://www.ibm.com/topics/predictive-analytics (access: 15.3.2023).

[6] See H. Stanley, *The Future of Personalization and How to Get Ready For It*, 20.10.2022, https://tiny.pl/cq52q (access: 15.3.2023).

[7] It is possible, among others, due to analyzing social media in terms of users' voting preferences. For example, see M. Rosenberg, N. Confessore, C. Cadwalladr, *How Trump Consultants Exploited the Facebook Data of Millions*, 17.3.2018, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (access: 15.3.2023).

diabetes, asthma or heart diseases[8]) as well as in the sphere of public security (i.a. detecting persons planning terrorist activities, determining individual inclination to become involved in criminal activity or the likelihood of someone committing repeated offenses). Even law firms use predictive analytics to assess the chances of a given party to win the case, based e.g. on comparing facts in the case with the body of rulings in search of similarities to other cases.[9]

To obtain the information needed, entrepreneurs use a variety of ways, both legal and illegal. The commonly mentioned legal methods include making use of the options allowed by the personal data protection regulations in force, e.g. the information has been obtained directly from the person it concerns or is available to the public. On the other hand, customers' data can be taken over without authorization or against the authorization issued by another data controller; they can be collected by various spy programs, obtained by breaking into databases,[10] or purchased from entities dealing with illegal acquisition of personal data.

Attempts to qualify the way of collecting information as legal or illegal are particularly problematic in the case of processing information in big data. The first source of difficulties is the fact that such a dataset is large in volume, heterogeneous, complex and changeable, and a stream of new information comes in real time, usually without any assessment of the type of the source of such data. Datasets of this type combine, e.g., information from the Internet, data obtained from different institutions and organizations (e.g. medical data), information published on social media, as well as bookkeeping or transaction data.

Another source of difficulties is the possibility that the controller of big data can use a variety of methods and technologies to obtain, sort and file data, and these methods can intentionally or accidentally lead to the identification of a natural person. Such connection between pieces of information can be made, e.g., when the database is expanded by including new information which, combined with the data already in the base, can become identifying data. Personal data can also be obtained as a result of combining various, seemingly unrelated pieces of information already included in a database. For example, data on Internet shopping can be compared both to traditional personal data and to a digital shadow. Another method of combining data from different sources is building an Internet behavioral profile of a customer by identification of IP numbers associated with a customer's account (e.g. by following

[8] See S. Buczyński, *Działania na zbiorach typu big data z perspektywy rozwoju i ochrony rynku usług zdrowotnych, detekcja white coat crime*, [in:] *Przeciwdziałanie patologiom na rynku medycznym i farmaceutycznym*, eds. A. Dobies, W. Pływaczewski, Warszawa 2019, pp. 155–162.

[9] For example, see Predictice, https://predictice.com (access: 15.3.2023).

[10] Serious security breaches occurred in the largest global companies such as Equifax, Target, Yahoo, Home Depot, and the United States Office of Personnel Management. For more information, see OPM U.S. Office of Personnel Management, *Cybersecurity Incidents*, https://www.opm.gov/cybersecurity/cybersecurity-incidents (access: 15.3.2023).

their Internet banking).[11] The possibility of identifying someone by connecting information is often independent of whether these are pieces of information that never were personal data or whether they have been subjected to the process of anonymization, i.e. removal of data enabling identification of a natural person. Moreover, obtaining personal data can be the outcome of legally acceptable actions, such as introduction of a digital production system or other technological solutions aimed at streamlining the company; it can also result from conscious illegal trading in personal data.

The third source of problems is the fact that information processed in big data is automatically analyzed in real time, with the use of a variety of – often imperfect – methods of data collection.[12] Solutions used in this process are based on imperfect algorithms of machine learning, which translates to significant difficulties with transforming information obtained from different sources into useful data, including also personal data (the so-called data cleaning). On the one hand, this process involves the risk of connecting specific pieces of information incorrectly, which may lead e.g. to ascribing to someone features this person does not possess, and thus may make the controller responsible for unauthorized disclosure of data as well as for injuring the customer's good name and violating their honor and dignity. On the other hand, information can be combined in a way that enables identification of a natural person, although such a possibility has not been foreseen or planned by the entity administering the dataset.

The problem of personal data protection has garnered widespread attention,[13] in particular such aspects as its essence and types as well as related threats and how this all is regulated by the 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 (GDPR),[14] the cases when an Internet user is traced and when personal data are

[11] S. Buczyński, *Działania na zbiorach typu big data z perspektywy ochrony praw e-konsumenta*, [in:] *Ochrona prawna konsumenta na rynku mediów elektronicznych*, eds. M. Królikowska-Olczak, B. Pachuca-Smulska, Warszawa 2015, pp. 129–136.

[12] For example, see K. Racka, *Big data – znaczenie, zastosowania i rozwiązania technologiczne*, "Zeszyty Naukowe PWSZ w Płocku. Nauki Ekonomiczne" 2016, vol. 23, pp. 319–320.

[13] Cf. J. Barta, P. Fajgielski, R. Markiewicz, *Ochrona danych osobowych. Komentarz*, Kraków 2011; P. Fajgielski, *Ochrona danych osobowych w telekomunikacji – aspekty prawne*, Lublin 2003; A. Mednis, *Cechy zgody na przetwarzanie danych osobowych w opinii Grupy Roboczej Art. 29 dyrektywy 95/46 nr 15/2011 (WP 187)*, "Monitor Prawniczy" 2012, no. 7; idem, *Ustawa o ochronie danych osobowych. Komentarz*, Warszawa 2001; *Ogólne rozporządzenie o ochronie danych osobowych. Komentarz*, ed. M. Sakowska-Baryła, Warszawa 2018; *Ogólne rozporządzenie o ochronie danych osobowych. Ustawa o ochronie danych osobowych. Wybrane przepisy sektorowe. Komentarz*, ed. P. Litwiński, Warszawa 2021; *Ustawa o ochronie danych osobowych przetwarzanych w związku z zapobieganiem i zwalczaniem przestępczości. Komentarz*, ed. A. Grzelak, Warszawa 2019.

[14] OJ L 119/1, 4.5.2016. See X. Konarski, *Profilowanie danych osobowych na podstawie ogólnego rozporządzenia o ochronie danych osobowych – dotychczasowy i przyszły stan prawny w UE*

processed in cyberspace,[15] legal basis for using artificial intelligence,[16] and the signs of a breach of the right to privacy in digital environment.[17] However, the literature only mentions predictive analytics and big data profiling,[18] while these issues are a necessary element in the activity of business which administer and use big data and thus deserve much more attention than it has received so far.

The purpose of this study is to answer the question if normative regulations adopted by GDPR properly balance the interests of the entities that use predictive analytics in their economic activity with the interests of persons whose data are thus processed. The paper utilizes dogmatic and analytical methods for the process of interpretation of the normative material and for the analysis of case law. The analyzed material included selected normative regulations and available literature on the subject.

RESULTS AND DISCUSSION

As assumed by EU law-makers, the GDPR is to contribute to creating "an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market, and to the well-being of natural persons" (Recital 2). This suggests that ensuring the rights of an individual regarding access to information about the said individual by other subjects is as important as striving to ensure free flow

oraz w Polsce, "Monitor Prawniczy" 2016, no. 20(Suppl.); P. Leja, *Ochrona danych osobowych a Internet rzeczy, profilowanie i repersonalizacja danych*, "Prawo Mediów Elektronicznych" 2017, no. 3; K. Szymielewicz, *Reforma europejskiego prawa o ochronie danych osobowych z perspektywy praw obywateli – więcej czy mniej ochrony?*, "Monitor Prawniczy" 2016, no. 20(Suppl.); M. Czerniawski, *Obowiązki administratora danych wynikające z prawa do przenoszenia danych*, "Monitor Prawniczy" 2017, no. 20(Suppl.).

[15] J. Byrski, H. Hoser, *Social media oraz technologie umożliwiające śledzenie użytkowników Internetu a współadministrowanie danymi osobowymi*, "Monitor Prawniczy" 2019, no. 21(Suppl.); J. Taczkowska-Olszewska, K. Chałubińska-Jentkiewicz, M. Nowikowska, *Retencja, migracja i przepływy danych w cyberprzestrzeni. Ochrona danych osobowych w systemie bezpieczeństwa państwa*, Warszawa 2019; J. Kurek, J. Taczkowska-Olszewska, *Ochrona danych osobowych jako realizacja zadań w obszarze bezpieczeństwa państwa*, Warszawa 2020.

[16] A. Krasuski, *Status prawny sztucznego agenta. Podstawy prawne zastosowania sztucznej inteligencji*, Warszawa 2021; *Prawo sztucznej inteligencji*, eds. L. Lai, M. Świerczyński, Warszawa 2020; E. Milczarek, *Prywatność wirtualna. Unijne standardy ochrony prawa do prywatności w internecie*, Warszawa 2020.

[17] W. Lis, *Zjawisko profilowania jako przejaw naruszenia prawa do prywatności w środowisku cyfrowym*, [in:] *Prawo prywatności jako reguła społeczeństwa informacyjnego*, eds. K. Chałubińska-Jentkiewicz, K. Kakareko, J. Sobczak, Warszawa 2017.

[18] P. Drobek, *Zasada celowości w dobie wielkich zbiorów danych (big data)*, "Monitor Prawniczy" 2014, no. 9(Suppl.).

of information, including personal data.[19] Like the right to protect personal data, the right to privacy is not absolute. In particular, the right to protect personal data should be considered in the context of its social function and balance against other fundamental laws according to the principle of proportionality (Recital 4). This means that the boundaries of these laws, the ways in which they are implemented as well as the scope of access to specific information are strictly related to the content of the distributed information.

As established by Article 2 (1), the GDPR applies to "the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system". As this regulation distinguishes two ways of processing (automatic and other), this has frequently translated into different shaping of the scope of GDPR application. The possibility to use the specific mechanisms of personal data protection differs also depending on the context, scope and aim of their processing.

1. Personal data and the criteria of identifiability

In order to determine when the information processed in big data sets should be considered to be personal data, we should recall the concept of personal data, which according to Article 4 (1) GDPR mean "any information relating to an identified or identifiable natural person (…) an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".[20] At the same time, the GDPR explains that online identifiers are identifiers such as internet protocol addresses or cookie identifiers, generated by devices, applications, tools and protocols, or other identifiers such as RFID tags.

According to the mentioned Article 4 (1) GDPR, personal data include not only the pieces of information that make it possible to identify an individual but also those that enable indirect determination of identity, in particular such information that identifies a person directly or indirectly only when combined with other data, e.g. information on family situation, medical history, financial status or education. In practice, pointing out criteria allowing indirect identification is particularly difficult.

---

[19] Cf. M. Jagielski, *Prawo do ochrony danych osobowych. Standardy europejskie*, Warszawa 2012, p. 29 ff. See also K. Szymielewicz, A. Walkowiak, *Autonomia informacyjna w kontekście usług internetowych: o znaczeniu zgody na przetwarzanie danych i ryzykach związanych z profilowaniem*, "Monitor Prawniczy" 2014, no. 9(Suppl.).

[20] Already the law on personal data protection from 1997 considered the following as information allowing identification: "An ID number as well as one or more specific features determining one's physical, physiological, mental, economic, cultural or social characteristics".

The problem with defining the indirect identification criteria was partially solved by Recital 26 GDPR, in sentences 3 and 4, which complement Article 4 GDPR. According to this regulation, to determine whether a natural person is identifiable, all the means reasonably likely to be used (such as singling out entries referring to the same person) to identify the natural person directly or indirectly by the data controller or another person. To establish whether means are reasonably likely to be used to identify the natural person, all objective factors should be taken into consideration, such as the costs of and the amount of time required for identification, as well as the technology available at the time of the processing and technological developments.

In the case of processing big data, all the mentioned factors can change, particularly when such a process is complex and long-lasting. As expected, this poses the risk that the possibility to identify specific persons will grow dynamically – the larger the data set, the higher the risk.

In many cases these changes are gradual, which begs the question: When will the identifiability threshold be crossed? G. Hornung and B. Wagner point out that the GDPR does not define when we can begin to consider a particular piece of information as personal data due to its character or context in which it appears. Nor does the GDPR answer whether the assessed information – due to its content, purpose or effect – must from the very start offer the possibility to identify a specific person (i.e. be about that person) or whether it can acquire this characteristic later.[21] In practice, the doubts are related to determining the stage of information processing and/or collecting at which the data have been enriched so far that they have become personal data, thus meeting the condition of identifiability of a natural person.

P. Litwiński argues that both the law and court rulings on the one hand subscribe to the opinion that the premise of a natural person's identifiability included in the definition of personal data should be understood objectively, which means that the possibility of identifying a person should be analyzed independently of the capabilities of the entity that is to conduct the identification. On the other hand, there are voices that it is also necessary to examine whether the entity with access to the data (which are in possession of a third party) is capable of using such information within their own means in order to identify a specific person.[22] Depending on the adopted stance, this may mean the necessity of investigation the conditions in a particular case only on the basis of objectivized criteria or with the inclusion of the subjective opportunity of a specific service provider/data controller to act.

[21] G. Hornung, B. Wagner, *Der schleichende Personenbezug: Die Zwickmühle der Re-Identifizierbarkeit in Zeiten von Big Data und Ubiquitous Computing*, "Computer und Recht 2019", vol. 35(9). This source further discusses the question of secondary identification in German legislation.

[22] P. Litwiński, *Pojęcie danych osobowych w ogólnym rozporządzeniu o ochronie danych osobowych. Glosa do wyroku TS z dnia 19 października 2016 r., C-582/14*, "Europejski Przegląd Sądowy" 2017, no. 5, pp. 49–54. A more extensive reference list can be found there.

Both the potential for the secondary identifiability of persons whom the data concern and the unintended ability to identify a person through combining pieces of information pose particular challenges for the entities processing information in big data. Ascertaining whether the regulations on personal data protection are applicable in a given case if the person responsible has an abstract possibility to collect information about a natural person, but this possibility is neither specific nor actively made use of. Clearly, resolving this question will depend on the stance adopted with regard to the subjective or objective assessment of the premise of the identifiability of a natural person.

According to P. Litwiński, in Poland the subjective understanding of this premise seems dominant. He also thinks that in the GDPR, the European Parliament also adopted the subjective approach by referring not only to the means of identification that are "reasonably probable" but also to the cases when there is "reasonable likelihood" that such means of identification will be used.[23]

The above issue also appeared on the margins of the rulings of the European Court of Justice (ECJ) on personal data protection in the context of providing services of the information society.[24] However, even those few rulings did not decide unequivocally how the premise of identifiability should be understood.

One of the best-known rulings related to this question is the judgment of the ECJ in the case *Patrick Breyer v. Bundesrepublik Deutschland*,[25] In the light of the Directive 95/46/EC,[26] the ECJ pointed out that "a dynamic IP address registered by an online media services provider when a person accesses a website that the provider makes accessible to the public constitutes personal data within the meaning of that provision, in relation to that provider, where the latter has the legal means which enable it to identify the data subject with additional data which the internet service provider has about that person". At the same time the ECJ pointed out that the ability to combine an IP address with additional data offering the possibility to identify a specific person should be assessed "rationally", considering whether the identification of the person whom the data concern is "prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power". In this context, it should be evaluated whether "the risk of identification appears in reality to be insignificant". The significance of this judgment lies in the fact that the ECJ emphasized here that what is most

[23] *Ibidem*.

[24] *Ibidem*.

[25] Judgment of the ECJ of 19 October 2016, case C-582/14, *Patrick Breyer v. Bundesrepublik Deutschland*, ECLI:EU:C:2016:779.

[26] 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 (OJ L 281/31, 23.11.1995.

important in assessing the premise of identifiability is the rational evaluation of the context of data processing (i.e. time required, costs, etc.).

A similar stance was expressed in the European Data Protection Board's *Guidelines 04/2020 of 21 April 2020*.[27] According to this document, the assessment of a specific anonymization requires "objective aspects (time, technical means) and contextual elements that may vary case by case (rarity of a phenomenon including population density, nature and volume of data)".

The above remarks demonstrate that the ascertainment of the premise of identifiability cannot be based only on evaluation of objective aspects but must also rationally assess the entire context of data processing.

Rationality above all means being guided by logic. Thus, a rational assessment cannot be based solely on an "assumption" that specific pieces of information will make it possible to identify a specific person when combined, but it must consider the current level of knowledge and technical solutions used by the data controller. A rational assessment not only avoids going against logic but also complies with the commonly accepted standards of a "reasonable person". Again, this means that an abstract possibility of combining pieces of information in a way enabling identification of a natural person is insufficient as it may turn out that in a given case it is not feasible due to non-proportional efforts (time, costs, labor, etc.) or due to significant obstacles installed by a given service provider/administrator of a technical solution.

Nevertheless, considering Recital 26 GDPR, an assessment whether during the processing of big data we encounter personal data should focus on the following factors:

- time needed to search for additional information as well as incurred costs/other invested resources,
- technical characteristics of the tools used, including hardware (such as computing capabilities of a unit owned) and software (the specifics of algorithm operation),
- characteristics of a dataset and the possibilities of accessing additional information (e.g. to widely available sources),
- human resources, including the knowledge and experience of the personnel.

Particularly in the case of information processing in big data, due to the technological and content-related diversity,[28] a rational assessment should mean the analysis of the specific context of data processing. At the same time, rationality

[27] European Data Protection Board, *Guidelines 04/2020 on the Use of Location Data and Contact Tracing Tools in the Context of the COVID-19 Outbreak*, 21.4.2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf (access: 15.3.2023).

[28] Such sets include, e.g., publicly accessible information (such as data from social media including the time of publishing an entry, language, interactions with other users, user's geolocation, clickstreams from webpages, information published in blogs, portals, etc.), information from databases

excludes formulating a fixed definition of how the premise of identifiability should be understood. On the contrary, in each case it requires considering the practices that specific entrepreneurs use in their economic activity.

2. Obligations of entities processing personal data

Data processing is understood as operations or sets of operations performed on personal data or sets of personal data in an automated or non-automated manner (Article 4 (2) GDPR). At the same time, it is obvious that there is no data processing if the operations of collecting, recording, organizing, ordering, etc. of data are performed on information that does not enable the direct or indirect identification of a natural person. This is indicated by Recital 26 GDPR, according to which the principles of data protection should not apply to anonymous information or to anonymized personal data, including processing for statistical or scientific purposes. This position also results indirectly from Article 5 (1) (b) GDPR (the purpose limitation principle), according to which the collection and recording of personal data must take place for an explicit and legitimate purpose. The key to determining the fact of processing is the intention with which a given entity came into possession of personal data (purpose of processing).

Even if the processing of personal data is accidental, e.g. as a result of an unintentional joining of information into personal data, there are numerous obligations on the part of such a service provider, for which they shall be responsible as a controller. For instance, according to Article 5 (1) GDPR specifying the rules regarding the processing of personal data, the person responsible for the processing of previously non-personal data, in the case of linking them with information enabling identification, shall be obliged to:

a) process them in accordance with the requirements of transparency, fairness and law – Article 5 (1) (a) GDPR,
b) specify an explicit and legitimate purpose – Article 5 (1) (b) GDPR,
c) limit the scope of processing in terms of quantity and content to the extent needed to achieve the purposes of their processing – Article 5 (1) (c) and (e) GDPR,
d) correct and update data, if necessary – Article 5 (1) (d) GDPR,
e) implement appropriate technical or organizational measures – Article 5 (1) (f) GDPR.

The controller can easily fulfill certain obligations without knowing specifically which person the data concern. This applies, e.g., to the implementation of data protection measures at the stage of designing technological solutions (Article 25

collected by businesses (e.g. big operators such as eBay.com, Amazon.com, Google, and Facebook), information made available by the public administration, as well as data produced by smart devices.

(1) GDPR), concluding agreements with entities that process contracts (Article 28 GDPR), maintaining a record of processing activities (Article 30 GDPR), taking data security measures (Article 32 GDPR), carrying out data protection impact assessment (Article 35 GDPR) and certification (Article 42 GDPR), designating the data protection officer (Article 37 GDPR) and complying with restrictions on the transfer of data to a third country (Article 44 ff. GDPR). In these cases, the identification of persons whose data is in question comes as useful, but not necessary.

However, there are also such obligations the implementation of which will require the controller to have/obtain additional, detailed information. This applies in particular to the implementation of the principle of transparency related to the need to provide the data subject with information about the purpose and recipients of the personal data (Articles 12–14 GDPR), and the principle of data correctness (accuracy), which requires the controller to ensure compliance with the actual state, completeness and validity of the data.

Many provisions require the controller to know specific facts, e.g. in some cases, in order to legalize the processing of personal data, the controller must obtain consent from a specific entity (Article 6 (1) (a), Article 9 (2) (a) in conjunction with Articles 7 and 8 GDPR), or in order to apply the premise of Article 6 (1) (f) GDPR, the controller must assess whether a negative condition is met in the form of the existence of interests or fundamental rights and freedoms of the data subject in a given actual state, which override the legitimate interests of the controller or a third party.

Without the knowledge of the relevant contact information, it is not possible to meet the information requirements under Articles 13 and 14 GDPR, or the secondary information obligation under Article 15 GDPR, in particular making the information publicly available (Article 14 (5) (b) GDPR). Finally, the regulation imposes on the controller an additional obligation to know the nature, scope and context of personal data processing and the risk of violating the rights or freedoms of data subjects. Due to the risk, the controller is subject to additional obligations related to data processing (e.g. Articles 24, 25, 32, 33, 34, 35 GDPR).

This means that in certain circumstances, in order to comply with the provisions on the protection of personal data, the controller must obtain additional information to identify the data subject only to comply with the provisions of the regulation (e.g. in order to obtain from their consent to the processing of data, informing them of their rights, and assessing the risk related to the processing of their data).

If the circumstances where the above obligations arise, the question is how much time the controller has to implement them or in what phase of processing they should be implemented at the latest?

When answering the above question, it is helpful to refer to Article 13 (1) and Article 14 (1) GDPR. A literal interpretation leads to the conclusion that the information obligations indicated in these provisions shall be imposed on the controller when collecting personal data from the data subject as well as when obtaining them

from other sources, also those publicly available. The above-mentioned provisions are in no way related to a specific stage of processing. This obligation shall be imposed on the controller when obtaining personal data from third parties as well as when obtaining personal data as a result of extending the already possessed information.

Article 14 (3) GDPR sets out different deadlines for the information transfer. The controller has the possibility to choose from three options:

1) within a reasonable time after obtaining the personal data – within a month at the latest – having regard to the specific circumstances of personal data processing,
2) if personal data are to be used for communication with the data subject – at the first such communication to that data subject at the latest, or
3) if it is planned to disclose personal data to another recipient – at the latest when they are first disclosed.

Article 14 (3) GDPR uses the term "after obtaining", but it is obvious that in the cases in question it will be the moment when the given information becomes identifiable. Such a reference to a "reasonable time" is not stated in many provisions imposing various obligations on the controller. From their literal wording, the results directly state that they apply immediately from the moment of ascertaining the fact of personal data processing.[29]

In order to legally process personal data from the very start, at least one of the conditions set out in Article 6 (1) GDPR must be observed. Also from the moment when the controller is dealing with personal data, the processing must be lawful, fair and transparent for the data subject (Article 5 (a) GDPR). The controller shall guarantee that the processing is carried out in a manner that ensures adequate security of personal data (Article 5 (f) GDPR), as well as provide adequate data protection already at the design stage, which requires learning the full context of data processing, including external and internal threats (Article 25 GDPR). Moreover, the controller must appoint a representative (Article 27 GDPR), conclude specific agreements with processor (Article 28 GDPR), record processing activities (Article 30 GDPR), ensure security of data processing adequate to the risk (Article 32 GDPR), carry out data protection impact assessment (Article 35 GDPR), designate a data protection officer (Article 37 GDPR) and comply with the requirements for transfers to third countries (Article 44 ff. GDPR).

The requirement of an immediate fulfillment of the above obligations if there arises a connection to a specific person seems unjustified, particularly when the purpose of the activity of a given entity is not to obtain personal data from information that was, e.g., subject to anonymization. Despite the lack of appropriate normative regulations, it is obvious that the entity responsible for data processing should have adequate time to fulfill their obligations. This should depend on the

[29] This fact is also discussed by G. Hornung and B. Wagner (*op. cit.*, p. 565 ff.).

nature of the given obligation and should be shorter in the case of sensitive data processing.[30] The person in charge should also seek to determine the legal status of the processed information.

Fulfillment of some of the above obligations can often be in conflict with the principle of minimization. Noting this fact, Recital 57 GDPR states that where the personal data processed by a controller do not enable them to identify a natural person, they shall not be required to obtain additional information to identify the data subject solely for the purpose of complying with provisions of the Regulation.

The principle of minimization is also expressed in Article 11 (1) GDPR. According to it, if the purposes for which the controller processes personal data do not or no longer require the identification of the data subject by the controller, the controller is not required to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with the GDPR. The principal effect of the lack of ability to identify persons is the possibility for the controller to further dispose of information that, by definition, no longer constitute personal data. In the event of possessing information that does not ensure the identification of data subjects, this provision exempts the controller from the obligation to acquire additional information (Article 11 (1) GDPR).

In practice, the application of this provision in the case of data processing in large sets (big data) raises numerous doubts. This provision talks about "additional information" and about "obligation to maintain, acquire or process", but it does not make it clear whether these exemptions cover the data that come from a third party or are the result of operations on information processed in one large data set. In the German-language literature on the subject, there is a view that Article 11 (1) GDPR refers to the situation of obtaining data from third parties, and not obtained as a result of actions taken within the same data set.[31]

In the light of the above doubts, it seems unjustified to exempt from the provisions of the GDPR in a situation when the controller has all the information enabling the identification of an individual, but it requires additional work related to a certain organizational and/or technical effort. However, the requirement of equal treatment of all entities involved in the processing of personal data and the reference to the purposes of Article 11 GDPR are in contrast to the above interpretation. The doubt is mainly about the basis on which entities that obtain additional information from third parties are favored over those who obtain it on their own as a result of additional activities, particularly since the result of the activities of both entities is the same.

It should be emphasized that Article 11 GDPR does not in any way exempt from all obligations arising from the Regulation. From the point of view of the meaning and purpose of the rule, this provision may refer only to those provisions

[30] See *ibidem*.

[31] *Ibidem*.

that require the controller to have detailed knowledge of the data subject. This applies to all those provisions that require communication with the data subject or when it is necessary to assign specific data to them. In fact, Article 11 GDPR is a special provision that contains controversial solutions without specifying in detail the scope of its validity. Its analysis also does not unequivocally answer the aforementioned doubts.

CONCLUSIONS

One of many reasons behind the increase in illegal circulation of personal data is that data analysts use this information to develop business and advertising strategies.[32] In order to prevent these activities, numerous obligations have been imposed on data controllers. In the case of entities storing big data, the arising obligations can lead to their resignation from acquiring data that can identify the subject and alternatively to certain solutions significantly limiting the service provider/controller's ability to know the identity of natural persons (e.g. anonymization based on partial erasing of information that can identify an individual). On the other hand, these entities can take actions aimed at hiding all or part of their activities from the authorities responsible for personal data protection, or transfer their activities to a country where legal solutions are most favorable for them.

Having in mind the assumptions of the GDPR, it seems obvious that the controller should not always be expected to actively search information in order to identify an individual, as this would lead to numerous absurdities. In practice, particularly in the case of indirect and secondary identification, there may be a significant risk of assigning additional information to a wrong person, which for the controller will entail the risk of incorrect data processing, and for third parties the risk of receiving incorrect data or transferring them to unauthorized persons.

In order to avoid the above problems, it is necessary to regularly analyze the possible scenarios of combining information, particularly in the case of big data. An option here could be to introduce certain solutions that would inform the controller about the risk of possible identification of an individual. This might be, e.g., an alert system enabling actions to re-anonymize the data or adapt the controller's activities to the requirements of the GDPR. However, such solutions will also be difficult to implement since the systems that check identifiability should be actually perceived as a tool fulfilling the intentions of the service provider to take actions leading to the ongoing identification of persons whose data is processed. If regular checks

[32] P. Mudgal, *Illegal Acquisition of Data/Information by Authorities, Apps and Social Media*, 16.10.2020, https://blog.ipleaders.in/illegal-acquisition-data-information-authorities-apps-social-media (access: 15.3.2023).

are made as to whether specific data subjects are identifiable, the purpose of such activity is indirectly to undermine the anonymity of such persons. In this case, it is not possible to talk about some systemic resignation from acquiring data that can identify individuals, or about data processing that does not allow the controller to know the identity of such persons, i.e. the situation referred to in Article 11 GDPR. This limitation means that it is only possible to introduce such technical mechanisms that will check the framework operation of individual systems, or that will refer to the structure and size of databases, introduced categories, and metadata specificity in order to assess whether risks leading to deanonymization can occur in a given case.

REFERENCES

Literature

Barta J., Fajgielski P., Markiewicz R., *Ochrona danych osobowych. Komentarz*, Kraków 2011.

Buczyński, *Działania na zbiorach typu big data z perspektywy ochrony praw e-konsumenta*, [in:] *Ochrona prawna konsumenta na rynku mediów elektronicznych*, eds. M. Królikowska-Olczak, B. Pachuca-Smulska, Warszawa 2015.

Buczyński S., *Działania na zbiorach typu big data z perspektywy rozwoju i ochrony rynku usług zdrowotnych, detekcja white coat crime*, [in:] *Przeciwdziałanie patologiom na rynku medycznym i farmaceutycznym*, eds. A. Dobies, W. Pływaczewski, Warszawa 2019.

Byrski J., Hoser H., *Social media oraz technologie umożliwiające śledzenie użytkowników Internetu a współadministrowanie danymi osobowymi*, "Monitor Prawniczy" 2019, no. 21(Suppl.).

Czerniawski M., *Obowiązki administratora danych wynikające z prawa do przenoszenia danych*, "Monitor Prawniczy" 2017, no. 20(Suppl.).

Drobek P., *Zasada celowości w dobie wielkich zbiorów danych (big data)*, "Monitor Prawniczy" 2014, no. 9(Suppl.).

Fajgielski P., *Ochrona danych osobowych w telekomunikacji – aspekty prawne*, Lublin 2003.

Grzelak A. (ed.), *Ustawa o ochronie danych osobowych przetwarzanych w związku z zapobieganiem i zwalczaniem przestępczości. Komentarz*, Warszawa 2019.

Hornung G., Wagner B., *Der schleichende Personenbezug: Die Zwickmühle der Re-Identifizierbarkeit in Zeiten von Big Data und Ubiquitous Computing*, "Computer und Recht 2019", vol. 35(9), **DOI: https://doi.org/10.9785/cr-2019-350910**.

Jagielski M., *Prawo do ochrony danych osobowych. Standardy europejskie*, Warszawa 2012.

Konarski X., *Profilowanie danych osobowych na podstawie ogólnego rozporządzenia o ochronie danych osobowych – dotychczasowy i przyszły stan prawny w UE oraz w Polsce*, "Monitor Prawniczy" 2016, no. 20(Suppl.).

Konarski X., *Profilowanie danych osobowych na podstawie ogólnego rozporządzenia o ochronie danych osobowych – dotychczasowy i przyszły stan prawny w UE oraz w Polsce*, [in:] *Polska i europejska reforma ochrony danych osobowych*, eds. E. Bielak-Jomaa, D. Lubasz, Warszawa 2016.

Krasuski A., *Status prawny sztucznego agenta. Podstawy prawne zastosowania sztucznej inteligencji*, Warszawa 2021.

Kurek J., Taczkowska-Olszewska J., *Ochrona danych osobowych jako realizacja zadań w obszarze bezpieczeństwa państwa*, Warszawa 2020.

Lai L., Świerczyński M. (eds.), *Prawo sztucznej inteligencji*, Warszawa 2020.

Leja P., *Ochrona danych osobowych a Internet rzeczy, profilowanie i repersonalizacja danych*, "Prawo Mediów Elektronicznych" 2017, no. 3.

Lis W., *Zjawisko profilowania jako przejaw naruszenia prawa do prywatności w środowisku cyfrowym*, [in:] *Prawo prywatności jako reguła społeczeństwa informacyjnego*, eds. K. Chałubińska-Jentkiewicz, K. Kakareko, J. Sobczak, Warszawa 2017.

Litwiński P., *Pojęcie danych osobowych w ogólnym rozporządzeniu o ochronie danych osobowych. Glosa do wyroku TS z dnia 19 października 2016 r., C-582/14*, "Europejski Przegląd Sądowy" 2017, no. 5.

Litwiński P. (ed.), *Ogólne rozporządzenie o ochronie danych osobowych. Ustawa o ochronie danych osobowych. Wybrane przepisy sektorowe. Komentarz*, Warszawa 2021.

Mednis A., *Cechy zgody na przetwarzanie danych osobowych w opinii Grupy Roboczej Art. 29 dyrektywy 95/46 nr 15/2011 (WP 187)*, "Monitor Prawniczy" 2012, no. 7.

Mednis A., *Ustawa o ochronie danych osobowych. Komentarz*, Warszawa 2001.

Milczarek E., *Prywatność wirtualna. Unijne standardy ochrony prawa do prywatności w internecie*, Warszawa 2020.

Racka K., *Big data – znaczenie, zastosowania i rozwiązania technologiczne*, "Zeszyty Naukowe PWSZ w Płocku. Nauki Ekonomiczne" 2016, vol. 23.

Sakowska-Baryła M. (ed.), *Ogólne rozporządzenie o ochronie danych osobowych. Komentarz*, Warszawa 2018.

Szymielewicz K., *Reforma europejskiego prawa o ochronie danych osobowych z perspektywy praw obywateli – więcej czy mniej ochrony?*, "Monitor Prawniczy" 2016, no. 20(Suppl.).

Szymielewicz K., Walkowiak A., *Autonomia informacyjna w kontekście usług internetowych: o znaczeniu zgody na przetwarzanie danych i ryzykach związanych z profilowaniem*, "Monitor Prawniczy" 2014, no. 9(Suppl.).

Taczkowska-Olszewska J., Chałubińska-Jentkiewicz K., Nowikowska M., *Retencja, migracja i przepływy danych w cyberprzestrzeni. Ochrona danych osobowych w systemie bezpieczeństwa państwa*, Warszawa 2019.

Wiewiórowski W.R., *Założenia wstępne dla zrównoważonego przetwarzania informacji ze źródeł publicznych w czasach big data*, [in:] *Jawność i jej ograniczenia*, vol. 12: *Model regulacji*, ed. T. Bąkowski, Warszawa 2016.

Online sources

Accenture, *We, the Post-Digital People. Can Your Enterprise Survive the "Tech-Clash"?*, 2020, https://www.accenture.com/pl-en/insights/technology/technology-trends-2020 (access: 15.3.2023).

Article 29 Data Protection Working Party, Opinion 2/2010 on Online Behavioural Advertising, adopted on 22 June 2010, 00909/10/EN WP 171, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf (access: 15.3.2023).

European Data Protection Board, *Guidelines 04/2020 on the Use of Location Data and Contact Tracing Tools in the Context of the COVID-19 Outbreak*, 21.4.2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_20200420_contact_tracing_covid_with_annex_en.pdf (access: 15.3.2023).

IBM, *What Is Predictive Analytics?*, https://www.ibm.com/topics/predictive-analytics (access: 15.3.2023).

Lohr S., *The Origins of 'Big Data': An Etymological Detective Story*, 1.2.2013, https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story (access: 15.3.2023).

Mudgal P., *Illegal Acquisition of Data/Information by Authorities, Apps and Social Media*, 16.10.2020, https://blog.ipleaders.in/illegal-acquisition-data-information-authorities-apps-social-media (access: 15.3.2023).

OPM U.S. Office of Personnel Management, *Cybersecurity Incidents*, https://www.opm.gov/cybersecurity/cybersecurity-incidents (access: 15.3.2023).

Predictice, https://predictice.com (access: 15.3.2023).

Rosenberg M., Confessore N., Cadwalladr C., *How Trump Consultants Exploited the Facebook Data of Millions*, 17.3.2018, https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (access: 15.3.2023).

Stanley H., *The Future of Personalization and How to Get Ready For It*, 20.10.2022, https://tiny.pl/cq52q (access: 15.3.2023).

Legal acts

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 (OJ L 281/31, 23.11.1995.

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) (OJ L 119/1, 4.5.2016).

Case law

Judgment of the ECJ of 19 October 2016, case C-582/14, *Patrick Breyer v. Bundesrepublik Deutschland*, ECLI:EU:C:2016:779.

ABSTRAKT

Niniejsze opracowanie poświęcone zostało analizie regulacji rozporządzenia Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych (RODO) w celu odpowiedzi na pytanie, czy we właściwy sposób wyważają one interesy zarówno podmiotów wykorzystujących w swojej działalności gospodarczej analizę predykcyjną i profilowanie, jak i osób, których dane są przez nich przetwarzane. Ze względu na to, że ten rodzaj przetwarzania opiera się na dużych zbiorach danych, właściwą analizę tego zagadnienia należało rozpocząć od określenia, jakie informacje przetwarzane w takich zbiorach i w jakich warunkach należy uznać za dane osobowe. W oparciu o te ustalenia przeprowadzona została analiza obowiązków nakładanych przez RODO na podmioty przetwarzające dane osobowe w sytuacji, gdy źródłem danych są informacje pozyskane ze zbiorów typu big data. Umożliwiło to dokonanie oceny przyjętych regulacji normatywnych oraz wskazanie możliwych rozwiązań i ścieżek rozwoju.

Słowa kluczowe: RODO; dane osobowe; profilowanie; big data; analiza predykcyjna