RUCH PRAWNICZY, EKONOMICZNY I SOCJOLOGICZNY
Rok LXXXI – zeszyt 1 – 2019

MARCIN ROJSZCZAK

DEFINICJA I GRANICE PRAWNEJ OCHRONY PRYWATNOŚCI W EPOCE ANALITYKI *BIG DATA*

I. UWAGI WPROWADZAJĄCE[1]

Po ponad stu latach od wprowadzenia do nauki prawa pierwszych definicji prywatności treść tego prawa, a co za tym idzie – i granice ochrony, nadal są przedmiotem dyskusji i sporów w doktrynie. W Powszechnej deklaracji praw człowieka (PDPC)[2] prywatność została zaliczona do katalogu podstawowych praw człowieka, co wiązało się między innym z nałożeniem na państwa obowiązku powstrzymania się od nieuprawnionej ingerencji (obowiązki negatywne), ale i zapewnienia poszanowania prywatności w relacjach horyzontalnych (obowiązki pozytywne). Współcześnie w większości państw wysoko rozwiniętych ochrona prywatności jest gwarantowana normami konstytucyjnymi. Także wszystkie najważniejsze międzynarodowe systemy ochrony praw człowieka uwzględniają prywatność w katalogu dóbr chronionych.

Pomimo powszechności ochrony zakres definicyjny prywatności nadal budzi kontrowersje. W redakcji przepisu art. 17 PDPC zdefiniowano prywatność za pomocą otwartego katalogu dóbr chronionych. W podobny sposób wyrażono treść prawa w innych systemach ochrony oraz przepisach konstytucyjnych. Takie podejście z jednej strony pozwala na dostosowanie zakresu prawa do zmieniających się potrzeb i oczekiwań społecznych, z drugiej jednak – zwiększa ryzyko stosowania interpretacji zbyt rozszerzającej, sprzecznej z *ratio legis*. Ekspansywność pojęcia „prywatność" jest zjawiskiem znajdującym szerokie omówienie zwłaszcza w nauce amerykańskiej, gdzie z uwagi na znaczenie precedensów treść prawa jest w dużej mierze kształtowana przez judykaturę[3].

Począwszy od lat siedemdziesiątych XX w. zagadnieniem dynamicznie zyskującym na znaczeniu jest ochrona danych osobowych. Impulsem do poszukiwania nowych środków ochrony prawnej była potrzeba zapewnienia ochrony prywatności jednostek w odniesieniu do danych przetwarzanych w systemach informatycznych. W początkowym etapie rozwoju nowe przepisy były postrzegane jako *lex specialis* względem regulacji dotyczących ochrony prywatności[4].

[1] Stan prawny oraz źródła internetowe aktualne na dzień 1.03.2019 r.

[2] Powszechna deklaracja praw człowieka z 10.12.1948 r., <http://cli.re/g21VMk>.

[3] Motyka (2010): 9–36.

[4] Wzajemne powiązanie pomiędzy prawem do prywatności a ochroną danych jest zagadnieniem badanym w nauce. Bez wątpienia nie są to pojęcia tożsame i należy zaakceptować pogląd,

https://doi.org/10.14746/rpeis.2019.81.1.8

Przepisy dotyczące ochrony danych ewoluowały od ogólnych wytycznych dla podmiotów przetwarzających do kompleksowych rozwiązań prawnych, definiujących cały model ochrony wzmacniany ustanowieniem właściwych organów nadzoru. Poszczególne generacje rozwiązań prawnych różniły się nie tylko szczegółowością praw i obowiązków, ale również odmienną rangą ustrojową zarówno przepisów materialnych[5], jak i instytucjonalnych[6]. Elementem niezmiennym było natomiast wyznaczenie materialnego zakresu za pomocą terminu „dane osobowe". Pojęcie to obejmuje każdą informację dotyczącą osoby o ustalonej tożsamości lub pozwalającą na jej identyfikację[7]. Przetwarzanie danych niepozwalających na identyfikację podmiotów co do zasady jest wyłączone spod stosowania regulacji prawnych dotyczących ochrony danych osobowych. Poszczególni prawodawcy mogą doprecyzowywać tę definicję, wskazując na przykład, że warunek identyfikowalności osoby jest spełniony niezależnie bądź że zbiór danych pozwala na wskazanie konkretnej osoby bezpośrednio (np. imię i nazwisko, numer PESEL) czy pośrednio (przez zbiór indywidualnych cech)[8].

Drugim ważnym terminem stosowanym w prawie ochrony danych już od lat siedemdziesiątych XX w. są tzw. szczególne kategorie danych (wcześniej określane jako dane wrażliwe). Są to informacje osobowe ujawniające pochodzenie rasowe, etniczne, poglądy polityczne, przekonania religijne, ale także stan zdrowia lub orientację seksualną. Definicja danych wrażliwych ewoluowała w kolejnych regulacjach prawnych i doprowadziła do objęcia szczególnym reżimem przetwarzania nowych kategorii informacji (np. dane o kodzie genetycznym bądź dane biometryczne). Pojęcie danych wrażliwych związane jest zwłaszcza z europejskim modelem ochrony danych. Z jego stosowaniem wiąże się wprowadzenie domyślnego zakazu przetwarzania z zamkniętym i enumeratywnie wymienionym katalogiem wyłączeń. Szczególny reżim ochronny przewidziany dla danych wrażliwych wyraża przekonanie prawodawców, że informacje, które mogą prowadzić do bezprawnej dyskryminacji, nie powinny być gromadzone i przetwarzane.

że żaden z tych terminów nie wyznacza zbioru nadrzędnego względem drugiego. W szczególności można wskazać przypadki naruszenia prawa do prywatności, niemające związku z przetwarzaniem danych osobowych. Analogicznie, bez trudu można wymienić przypadki naruszenia przepisów dotyczących ochrony danych, które *per se* nie przesądzają o ingerencji w prywatność jednostki (np. brak realizacji obowiązków informacyjnych względem podmiotu danych). W odniesieniu do zdarzeń mających miejsce w cyberprzestrzeni oba pojęcia często są stosowane zamiennie dla określenia prawa jednostki do realizacji jej autonomii informacyjnej.

[5] Współcześnie w europejskim modelu prawnym dane osobowe stanowią odrębny (niezależny) od prywatności przedmiot ochrony wymieniany w katalogu podstawowych praw człowieka. Por. np. art. 7 i 8 Karty praw podstawowych UE (Dz. Urz. UE 2016, nr C 202: 389) oraz art. 47 i 51 Konstytucji RP (Dz. U. Nr 78, poz. 483 ze zm.).

[6] Po reformie lizbońskiej kontrola przestrzegania prawa do ochrony danych osobowych ze strony niezależnego organu została zagwarantowana w przepisach traktatowych UE; por. art. 16 ust. 2 Traktatu o funkcjonowaniu UE (Dz. Urz. UE. 2016, nr C 202: 47).

[7] Nie obejmuje jednak danych osób zmarłych, z chwilą śmierci ustaje bowiem zdolność do czynności prawnych. Osoba zmarła nie może być zatem uznana za osobę fizyczną, a w efekcie dane jej dotyczące przestają być objęte zakresem stosowania unijnych przepisów o ochronie danych.

[8] Zob. np. art. 4 pkt 1 rozporządzenia 2016/679 z 27 kwietnia 2016 r. (Dz. Urz. UE 2016, nr L 119: 1).

Oparty na formalnej definicji danych osobowych oraz podziale na dane zwykłe oraz wrażliwe europejski model ochrony służy realizacji koncepcji autonomii informacyjnej, zgodnie z którą jednostki, kontrolując udostępnianie informacji na swój temat, mogą w sposób skuteczny limitować do nich dostęp, a w konsekwencji wyznaczać granicę ochrony własnej prywatności[9]. Zmiany technologiczne, w szczególności pojawienie się nowych form przetwarzania dużych zbiorów danych (*big data*), spowodowały, że to założenie straciło na aktualności. Technologia *big data* pozwala bowiem na przeprowadzenie identyfikacji jednostki nawet na podstawie informacji poddanych wcześniejszej anonimizacji, a także ustalenie wrażliwych danych osobowych na podstawie danych publicznie dostępnych. W konsekwencji upowszechnienia analiz *big data* konieczne jest ponowne przeanalizowanie, czy podstawy europejskiego modelu ochrony danych w sposób prawidłowy definiują zarówno przedmiot ochrony, jak i sposób jej zapewnienia. Dotychczasowe reformy przepisów – nie wyłączając ogólnego rozporządzenia o ochronie danych[10] – skutkowały rozbudowaniem uprawnień jednostek oraz zwiększeniem obowiązków po stronie podmiotów przetwarzających. Wydaje się, że nowe technologie – takie jak *big data* – mogą pozwolić na łatwe ominięcie istniejących przepisów prawnych (w tym także wynikających z ogólnego rozporządzenia). Jednocześnie, z uwagi na masowy charakter przetwarzania, technologie te wykorzystywane w sposób pozbawiony nadzoru i regulacji mogą stanowić poważne zagrożenie dla prywatności nie tylko pojedynczych osób, ale całych społeczności. Nie bez przyczyny magazyn „Science", wydając w roku 2015 numer specjalny poświęcony nowym zagrożeniom dla ochrony prywatności – w tym analityce *big data* – zatytułował go *The End of Privacy* [Koniec prywatności][11].

Celem niniejszego artykułu jest rozważenie, w jaki sposób upowszechnienie analityki *big data* może wpłynąć na dalszą ewolucję przepisów o ochronie danych, a przede wszystkim czy kluczowe założenia europejskiego modelu ochrony ustanowione kilkadziesiąt lat temu pozostają nadal aktualne.

II. PRZETWARZANIE DUŻYCH ZBIORÓW DANYCH

W literaturze brakuje powszechnie akceptowalnej definicji *big data*. Część autorów wskazuje na powiązanie *big data* ze spopularyzowaną w latach dziewięćdziesiątych XX w. analityką *data mining*[12]. Celem *data mining* było budo-

[9] Potrzeba zagwarantowania autonomii informacyjnej w relacjach jednostka–państwo wynika także z krajowych norm konstytucyjnych (art. 51 Konstytucji) i była wielokrotnie przedmiotem rozważań Trybunału Konstytucyjnego (zob. np. wyroki U 3/01, K 8/04, K 33/08).

[10] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (Dz. Urz. UE 2016, nr L 119: 1).

[11] *The End of Privacy*, Science 347(6221), 2015.

[12] Rubinstein (2013): 74.

wanie nieoczywistych wniosków na podstawie informacji pochodzących z dużych baz danych. W ten sposób możliwe było obrazowanie zależności i trendów, a w konsekwencji przewidywanie przyszłych zdarzeń. Technologia ta znalazła szerokie zastosowanie zwłaszcza w sektorze e-usług do prognozowania zachowań użytkowników (np. preferencji zakupowych). *Big data* pozwala na wykorzystanie tych samych zaawansowanych algorytmów, ale w odniesieniu do zbiorów powstałych w wyniku korelacji (połączenia) wielu dużych baz danych. Podstawą zarówno *data mining*, jak i *big data* jest proces zwanym odkrywaniem wiedzy, który można zdefiniować jako ujawnianie nowych informacji na podstawie danych już posiadanych. Wyniki analizy nie są jednak faktami, ponieważ ich prawdziwość zależna jest zarówno od jakości danych źródłowych, jak i poprawności wnioskowania (działania algorytmu).

Kluczowe dla zrozumienia potencjału *big data* jest zrozumienie specyficznych cech tej technologii:

- bazowania na dużych zbiorach danych o różnej jakości,
- wykorzystaniu algorytmów heurystycznych oraz uczenia maszynowego,
- efektu przyrostowego,
- braku pewności co do prawdziwości wniosków.

Jednym z głównych źródeł danych do analityki *big data* są informacje publicznie dostępne w Internecie. Wszyscy główni operatorzy e-usług, nie wyłączającą portali społecznościowych, świadczą usługi dostępu do jawnych informacji publikowanych przez swoich użytkowników. Dla przykładu każdego dnia w serwisie Twitter zamieszczanych jest ok. 400 mln wiadomości, na Facebooku – 350 mln zdjęć, a za pośrednictwem portalu YouTube oglądanych jest ok. 4 mld filmów[13]. Duża część z tych informacji dostępna jest za pośrednictwem specjalnego interfejsu programowego (API) dla zewnętrznych dostawców treści, agregujących dane na potrzeby *big data*. Informacje na temat aktywności użytkowników (w szczególności – ich wpisy) uzupełniane są przez operatorów serwisów o szereg dodatkowych znaczników (parametrów) ułatwiających ich dalszą analizę – takich jak: dokładny czas umieszczenia wpisu; geolokalizacja użytkownika w momencie opublikowania wpisu; język, w jakim wpis został dodany; interakcję z innymi użytkownikami itp. Możliwe jest nie tylko analizowanie treści zamieszczonych informacji, ale również badanie powiązań pomiędzy użytkownikami czy ich wzajemnych interakcji. Pamiętając o ogromnej popularności portali społecznościowych, analizy *big data* każdego dnia mogą być zasilane setkami milionów nowych wpisów.

Istotnym ograniczeniem, immamentnie związanym z tą technologią, jest jednak różna jakość informacji źródłowych. Informacje publikowane w portalach społecznościowych mogą zawierać informacje nieprecyzyjne lub nieprawdziwe, nierzadko błąd ten jest wprowadzany celowo i ma na celu dystrybucję fałszywych danych. Przedsiębiorstwa zajmujące się profesjonalnie dostarczaniem narzędzi do analityki *big data* (tzw. brokerzy danych) stosują zaawan-

[13] Social Media and Big Data, The Parliamentary Office of Science and Technology, POSTnote 2014, nr 460, <http://cli.re/GWpA1q>.

sowane algorytmy oczyszczania i poprawiania jakości danych źródłowych[14]. Algorytmy te w dużej części bazują na wykrywaniu i poprawianiu błędów dzięki korelacji informacji z innymi zbiorami danych[15]. Ponieważ są to także działania automatyczne, ich skuteczność jest ograniczona.

Skoro dane źródłowe zawierają informacje o różnej jakości, to błąd ten może być tylko zwiększony w wyniku dalszego przetwarzania. Oznacza to, że wnioski pochodzące z takich badań zawsze będą obarczone błędem nie mniejszym niż dane źródłowe. Jeżeli wnioski dotyczą prognoz dotyczących dużych grup społecznych (np. preferencje wyborcze, decyzje zakupowe, badania sondażowe), to możliwe konsekwencje wydają się mniej istotne, niż gdy celem działania algorytmu jest wnioskowanie w sprawach indywidualnych (np. badanie profili online osób ubiegających się o wydanie wizy wjazdowej[16] lub kandydatów do pracy[17]).

Z zagadnieniem jakości danych bezpośrednio wiąże się problem zastosowanych algorytmów analitycznych. Celem technologii *big data* jest poszukiwanie nieznanych wcześniej zależności w dużych zbiorach informacji, które mogą wskazywać na istnienie nowych zależności (odkrywanie wiedzy). Dlatego wykorzystywane algorytmy bazują na heurystyce i mechanizmach uczenia maszynowego. Efekt działania tego typu algorytmów nie może być łatwo przewidziany nawet przez ich twórców – nie tylko z uwagi na skalę analizowanych danych, ale również na to, że uzyskiwane wyniki zależą od wniosków z analiz historycznych.

Ponieważ źródła danych wykorzystywane w *big data* są stale uzupełniane (o potencjalnie nawet kilkaset milionów nowych wpisów dziennie), to z uwagi na sposób działania wykorzystywanych algorytmów interpretacja wcześniejszych danych może ulec zmianie w wyniku wprowadzenia nowych informacji. Ta cecha *big data* nazywana jest efektem przyrostowym i w praktyce oznacza, że wnioskowanie jest zmienne w czasie – im więcej danych jest dostarczanych, tym bardziej precyzyjne wyniki i mniejszy błąd. Uwzględniając wieloletnią dostępność danych z portali społecznościowych, można oczekiwać, że prognozy generowane przez *big data* będą wiernie odzwierciedlały preferencje i oczekiwania użytkowników. Hipoteza ta znajduje liczne potwierdzenia praktyczne, np. związane z tzw. aferą Cambridge Analytica[18].

Połączenie omówionych powyżej cech – a więc korzystanie z olbrzymich, stale aktualizowanych baz o różnej jakości danych z zaawansowaną analityką wykorzystującą algorytmy uczenia maszynowego i sztucznej inteligencji rodzi problem wiarygodnej oceny prawdziwości wniosków. Z uwagi na skom-

---

[14] Maletic, Marcus (2009).

[15] Rahm, Do (2000): 5–8.

[16] 14 Million Visitors to U.S. Face Social Media Screening, The New York Times 30.03.2018, <http://cli.re/LvzjyV>.

[17] This Artificial Intelligence Can Predict How You'll Behave at Work Based on Social Media, Inc. 3.11.2017, <http://cli.re/gxrqy1>.

[18] Historia sporu D. Carrolla z Cambridge Analytica, zakres informacji posiadanych przez Cambridge Analytica oraz podmioty powiązane (m.in. SCL Elections Limited), a także cel przetwarzania zob. Carroll (2018).

plikowanie realizowanych obliczeń oraz wielkość danych przy rozbudowanych badaniach (np. obejmujących miliony użytkowników) nie ma możliwości zweryfikowania, czy uzyskany wynik cząstkowy jest prawidłowy. Co więcej, próba oszacowania stopnia pewności (błędu) w zakresie wyniku także jest problematyczna z uwagi na fakt, że zakres badań zazwyczaj obejmuje zagadnienia trudno weryfikowalne w inny sposób. Dlatego w literaturze przedmiotu wskazuje się, że analityka *big data* ma cechy samo spełniającej się przepowiedni[19]. Dotyczy to zwłaszcza tzw. analityki predykcyjnej, a więc próby określenia przyszłych zachowań na podstawie danych historycznych. Analityka predykcyjna wykorzystywana jest nie tylko w zakresie profilowania preferencji zakupowych czy próby wpływania na sympatie polityczne, ale również w obszarze bezpieczeństwa publicznego (np. wykrywanie ekstremizmów, osób planujących działania terrorystyczne itp.).

III. KONSEKWENCJE ANALITYKI *BIG DATA* DLA AUTONOMII INFORMACYJNEJ JEDNOSTKI

Źródłem danych na potrzeby *big data* mogą być nie tylko informacje publicznie dostępne – na potrzeby analiz wykorzystywane mogą być także bazy gromadzone przez przedsiębiorców (np. dużych operatorów e-usług, takich jak Google czy Amazon), ale również organy władzy publicznej[20]. W dalszych rozważaniach omówiony zostanie problem wykorzystania w *big data* informacji publicznie dostępnych – czy to publikowanych przez samych użytkowników w serwisach internetowych (portale społecznościowe, a także blogi, fora użytkowników itp.), czy udostępnianych przez państwa lub przedsiębiorców po ich wcześniejszym odpersonalizowaniu (anonimizacji).

Możliwość korelowania wielu informacji, pochodzących z różnych baz danych, tworzy przestrzeń do ujawnienia nowych, niewystępujących w danych źródłowych informacji. W przypadku analizy odpersonalizowanych zbiorów danych taką nową informacją może być tożsamość danej osoby. W ten sposób zestawiając kilka zanonimizowanych zbiorów, możliwe jest przeprowadzenie ponownej identyfikacji danych podmiotów. Pierwsze tego typu badania zostały przeprowadzone w czasach poprzedzających rozwój technologii *big data* i pokazały, że posiadanie trzech informacji na temat osoby – kodu pocztowego, daty urodzenia oraz płci – jest wystarczające do identyfikacji 87% populacji mieszkańców Stanów Zjednoczonych[21].

W 2006 r. Netflix przeprowadził konkurs wśród swoich użytkowników, w ramach którego opublikowany został zbiór 100 mln anonimowych ocen dokonanych przez 500 tys. użytkowników serwisu. Baza ta została poddana analizie, w efekcie czego badacze wykazali możliwość ponownej identyfika-

[19] Hert, Lammerant (2016): 126.
[20] Mucha (2012): 394–398.
[21] Tene (2011): 18.

cji użytkowników. W tym celu wykorzystano informacje pochodzące z innych publicznie dostępnych systemów (w tym bazy Internet Movie Database, czy portali społecznościowych, takich jak Facebook). W podsumowaniu przeprowadzonych badań autorzy skwantyfikowali złożoność ponownej identyfikacji podmiotu danych w zależności od liczby korelowanych baz danych oraz jakości zawartych w nich informacji[22].

Kolejne interesujące wyniki pochodzą z opublikowanego w 2015 r. badania próbki 1,1 mln zanonimizowanych transakcji dokonanych kartami płatniczymi. Także w tym przypadku możliwe stało się przeprowadzenie identyfikacji ok. 90% osób dokonujących płatności. Powyższe przykłady mogą wskazywać, że anonimizacja dużych zbiorów danych w rzeczywistości nie prowadzi do trwałego pozbawienia zbiorów cech pozwalających na ustalenie tożsamości. Dane zawierają bowiem cały czas powiązania i identyfikatory logiczne[23], które po ich skorelowaniu z zewnętrznym źródłem danych mogą pozwolić na ustalenie tożsamości użytkownika, a w efekcie ujawnienie znaczenia informacji z bazy pierwotnej (poddanej wcześniejszej anonimizacji). Naturalnie istnieje możliwość usunięcia tego typu powiązań wewnętrznych (np. przez usunięcie unikalnych identyfikatorów użytkowników lub identyfikatorów transakcji), jednak dane przetworzone w ten sposób będą nieprzydatne do dalszych analiz. Jak słusznie zauważył Michael Mattioli, „dane pozbawione kontekstu mogą być także pozbawione sensu"[24].

Odrębnym zagadnieniem jest możliwość przeprowadzenia ponownej identyfikacji wyłącznie na podstawie danych pochodzących z mediów społecznościowych – a więc tylko na podstawie wpisów samych użytkowników. Problematyka ta ma kluczowe znaczenie praktyczne z uwagi na rozwijający się rynek brokerów danych i usług przez nich świadczonych. Arvind Narayanan i Vitaly Shamatikov wykazali, że korelowanie informacji publikowanych w dwóch tylko serwisach – Twitter i Flickr – wystarczy do ustalenia tożsamości osób posługujących się kontami anonimowymi. Do uzyskania tych informacji wykorzystano technikę poszukiwania trendów (podobieństw) w grafach znajomych na obu portalach. Profesjonalni dostawcy usług *big data* oferują dostęp do strumieni danych pochodzących z kilkudziesięciu e-usług, w tym wszystkich czołowych portali społecznościowych[25].

Wykazana w szeregu badań przydatność *big data* do ponownej identyfikacji prowadzi do kilku zasadniczych wniosków. Po pierwsze, autonomia informacyjna jednostki nie może być w pełni realizowana w sytuacji, gdy dostępna technologia pozwala na określenie jej cech indywidualnych i informacji prywatnych wyłącznie na podstawie publicznie dostępnych danych. Po drugie,

---

[22] Narayanan, Shmatikov (2008).

[23] Dla przykładu w bazie Netflix możliwe było wyodrębnienie ocen dokonanych przez jednego użytkownika. Im więcej było takich ocen, tym łatwiej badacze mogli przeprowadzić korelację tych informacji w sposób prowadzących do jednoznacznej identyfikacji użytkownika.

[24] Mattioli (2014): 547.

[25] Przykładem może być serwis PeekYou, który wykorzystuje informacje pochodzące z sześćdziesięciu różnych publicznych baz danych (portale społecznościowe, wiadomości, blogi) do tworzenia szczegółowych profili konsumenckich; Federal Trade Commission (2014): 9.

ochrona prywatności w cyberprzestrzeni nie może opierać się wyłącznie na przepisach o ochronie danych osobowych, ponieważ nawet zbiory niepozwalające na identyfikację podmiotów danych niosą ze sobą wartość informacyjną mogącą prowadzić do naruszeń w sferze prywatności. Gromadzenie i obrót dużymi bazami danych zawierającymi informacje na temat jednostek powinien być ograniczany i kontrolowany w sposób nawet bardziej rygorystyczny, niż wynika z przepisów o ochronie danych – ponieważ potencjalnie może prowadzić do naruszenia prywatności wielu milionów osób.

Wprowadzenie postulowanych przepisów wydaje się najlepszym rozwiązaniem, ograniczającym ryzyko związane z utratą kontroli nad rynkiem Big Data. We współczesnych demokracjach nie kwestionuje się konieczności stosowania interwencjonizmu w obszarach szczególnie istotnych dla funkcjonowania państwa. W dzisiejszym świecie informacje pełnią rolę dla gospodarki nie mniejszą, jak pieniądze sto lat temu. Skoro więc państwa uznały za konieczne wdrożenie mechanizmów nadzoru nad sektorem finansowym, podobne działania powinny być obecnie podjęte w zakresie nadzoru nad podmiotami prowadzącymi obrót dużymi zbiorami danych.

Odrębnym problemem wymagającym dalszej analizy jest zasadność trwania unijnego prawodawcy przy wyznaczaniu granic stosowania przepisów o ochronie danych od spełnienia warunku „identyfikowalności" podmiotu danych. Rynek danych ewoluuje bardzo szybko, *big data* obrazuje, w jaki sposób nowe algorytmy mogą wprowadzić analizę danych na nowy, nieznany wcześniej poziom szczegółowości. W funkcjonującym w UE modelu podmiot przetwarzający sam powinien określić, czy zastosowane formy przetwarzania pozwalają na identyfikację podmiotów danych – i jeżeli tak, dostosować się do wymagań wynikających z RODO. Przyjmując takie podejście za właściwe, prawodawca unijny musiał założyć, że podmioty zajmujące się przetwarzaniem danych będą zainteresowane zwiększaniem obciążeń prawnych ograniczających ich profesjonalną aktywność. Trafność takiej oceny wydaje się wysoce wątpliwa, a dalsze trwanie przy niej może prowadzić nie do wzmocnienia, ale wręcz osłabienia unijnych standardów ochrony prywatności w cyberprzestrzeni.

IV. KONSEKWENCJE ANALITYKI *BIG DATA* DLA OCHRONY ZBIORÓW WRAŻLIWYCH

Odrębnym zagadnieniem jest kwestia zasadności utrzymywania w europejskim modelu ochrony danych odrębnego reżimu przetwarzania dla danych wrażliwych. W ten sam sposób, w jaki *big data* pozwala na określenie tożsamości użytkowników, możliwe jest wnioskowanie na temat innych cech indywidualnych – w szczególności dotyczących stanu zdrowia, orientacji seksualnej bądź sympatii politycznych.

Przykładem mogą być badania przeprowadzone przez Latanyę Sweeney z wykorzystaniem publicznie dostępnej bazy usług medycznych. Stan Wa-

szyngton co roku publikuje bazę zawierającą zanonimizowane dane dotyczące świadczeń medycznych. Baza, która może być kupiona przez każdego zainteresowanego za 50 USD, zawiera informacje na temat wszystkich przeprowadzonych hospitalizacji, w tym wieku pacjenta, postawionej diagnozy, przeprowadzonych procedur, a także kosztów poszczególnych usług. Sweeney wykorzystała informacje prasowe dotyczące zdarzeń skutkujących potrzebą hospitalizacji poszkodowanych do deanonimizacji bazy świadczeń medycznych. W efekcie była ona w stanie określić tożsamość 43% analizowanych przypadków. Korzystając z podobnej techniki, australijscy naukowcy z Uniwersytetu w Melbourne określili listę świadczeń medycznych udzielonych siedmiu wybranym osobom publicznym, w tym trzem byłym i obecnym parlamentarzystom[26].

Omówienia wymaga możliwość uzyskania informacji wrażliwych – w tym na temat stanu zdrowia czy preferencji politycznych – wyłącznie na podstawie analizy aktywności samych użytkowników, bez korzystania ze zanonimizowanych rejestrów medycznych. Dla przykładu koncern Target wykorzystywał informacje na temat podejmowanych decyzji zakupowych do prognozowania, które klientki będą w ciąży. Technologia pozwalała także na określenie zaawansowania ciąży oraz wskazania przewidywanego terminu porodu. Informacje te były następnie wykorzystywane do promocji produktów przeznaczonych dla kobiet w określonym trymestrze ciąży[27].

Z innej techniki skorzystała polska firma Selectivv, która na podstawie informacji pochodzących z telefonów użytkowników (m.in. geolokalizacji oraz listy zainstalowanych aplikacji) przeprowadziła badanie uczestników Open'er Festival 2017. Wyłącznie na podstawie informacji pochodzących z obserwacji aktywności użytkowników – a więc bez przeprowadzania wywiadów – ustalono, że 11% uczestników należy do mniejszości seksualnych, a 14% to kobiety starające się o dziecko[28]. Jest to przykład dowodzący, że nawet informacja o aplikacjach zainstalowanych przez użytkownika na telefonie komórkowym w połączeniu z odpowiednio dużą próbą jest wystarczająca, aby przeprowadzić profilowanie użytkowników i określić indywidualne cechy jednostek.

Przykład wątpliwie etycznych analiz prowadzonych przez Cambridge Analytica obrazuje, w jaki sposób aktywności online użytkowników mogą być wykorzystane do badań sympatii politycznych. Należy wskazać, że casus Cambridge Analytica – chociaż spotkał się z wyjątkowym zainteresowaniem mediów – nie należy do wyjątków. Według dostępnych informacji Cambridge Analytica uzyskała dostęp do ok. 40–50 mln profili użytkowników portalu Facebook[29]. Na rynku brokerów danych od lat funkcjonują przedsiębiorcy zarządzający setkami milionów profili uzupełnianych danymi pochodzącymi z wielu portali społecznościowych. Dla przykładu Acxiom wskazuje, że zarzą-

---

[26] Teague, Culnane, Rubinstein (2017).

[27] Richards (2013): 1939–1940.

[28] Wyniki i opis metodyki badawczej: <http://cli.re/g9ZDm7>.

[29] Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach, The Guardian 17.03.2018, <http://cli.re/gmAZB7>.

dza bazą ok. 700 mln profili[30]. Według raportu Federalnej Komisji Handlu (Federal Trade Commission, FTC), podsumowującego badanie procedur gromadzenia i zarządzania danymi przez największych amerykańskich brokerów danych, jedną z informacji uwzględnianych w profilach jest wskazanie preferencji wyborczych. Nie ma żadnego powodu, aby uznać, że zakres analiz, jakie przeprowadziła Cambridge Analytica w odniesieniu do kilkudziesięciu milionów osób, nie może zostać wykonany przez Acxiom lub innych podobnych przedsiębiorców, z tą różnicą, że w odniesieniu do 500 mln użytkowników (populacji większej niż ludność UE). Zasady etyczne i moralne stosowane przez brokerów danych opierają się w wielu przypadkach na ich wewnętrznych regulacjach, a nie przepisach powszechnie obowiązującego prawa.

Po analizie znaczenia i roli, jaką unijny prawodawca przywiązuje do przetwarzania danych wrażliwych, pojawia się wątpliwość, w jakiej mierze deklaracje te mają faktyczne znaczenie w ochronie jednostek przed nieuprawnioną dyskryminacją. Globalny charakter rynku *big data* powoduje, że przedsiębiorcy bez problemu mogą kupować duże zbiory danych i przekazywać je do dalszego przetwarzania w jurysdykcjach pozbawionych ograniczeń prawnych. Wprowadzenie wysokiego reżimu przetwarzania dla przedsiębiorców europejskich nie prowadzi do zwiększenia ochrony przed naruszeniami ze strony podmiotów działających poza UE. Rynek analiz *big data* opiera się na dużych zbiorach danych pozyskiwanych od innych przedsiębiorców. Tego typu pośrednicy nie znają celu przetwarzania ani nie mają na niego wpływu. Ponieważ bazy te nie zawierają danych osobowych, nie muszą być chronione w sposób wynikający z ogólnego rozporządzenia dla przypadków transgranicznego przekazywania danych. Z kolei przedsiębiorca zagraniczny, nabywając tego typu duże zbiory zanonimozowanych informacji, może dzięki analityce *big data* przeprowadzić ponowną identyfikację lub określić cechy jednostkowe użytkowników w sposób, który spowoduje naruszenie ich prywatności.

V. PODSUMOWANIE

Upowszechnianie nowych technik przetwarzania informacji prowadzi do konieczności weryfikacji wielu paradygmatów leżących u podstaw prawa ochrony danych osobowych. Możliwość identyfikacji jednostki jako wyróżnik zbioru danych osobowych czy domyślny zakaz przetwarzania danych wrażliwych to zasady funkcjonujące w europejskim modelu ochrony prywatności od przeszło czterdziestu lat. Zostały ustanowione w czasach, gdy systemy informatyczne były wykorzystywane jako uzupełnienie przetwarzania tradycyjnego, opartego na kartotekach i papierowych zbiorach dokumentów. Obecnie skala przetwarzania prowadzona w wersji elektronicznej – zarówno w wymiarze geograficznym, jak i wolumetrycznym – przekracza wielokrotnie zakres przetwarzania realizowanego z wykorzystaniem dokumentacji papierowej.

[30] Federal Trade Commission (2014): 9.

Ponadto coraz więcej informacji dostępnych jest wyłącznie w postaci elektronicznej i dotyczy aktywności prowadzonych tylko w cyberprzestrzeni.

Dalsze trwanie przy formalnym podziale na dane zwykłe i wrażliwe może tworzyć błędne przeświadczenie o większych gwarancjach związanych z ochroną szczególnych kategorii danych. Przedstawione powyżej badania obrazują, że jest to przekonanie błędne. Dostępne współcześnie technologie – w szczególności analityka *big data* – pozwala na przeprowadzanie różnego typu badań prowadzących do ujawnienia chronionych faktów z życia setek milionów użytkowników Internetu wyłącznie na podstawie informacji publicznie dostępnych. Nie ma przy tym znaczenia, jak wiele z tych informacji zostało opublikowanych przez poszczególne osoby. Nie jest nawet istotne, czy dana osoba opublikowała cokolwiek – technologia *big data* jest równie przydatna w profilowaniu osób, które rzadko lub w ogóle nie korzystają z e-usług (wówczas wnioskowanie odbywa się wyłącznie na podstawie informacji pochodzących od innych użytkowników, a dotyczących analizowanej osoby oraz z licznych rejestrów publicznych). Słusznie wskazuje zatem Arwid Mednis, że stosowanie *big data* może prowadzić do naruszenia prywatności nawet wtedy, gdy dana osoba nie została poddana profilowaniu. Sam bowiem fakt podejmowania decyzji na podstawie podobieństwa jej cech do profili innych osób może pozwolić na przewidywanie jej zachowań, a w efekcie prowadzić do ograniczenia autonomii woli[31].

Swobodna wymiana dużych zbiorów danych, zawierających zanonimizowane informacje na temat aktywności użytkowników nieuchronnie prowadzi do zmniejszenia gwarancji, że prywatność tych osób zostanie poszanowana. Wprowadzone w Unii rozbudowane wymagania dla przedsiębiorców przetwarzających niewielkie zbiory danych wydają się nieadekwatne, gdy uwzględni się, że przepisy te nie obejmują działalności funkcjonujących w państwach trzecich podmiotów mogących swobodnie gromadzić i przetwarzać zbiory pozwalające na ustalanie szeregu prywatnych informacji na temat setek milionów użytkowników Internetu[32]. Ten dysonans, obrazujący lukę w istniejącym prawodawstwie, powinien zostać zlikwidowany przez wprowadzenie nowych przepisów regulujących obrót dużymi zbiorami danych.

Problem ten został dostrzeżony w Stanach Zjednoczonych, gdzie dwukrotnie przedkładano projekt ustawy federalnej regulującej działalność brokerów

---

[31] Mednis (2016).

[32] Chociaż zgodnie z art. 27 RODO zagraniczni administratorzy powinni wskazać na obszarze UE swojego przedstawiciela, a więc podmiot, do którego m.in. mogą być kierowane skargi i wnioski wynikające z rozporządzenia, to w rzeczywistości powstaje trudność, gdy administrator nie wykona tego obowiązku, a jednocześnie nie prowadzi działalności gospodarczej kierowanej bezpośrednio do użytkowników z obszaru UE. W takim bowiem przypadku istnieje ograniczona możliwość egzekwowania obowiązków publicznoprawnych (w tym kar pieniężnych) od administratora, który prowadzi działalność wyłącznie na obszarze obcej jurysdykcji. Także szereg innych szczegółowych obowiązków wynikających z RODO (np. dotyczących bezpieczeństwa przetwarzania) nie może być skutecznie weryfikowanych, np. na podstawie kontroli prowadzonych przez organ nadzorczy, a potwierdzenie ich spełnienia może przybrać formę w dużej mierze deklaratywną.

danych[33]. Z kolei w Australii zaproponowano wprowadzenie penalizacji czynności ponownej identyfikacji podmiotów danych w zanonimizowanych bazach publicznych[34]. Z kolei Komisja Europejska stoi na stanowisku, że wystarczającą regulację rynku *big data* stanowią przepisy ogólnego rozporządzenia[35]. Argumentacja ta, w odniesieniu do przedsiębiorców unijnych oraz części zagranicznych (zwłaszcza mających silne związki gospodarcze z rynkiem UE) adekwatna, jest jednak chybiona w odniesieniu do podmiotów funkcjonujących w państwach trzecich, niepodlegających europejskiej jurysdykcji, w tym możliwości władczego wpływania ze strony europejskich organów nadzoru[36].

Uwzględniając stale rosnące skomplikowanie e-usług, cyfryzację kolejnych obszarów życia oraz ogrom informacji publikowanych w Internecie, uzasadniony wydaje się pogląd, że upowszechnienie analityki *big data* (lub nowych, bardziej zaawansowanych technologii) sprawi, że dalsze funkcjonowanie znanych obecnie prawnych mechanizmów ochrony prywatności stanie się bezcelowe. Należy oczekiwać, że rozumienie prywatności będzie ewoluować i wprowadzi zamiast autonomii informacyjnej oraz wolności od obserwacji nowy zakres ochrony. Obecnie w nauce prawa trwa dyskusja na temat przyszłości prywatności – przedstawiane są różne propozycje, w jaki sposób ochrona prywatności powinna być realizowana w erze masowego przetwarzania dużych zbiorów danych[37]. Interesującą koncepcję przedstawił David Brin, proponując, aby prywatność nie była realizowana przez zapewnianie poufności informacji – ale wręcz przeciwnie, przez ich całkowitą jawność, połączoną wszakże z rozliczalnością, pozwalającą w łatwy sposób na ustalanie występujących nadużyć i osób za nie odpowiedzialnych[38]. Pewną modyfikacją ten propozycji jest postulat Wacława Iszkowskiego dotyczący wyłączenia spod prawnej ochrony

---

[33] Zob. projekt ustawy z 12 lutego 2014 r. o odpowiedzialności i transparentności działania brokerów danych (Data Broker Accountability and Transparency Act of 2014); <https://www.congress.gov/bill/113th-congress/senate-bill/2025>; projekt ustawy z 3 kwietnia 2015 r. o odpowiedzialności i transparentności działania brokerów danych (Data Broker Accountability and Transparency Act of 2015); <https://www.congress.gov/bill/114th-congress/senate-bill/668>.

[34] Projekt ustawy z 12 października 2016 r. o zmianie ustawy o ochronie prywatności z 1988 (Privacy Amendment Bill 2016), <http://cli.re/gYeeXo>.

[35] Komisja Europejska (2016).

[36] Należy rozróżnić sytuację, w której zagraniczny przedsiębiorca prowadzi działalność gospodarczą na terenie UE (posiada jednostkę organizacyjną, ustanowionego przedstawiciela, zasoby itp.), i sytuację, w której nie prowadzi takiej działalności, ale realizuje czynności przetwarzania danych, obejmujące także dane mieszkańców UE. W tym drugim przypadku trudno oczekiwać, aby przedsiębiorca zagraniczny w sytuacji kolizji norm prawa krajowego z regulacjami UE odrzucił przepisy krajowe i dostosował się do sprzecznych norm prawa UE. Rozwiązanie takie byłoby także niepraktyczne, ponieważ tworzyłoby barierę dla dalszego rozwoju e-usług. Przyjęcie modelu forsowanego przez UE mogłoby bowiem doprowadzić do sytuacji, w której każde państwo oczekiwałoby, że ustanowione przez nie normy krajowe w zakresie przetwarzania danych osobowych będą stosowane globalnie w odniesieniu do czynności przetwarzania dotyczących użytkowników podlegających danemu systemowi prawnemu. W rezultacie przedsiębiorcy unijni prowadzący działalność na terenie UE musieliby stosować chińskie prawo w zakresie przetwarzania danych użytkowników pochodzących z Chin lub brazylijskie w zakresie mieszkańców Brazylii.

[37] Zob. np. Tene, Polonetsky (2013): 256–272; Froomkin (2000): 1524–1539; Solove (2009).

[38] Brin (1998). Należy pamiętać, że koncepcja Brina nie jest powszechnie akceptowana i spotkała się także z krytyką – zob. np. Schneier (2008).

przetwarzania podstawowych danych osobowych[39]. Chociaż propozycja Brina została sformułowana ponad dwadzieścia lat temu, to nadal jest inspirującym przykładem wskazującym, że era prywatności nie skończy się wraz z globalizacją przetwarzania informacji – istniejące standardy ochrony bez wątpienia ewoluują, a początków tej zmiany można szukać w trwającej dyskusji na temat prawnej regulacji rynku *big data*.

Marcin Rojszczak
Uniwersytet Warszawski
marcin.rojszczak@gmail.com
https://orcid.org/0000-0003-2037-4301

Brin, D. (1998). The Transparent Society: Will Technology Force Us to Choose Between Privacy and Freedom? New York.
Carroll, D. (2018). Cambridge Analytica is dead, long live our data. Boston Review. <http://cli.re/g3mPAo> [dostęp: 1.03.2019].
Federal Trade Commission (2014). Data Brokers: A Call for Transparency and Accountability. Washington. <https://goo.gl/ig9tEp> [dostęp: 1.03.2019].
Froomkin, A. (2000). The death of privacy. Stanford Law Review 52(5): 1461–1543.
Grunebaum, M. (2015). Suicidology meets Big Data. Journal of Clinical Psychiatry 76(3): 383–384.
Iszkowski, W. (2016). O „nieciągłościach" ochrony danych osobowych, [w:] A. Mednis (red.), Prywatność a jawność. Bilans 25-lecia i perspektywy na przyszłość, Warszawa: 33–46.
Komisja Europejska (2016). The EU Data Protection Reform and Big Data. Brussels. <http://cli.re/LywNE5> [dostęp: 1.03.2019].
Lammerant H., Hert, P. de (2016). Predictive profiling and its legal limits: effectiveness gone forever? [w:] B. van der Sloot, D. Broeders, E. Schrijvers (eds.), Exploring the Boundaries of Big Data. Amsterdam: 145–168.
Lazer, D., Kennedy R., King G., Vespignani A. (2014). The parable of Google flu: traps in Big Data analysis. Science 343(6176): 1203–1205.
Maletic, J., Marcus, A. (2010). Data cleansing: a prelude to knowledge discovery, [w:] O. Maimon, L. Rokach (eds.), Data Mining and Knowledge Discovery Handbook. Boston: 19–32.
Mattioli, M. (2014). Disclosing Big Data. Minnesota Law Review 99(2): 535–583.
Mednis, A. (2016). Prywatność od epoki analogowej do cyfrowej – czy potrzebna jest redefinicja? [w:] Prywatność a jawność. Bilans 25-lecia i perspektywy na przyszłość. Warszawa: 3–16.
Motyka, K. (2010). Prawo do prywatności, Zeszyty Naukowe Akademii Podlaskiej w Siedlcach, Seria Administracja i Zarządzanie 85(12): 9–36.
Mucha, B. (2012). Data mining a współczesny kształt prawa do prywatności w Stanach Zjednoczonych Ameryki, [w:] J. Jaskiernia (red.), Efektywność europejskiego systemu ochrony praw człowieka. Ewolucja i uwarunkowania europejskiego systemu ochrony praw człowieka. Toruń: 392–436.
Narayanan, A., Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. <http://cli.re/L9DkB2> [dostęp: 1.03.2019].
Teague V., Culnane C., Rubinstein B. (2017). Health Data in an Open World. <https://arxiv.org/abs/1712.05627> [dostęp: 1.03.2019].

[39] Iszkowski wymienia w katalogu danych podstawowych: imiona, nazwisko, identyfikator personalny, wizerunek (zdjęcie paszportowe), adres mailowy oraz numer telefonu. Jednocześnie postuluje wzmocnienie „technicznej, organizacyjnej i prawnej" ochrony wrażliwych danych osobowych. Realizacja koncepcji Iszkowskiego w praktyce skutkowałaby dalszym rozwojem rynku *big data*: zniesienie ograniczeń w zakresie przetwarzania danych identyfikujących jednostkę doprowadziłoby do likwidacji jednej z istotnych barier rozwoju tej technologii – a więc ograniczonej jakości danych. Szczegóły: Iszkowski (2016).

Tene, O., Polonetsky, J. (2013). Big Data for all: privacy and user control in the age of analytics. Northwestern Journal of Technology and Intellectual Property 11(5): 240–273.
Rahm, E., Do, H. (2000). Data Cleaning: Problems and Current Approaches. <http://cli.re/GWprwd> [dostęp: 1.03.2019].
Richards, N. (2013). The dangers of surveillance. Harvard Law Review 126(7): 1934–1965.
Rubinstein, I. (2013). Big Data: the end of privacy or a new beginning? International Data Privacy Law 3(2): 74–87.
Schneier B. (2008). The myth of the 'transparent society'. Wired. <http://cli.re/gBonaa> [dostęp: 1.03.2019].
Solove, D. (2009). Privacy: a new understanding, [w:] Understanding Privacy. Cambridge–London: 171–198.
Tene, O. (2011). Privacy: the new generations. International Data Privacy Law 1(1): 15–27.

THE DEFINITION AND LIMITS OF DATA PROTECTION LAWS IN THE ERA OF BIG DATA ANALYTICS

Summary

More than one hundred years after the first definitions of the right to privacy, the content of this right and the limits of its protection are still being discussed and disputed in the doctrine. The protection of human rights tends to define privacy by determining an open list of protected values. At the same time, in data protection law the scope of regulation is determined by terms 'personal data' and 'special categories of data'. The definition of these terms has remained unchanged for over thirty years. The division of vertical and horizontal intrusions in the area of privacy protection in cyberspace is no longer valid. The activities of public authorities and specialized entities such as data brokers have been increasingly complementing one another. Collecting vast amounts of data about hundreds of millions of users may lead to privacy intrusions not only of individuals, but also of entire societies. The purpose of this article is an attempt to determine whether the legal regulations already in force and being implemented, based on the definition of personal data adopted in the pre-Internet era, have the potential to effectively protect against the risks associated with modern data processing techniques such as Big Data. To achieve this goal, the most important features of Big Data are discussed, such as algorithmic knowledge building or incremental effect, and it is also explained how this technology allows legal restrictions related to the processing of different categories of personal data to be bypassed. In the summary, a postulate to develop regulations dedicated to regulating the market for the processing of large data sets is formulated.

Keywords: right to privacy; personal data; Big Data; re-identification; special categories of data