ალექსანდრე ბოდუინი[*]
ანდრეა ფ. ფერარისი[**]

სინთეზური მონაცემების, როგორც ანონიმიზაციის სტრატეგიის სამართლებრივი და მარეგულირებელი პერსპექტივები

მზარდ ციფრულ სამყაროში პერსონალური მონაცემების დაცვა უმნიშვნელოვანესია, როგორც ფიზიკური და იურიდიული პირებისთვის, ისე მარეგულირებელი ორგანოებისთვის. მონაცემთა შეგროვების ტექნოლოგიების განვითარებასთან ერთად, ასევე, უნდა გაუმჯობესდეს მონაცემთა კონფიდენციალობის უზრუნველყოფის მეთოდები. წინამდებარე ნაშრომი იკვლევს სინთეზურ მონაცემებს, როგორც პერსპექტივას კონფიდენციალობის გამაძლიერებელი ტექნოლოგიის ("PET") ანონიმიზაციის მიზნით. ნაშრომი ფოკუსირებულია იურიდიულ, თეორიულ და პრაქტიკულ ასპექტებზე, სადაც განხილულია მარეგულირებლის კონტექსტი, განსაკუთრებით "GDPR"-ის მიხედვით, და განსაზღვრულია კონფიდენციალობის რისკები და გამოწვევები, რომლებიც ანონიმიზაციით უნდა იყოს დაცული. ჩვენ ვამტკიცებთ, რომ სინთეზურ მონაცემებს, როდესაც ისინი სათანადოდ გენერირებულია, შეუძლიათ დააკმაყოფილონ ანონიმიზაციის სტანდარტები და უზრუნველყონ განლაგების რეკომენდაციები კონფიდენციალობის რისკების შესამცირებლად. ჩვენი დასკვნები ხელს უწყობს მონაცემთა სინთეზური კონფიდენციალობის უზრუნველყოფის სტანდარტიზებულ ჩარჩოს, რომელიც შეესაბამება მონაცემთა დაცვის მიმდინარე და მომავალ რეგულაციას.

საკვანძო სიტყვები: *სინთეზური მონაცემი, ანონიმიზაცია, კონფიდენციალობის გამაძლიერებელი ტექნოლოგია, "GDPR".*

[*] სამართლის დოქტორი; "Aindo SpA"-ში პირადი ცხოვრების ხელშეუხებლობის დარგის მკვლევარი.
[**] ბოლონიის უნივერსიტეტის დოქტორანტი და "Data Valley Consulting srl"-ში სამართლის, მეცნიერებისა და ტექნოლოგიების მკვლევარ-სპეციალისტი.

1. შესავალი

მუდმივად განვითარებად ციფრულ სამყაროში პერსონალური მონაცემების დაცვა სულ უფრო და უფრო დიდ გამოწვევას წარმოადგენს, როგორც ინდივიდების, ისე ორგანიზაციებისა და მარეგულირებელი ორგანოებისთვის. მონაცემთა შეგროვების და დამუშავების ტექნოლოგიების მუდმივად განვითარების პირობებში, აუცილებელია, ასევე, გაუმჯობესდეს პერსონალური ინფორმაციის დაცვის საშუალებები.

ერთ-ერთი ასეთი საშუალებას წარმოადგენს ანონიმიზაცია, რომელიც მიზნად ისახავს მონაცემის იმგვარ გარდაქმნას, რომლის შედეგადაც შეუძლებელი გახდება მონაცემთა სუბიექტის იდენტიფიცირება. ანონიმიზაციის მექანიზმი აუცილებელია ევროკავშირის მონაცემთა დაცვის ძირითად რეგულაციასთან ("GDPR") შესაბამისობის უზრუნველსაყოფად, იგი, ასევე, იცავს სუბიექტთა კონფიდენციალობას მათი მონაცემების ბოროტად გამოყენების შემთხვევაში.

წინამდებარე ნაშრომი იკვლევს სინთეზურ მონაცემს, როგორც კონფიდენციალობის გამაძლიერებელ ტექნოლოგიას ("PET")[1], უფრო კონკრეტულად ნაშრომი განიხილავს ანონიმიზაციას იურიდიული და თეორიული თვალსაზრისით.

სინთეზური მონაცემები არ გროვდება ემპირიულად, არამედ გენერირდება ალგორითმების მეშვეობით. ამასთან, ისინი არ არის დაკავშირებული კონკრეტულ პირთან, მაგრამ, შესაძლოა, სასარგებლო იყოს მონაცემთა შესახებ მეცნიერებისა და ხელოვნური ინტელექტის ("AI") განვითარებისთვის. ტექნოლოგიებთან და კონფიდენციალობის კონცეფციასთან დაკავშირებულ ნაშრომში მოცემული პერსპექტივები და ინტერპრეტაციები უზრუნველყოფს სინთეზური მონაცემების კონფიდენციალობის სტანდარტიზებული და მყარი ჩარჩოს შექმნას.

ნაშრომი სტრუქტურირებულია შემდეგნაირად: მეორე თავში განხილულია ანონიმიზაციის საკანონმდებლო განმარტება "GDPR"-ის შესაბამისად. უფრო კონკრეტულად, მოცემულია პერსონალური მონაცემის, ანონიმიზებული მონაცემის შესახებ ანალიზი, ასევე, განხილულია ანონიმიზაციის მოთხოვნები, მათ შორის, შესაძლო თავდასხმის ტიპები. მესამე თავში, განხილულია იურიდიული თეორია იმის საჩვენებლად, რომ შესაბამისი ტექნოლოგიების სწორი გამოყენების პირობებში, რეალური მონაცემებისაგან სინთეზური მონაცემების შექმნა გენერირებადი ხელოვნური ინტელექტის მეშვეობით შეიძლება იყოს ანონიმიზაციის ერთ-ერთი საშუალება. ნაშრომის მეოთხე თავში ჩვენ ვიყენებთ შემუშავებულ კონცეფციებს, რათა გავცეთ რეკომენდაციები სინთეზური მონაცემების სწორი გამოყენების შესახებ კონფიდენციალობასთან დაკავშირებულ პოტენციურ საფრთხეებსა და ღია კვლევებზე დაყრდნობით.

2. ანონიმიზაციის კონცეფცია

აღნიშნულ თავში მოცემულია ანონიმიზაციის კონცეფცია, როგორც ეს აღწერილია წინა ნაშრომში[2]. ანონიმიზაცია არის პერსონალური მონაცემების

[1] OECD, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches," OECD Digital Economy Papers (OECD, 2023).

[2] *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., & Chauvenet C. R.,* Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: *Hideyuki Matsumi, Paul De Hert et al. (ed)*, Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.

ტრანსფორმაციის პროცესი, რომლის შედეგადაც, მონაცემთა ბაზაში პირდაპირ თუ ირიბად შეუძლებელია ფიზიკური პირის იდენტიფიცირება. აღნიშნული ტექნიკა წარმოადგენს კონფიდენციალობის დაცვის აუცილებელ წინაპირობას მონაცემებზე ორიენტირებულ ("data-centric") სამყაროში, ამასთან, იგი შესაბამისობაშია "GDPR"-თან. იდენტიფიცირებადი ელემენტების ეფექტურად წაშლით ან შეცვლით, ანონიმიზაცია პერსონალურ ინფორმაციას იცავს არაავტორიზებული წვდომისა და არასათანადო გამოყენებისგან, ამავე დროს იძლევა მონაცემთა ანალიზისა და შემდგომი გამოყენების საშუალებას. მონაცემთა ანონიმიზაციის სხვადასხვა მიდგომების შესწავლამდე, მნიშვნელოვანია, პირველ რიგში, გავაანალიზოთ პერსონალური მონაცემების მახასიათებლები, "GDPR"-ის შესაბამისად.

2.1. პერსონალური მონაცემი

ევროკავშირის მონაცემთა დაცვის ძირითადი რეგულაცია ("GDPR")[3] პერსონალურ მონაცემს განმარტავს, როგორც „ნებისმიერ ინფორმაციას, რომელიც იდენტიფიცირებულ ან იდენტიფიცირებად ფიზიკურ პირს („მონაცემთა სუბიექტს") უკავშირდება. ფიზიკური პირი იდენტიფიცირებადია, როდესაც შესაძლებელია მისი იდენტიფიცირება პირდაპირ ან არაპირდაპირ, მათ შორის, სახელით, გვარით, საიდენტიფიკაციო ნომრით, ადგილმდებარეობით, ელექტრონული კომუნიკაციის მაიდენტიფიცირებელი მონაცემებით, ფიზიკური, ფიზიოლოგიური, ფსიქიკური, ფსიქოლოგიური, გენეტიკური, ეკონომიკური, კულტურული ან სოციალური მახასიათებლით".

აღნიშნული განმარტება მოიცავს ოთხ მნიშვნელოვან ელემენტს, როგორც ეს მოცემულია ლოპეზის და ელბის[4] ანალიზში:

1. „ნებისმიერი ინფორმაცია": ეს ტერმინი ყოვლისმომცველია, მოიცავს ყველა სახის ინფორმაციას პიროვნების შესახებ, განურჩევლად მისი ბუნებისა (ობიექტური თუ სუბიექტური) ან იმ კონტექსტისა, რომელშიც ადამიანი მოქმედებს (მაგ., როგორც მომხმარებელი, პაციენტი, თანამშრომელი). ის მოიცავს ფართო სპექტრს, მათ შორის, განსაკუთრებული კატეგორიის მონაცემებს, ასევე, ზოგად ინფორმაციას პირადი, ოჯახის ან პროფესიული ცხოვრების შესახებ.
2. „უკავშირდება": ინფორმაცია პიროვნებას უკავშირდება სამ გზით: მისი შინაარსით (კონკრეტული მონაცემები ინდივიდის შესახებ, როგორიცაა: სამედიცინო ჩანაწერები), მიზნით (თუ გამოიყენება ინდივიდის სტატუსის ან ქცევის შესაფასებლად ან გავლენას ახდენს მასზე) და შედეგით (თუ მისი გამოყენება გავლენას ახდენს ინდივიდის უფლებებსა და ინტერესებზე).
3. „იდენტიფიცირებული ან იდენტიფიცირებადი": აღნიშნული ფოკუსირდება ჯგუფში პიროვნების გარჩევის უნარზე სხვადასხვა იდენტიფიკატორის საშუალებით, რომლებიც შეიძლება იყოს პირდაპირი (მაგ.: სახელი) ან არაპირდაპირი (მაგ.: ინფორმაციის უნიკალური კომბინაცია). კანონმდებლობა ვრცელდება ნებისმიერ პერსონალურ მონაცემზე, იდენტიფიკაციის მეთოდის მიუხედავად.

[3] 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) (Text with EEA relevance), (GDPR) art. 4.1.

[4] *López C.A.F., Elbi A.*, On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.

4. „ფიზიკური პირი“: კანონმდებლობა ვრცელდება და იცავს მოიცავს ყველა ადამიანს, რომელთა იდენტიფიცირება შესაძლებელია.

აღნიშნული კომპონენტები ერთობლივად უზრუნველყოფენ მონაცემთა დაცვისათვის საჭირო თანმიმდევრულ და ყოვლისმომცველ მიდგომებს. ამასთან, ეს კომპონენტები აბალანსებს პერსონალური მონაცემების ფართო ასპექტს, მათ შორის, ინდივიდის იდენტიფიცირების ხარისხს.

2.2. ანონიმიზაცია და ანონიმიზებული მონაცემი

ანონიმიზებული მონაცემი გულისხმობს ინფორმაციას, რომელიც არ შეიცავს პირის იდენტიფიცირების ელემენტებს, რაც უზრუნველყოფს, რომ პიროვნების ამოცნობა შეუძლებელია პირდაპირ ან მესამე პირებისთვის ხელმისაწვდომი დამხმარე მონაცემების საშუალებით. ანონიმიზებულ მონაცემზე არ ვრცელდება “GDPR”, გამომდინარე იქიდან, რომ არ წარმოადგენს პერსონალურ მონაცემს. ამასთან, აღსანიშნავია, რომ ანონიმიზაცია იცავს ინდივიდების კონფიდენციალობას და მონაცემის უკანონო ან არაეთიკური გამოყენებისგან.[5] “GDPR”-ის 26-ე მუხლი უფრო ღრმად იკვლევს პერსონალურ მონაცემებს, რაც თვალსაჩინოს ხდის ანონიმიზაციის კონცეფციას. აღნიშნული მუხლის თანახმად, ანონიმიზებული მონაცემი არის მონაცემთა იმგვარი დამუშავების შედეგი, როდესაც შეუძლებელია იდენტიფიცირებულ ან იდენტიფიცირებად მონაცემთა სუბიექტთან მათი დაკავშირება.

ხელახალი იდენტიფიკაციის, ან ანონიმიზებული მონაცემების პერსონალურ მონაცემებად გადაქცევის ალბათობა ფასდება არსებული მონაცემთა ბაზის სპეციფიური მახასიათებლებიდან გამომდინარე. აღნიშნული, შესაძლოა, განხორციელდეს მონაცემთა დადარების ან სხვა მსგავსი მეთოდების გამოყენებით. “GDPR”-ის 26-ე მუხლის თანახმად, იმის დასადგენად, შესაძლებელია თუ არა პირის იდენტიფიცირება, მხედველობაში უნდა იქნას მიღებული ყველა ის მეთოდი, რომელიც შესაძლოა გამოიყენოს, როგორც დამუშავებისთვის პასუხისმგებელმა პირმა, ისე სხვა ნებისმიერმა მესამე პირმა.

„იმისათვის, რომ დადგინდეს არის თუ არა ფიზიკური პირი იდენტიფიცირებადი, მხედველობაში უნდა იქნას მიღებული ყველა ის საშუალება, რომელიც გონივრულად იქნება გამოყენებული ფიზიკური პირის პირდაპირ ან ირიბად იდენტიფიცირებისთვის, როგორც დამუშავებისთვის პასუხისმგებელი პირის, ისე სხვა ნებისმიერი მესამე პირის მიერ.

მოცემულ შემთხვევაში „გონივრული“ ცნება გადამწყვეტია, იგი გულისხმობს პიროვნების იდენტიფიცირებისთვის გამოყენებულ საშუალებებს, რაც შეიძლება მოიცავდეს მნიშვნელოვან ხარჯებს, დროსა და მოქმედებას. შეფასებისას მხედველობაში უნდა იქნას მიღებული, როგორც მონაცემთა დამუშავებისას არსებული და ხელმისაწვდომი, ისე შესაძლო სამომავლო ტექნოლოგიები.[6]

---

[5] *Foglia C.*, Il Dilemma (ancora aperto) dell’Anonimizzazione e il Ruolo della Pseudonimizzazione nel GDPR, in Circolazione e Protezione dei Dati Personali, tra Libertà e Regole del Mercato. Commentario al Regolamento UE n. 2016/679 (GDPR) e al Novellato d.lgs. n. 196/2003 (Codice Privacy), 2019. *Stalla-Bourdillon S., Knight A.,* Anonymous Data v. Personal Data-False Debate: an EU Perspective on Anonymization, Pseudonymization and Personal Data, Wisconsin International Law Journal, 2017.

[6] *Tempestini L., D’Acquisto G.,* Il dato personale oggi tra le sfide dell’anonimizzazione e le tutele rafforzate dei dati sensibili,in Le nuove frontiere della privacy nelle tecnologie digitali, Aracne, 2016.

მონაცემთა დაცვის კანონმდებლობით აღიარებული პრინციპები არ ვრცელდება ანონიმიზებულ მონაცემებზე იმ შემთხვევაშიც თუკი მონაცემთა სუბიექტის ინდენტიფიცირება არაპროპორციულად დიდ ძალისხმევას ან/და ხარჯებს საჭიროებს. ამგვარად, ანონიმიზებული მონაცემიდან პირის იდენტიფიცირების მთავარ გამოწვევას იდენტიფიცირების პოტენციური ელემენტების არსებობა წარმოადგენს.[7]

29-ე მუხლის სამუშაო ჯგუფის ("WP29") მოსაზრებაში 4/2007 (Opinion 4/2007), სიღრმისეულადაა განხილული პერსონალური მონაცემის კონცეფცია, მათ შორის, გაანალიზებულია „იდენტიფიცირებადობის" მნიშვნელობები[8], აღნიშნულის თანახმად, ანონიმიზაციის პროცესის მდგრადობა იზომება „გამოსაყენებელი საშუალების გონივრულობის" ("means reasonably to be used") ტესტით.

იმის დასადგენად, არის თუ არა გამოყენებული საშუალება „გონივრული" ფიზიკური პირის იდენტიფიცირებისთვის, მხედველობაში უნდა იქნას მიღებული ყველა ობიექტური ფაქტორი, მათ შორის, იდენტიფიკაციისთვის საჭირო ხარჯები და დრო. ამასთან, უნდა შეფასდეს, როგორც მონაცემთა დამუშავებისას არსებული და ხელმისაწვდომი, ისე შესაძლო სამომავლო ტექნოლოგიები.

„მიზანშეწონილობის" კრიტერიუმი განსაკუთრებით აქტუალურია სტატისტიკური მონაცემების კონტექსტში. თუკი შეჯამებული მონაცემი, მისი მცირე ზომის ან სხვა საიდენტიფიკაციო ინფორმაციის ხელმისაწვდომობის გამო, პოტენციურად იწვევს პირის იდენტიფიცირებას, რაც ხაზს უსვამს ანონიმიზაციის პროცესის არაადეკვატურობას.[9]

ეფექტიანი ანონიმიზაცია ხელს უშლის ნებისმიერ ორგანიზაციას, რომ მათ მიერ მონაცემთა ბაზაში მოხდეს ინდივიდის იზოლირება ან მოხდეს ორი ჩანაწერის დაკავშირება. ანონიმიზაციის უზრუნველსაყოფად მხოლოდ პირდაპირი იდენტიფიკატორების ამოღება არ კმარა.[10] ხშირ შემთხვევაში საჭიროა დამატებითი ზომების მიღება, რომელთა ბუნება დამოკიდებულია მონაცემთა დამუშავების კონტექსტსა და მიზანზე. მნიშვნელოვანია, ასევე, ანონიმიზაციისათვის საჭირო ძალისხმევისა და ხარჯების შედარება ინდივიდების იდენტიფიცირების მზარდ ტექნიკურ შესაძლებლობებთან, მონაცემთა საჯარო ხელმისაწვდომობასა და არასრული ანონიმიზაციის შემთხვევებთან.[11]

ტერმინი „იდენტიფიცირებადი" მიემართება ნებისმიერ პირს, რომელიც ამუშავებს მონაცემებს და არა მხოლოდ დამუშავებისთვის პასუხისმგებელ თავდაპირველ პირს. თუ დამუშავებისთვის პასუხისმგებელი პირი ინახავს იდენტიფიცირებად ნედლ მონაცემებს და აზიარებს შეცვლილ ვერსიას, აღნიშნული მაინც კლასიფიცირდება, როგორც პერსონალური მონაცემი. ამის საწინააღმდეგოდ თუ დამუშავებისთვის პასუხისმგებელი პირი მონაცემს სრულად არაიდენტიფიცირებადს გახდის და მხოლოდ საერთო სტატისტიკისთვის გააზიარებს, აღნიშნული ანონიმიზებულ მონაცემად ჩაითვლება. ანონიმიზაციის

[7] *D'Acquisto G., Naldi M.*, Big Data e Privacy by Design, Anonimizzazione, Pseudonimizzazione, Sicurezza, Giappichelli, 2017.

[8] *Article 29 Data Protection Working Party (WP29)*, Opinion 4/2007 on the Concept of Personal Data, 2007.

[9] *Information Commissioner's Office (ICO)*, Anonymisation: Managing Data Protection Risk Code of Practice, 2012.

[10] *Agencia Española de Protección de Datos (AEDP) and European Data Protection Supervisor (EDPS)*.

[11] Misunderstanding related to Anonymization, 2021, <https://www.edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en>[31.07.2024].

ტექნიკის გამოყენებისას დამუშავებისთვის პასუხისმგებელმა პირმა უნდა შეაფასოს არჩეული მეთოდის სანდოობის დონე არსებულ ტექნოლოგიურ საშუალებებთან მიმართებით.

2.3. კონფიდენციალობის დარღვევის და თავდასხმების სახეები

29-ე მუხლის სამუშაო ჯგუფის ("WP29") მიერ იდენტიფიცირებულია თავდასხმის სამი ძირითადი ტიპი ანონიმიზაციასთან მიმართებით: „გამორჩევა" (Singling Out), „დაკავშირება" ("Linkability"), „ვარაუდი" ("Inference")[12]:

- „გამორჩევა" ("Singling Out"): გულისხმობს ზოგიერთი ან ყველა ჩანაწერის იზოლირების შესაძლებლობას, რომლითაც იდენტიფიცირდება ინდივიდი მონაცემთა ბაზაში. აღნიშნული მოიცავს ინდივიდის ჯგუფიდან გამორჩევის უნარს, რამაც შეიძლება გამოიწვიოს იდენტიფიკაცია, თუნდაც ინდივიდის ზუსტი ვინაობა უცნობი იყოს.
- Singling Out თავდასხმის ტიპი ყურადსაღები გახდა 1990-იან წლებში, როდესაც "MIT"-ის სტუდენტმა, ლატანია სვინიმ, მოიპოვა მასაჩუსეტსის გუბერნატორის უილიამ უელდის,[13] სავარაუდოდ, ანონიმიზებული ჯანმრთელობის შესახებ ცნობები. მან შეადარა ეს ჩანაწერები ამომრჩეველთა საჯარო რეგისტრაციის მონაცემებთან, და აღნიშნა, რომ "ZIP" კოდი, დაბადების თარიღი და მხოლოდ სქესი აშშ-ს მოსახლეობის აბსოლუტურ უმრავლესობის იდენტიფიცირების შესაძლებლობას იძლევა.
- „დაკავშირება" ("Linkability") - აღნიშნული გულისხმობს ერთ მონაცემთა სუბიექტთან, ასევე ერთ ან ორ მონაცემთა ბაზაში არსებულ სუბიექტთა ჯგუფთან დაკავშირებული მინიმუმ ორი ჩანაწერის ერთმანეთთან დაკავშირების შესაძლებლობას. ეს ნიშნავს, რომ თუ ცალ-ცალკე შეგროვებული ინფორმაციის სხვადასხვა ნაწილის ერთმანეთთან დაკავშირებით შესაძლებელია პირის იდენტიფიცირება, ანონიმიზაციის პროცესი არაეფექტურია.
- „Linkability"-ის თავდასხმის ტიპი აქტუალური გახდა 2006 წელს „Netflix"-ში მომხდარი ინციდენტის დროს. კერძოდ, მკვლევრებმა ნარაიანანმა და შმატიკოვმა „Netflix"-ის მონაცემთა ბაზაში არსებული მომხმარებლის ანონიმიზებული მონაცემები დაუკავშირეს „Internet Movie Database" ("IMDb")-ში არსებულ საჯაროდ ხელმისაწვდომ მონაცემებს. მათ შეძლეს იმ ანონიმური მომხმარებლების იდენტიფიცირება, რომლებმაც "IMDb"-ზე თავიანთი სახელებით გამოაქვეყნეს ფილმების რეიტინგები. აღნიშნული ცხადყოფს, რომ ერთი შეხედვით ერთმანეთთან დაუკავშირებელმა ინფორმაციამ შეიძლება საფრთხე შეუქმნას მონაცემთა ანონიმიზაციას.
- „ვარაუდი" ("Inference") - სამართლებრივ კონტექსტში, ის ეხება ინდივიდის შესახებ უცნობი ინფორმაციის ცნობილი ინფორმაციიდან ამოღების შესაძლებლობას. "Inference" თავდასხმის ტიპი სარგებლობს მონაცემების

[12] Article 29 Data Protection Working Party (WP29). "Opinion 05/2014 on Anonymisation Techniques", 2014

[13] *Barth-Jones D.*, The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now, 2012. *Narayanan A., Shmatikov V.*, How to Break Anonymity of the Netflix Prize Dataset (arXiv preprint cs/0610105), 2006.

სტატისტიკური დამოკიდებულებებით, რათა გამოხშიროს განსაკუთრებული კატეგორიის მონაცემი ჩვეულებრივისგან.

- ამ თვალსაზრისით საინტერესოა 2014 წლის საქმე, რომელიც ეხება ნიუ-იორკის ტაქსისა და ლიმუზინის კომისიის მონაცემთა ბაზებს.[14] კერძოდ, ანონიმიზებული მონაცემთა ბაზა შეიცავდა დეტალებს ქალაქში ტაქსით ყველა მგზავრობის შესახებ. მიუხედავად ამისა, მკვლევრებს შეეძლოთ აღნიშნული მონაცემთა ბაზიდან არა მხოლოდ კონკრეტული მძღოლების არამედ სხვა პირების, მათ შორის, ცნობილი ადამიანების აქტივობების და შემოსავლის იდენტიფიცირება და იმის დემონსტრირება, რომ სავარაუდო ანონიმური მონაცემებიც კი შეიძლება იყოს დაუცველი ამგვარი შეტევების მიმართ. აღნიშნული ცხადყოფს, რომ „Inference"-ის თავდასხმის დროს ანონიმიზებულად მიჩნეული მონაცემებიც შესაძლოა იყოს დაუცველი.

ამგვარად, ანონიმიზაციის მძლავრი მექანიზმი იცავს პირებს მონაცემთა ბაზაში „გამორჩევიდან" ("Singling out"), ასევე, ერთ ან რამდენიმე მონაცემთა ბაზაში არსებული ჩანაწერების მათთან ასოცირებისგან ("Linkability") და მონაცემთა ბაზაში შემავალი კონკრეტული პირების შესახებ ინფორმაციის გამოხშირვისგან ("Inference")[15]. შესაბამისად, შეიძლება ითქვას, რომ მაიდენტიფიცირებელი ელემენტების მხოლოდ ამორიცხვა შესაბამის მექანიზმს არ წარმოადგენს, რადგან რიგ შემთხვევებში მონაცემთა დამუშავების კონტექსტი, მიზანი და აუცილებლობა დამატებით ნაბიჯებს მოითხოვს. აქვე უნდა აღინიშნოს, რომ მონაცემთა დაცვის კანონმდებლობა ვრცელდება მხოლოდ იმ შემთხვევაში, თუკი პირის იდენტიფიცირება შესაძლებელია, განურჩევლად დამუშავებისთვის პასუხისმგებელი პირის ან დამუშავებაზე უფლებამოსილი პირის მიზნებისა.

კონფიდენციალობასთან დაკავშირებული ბოლოდროინდელი შეხედულებები აღნიშნულს ყოფს ორ კატეგორიად: იდენტობის გამჟღავნება და პირის შესახებ უცნობი ინფორმაციის ცნობილი ინფორმაციიდან ამოღების შესაძლებლობა.[16] აღნიშნული მოიცავს 29-ე სამუშაო ჯგუფის მიერ განმარტებულ „გამორჩევის" ("Singling Out") და „ვარაუდის" ("Inference") ტიპს. ორივე შემთხვევაში თავდასხმა წარმატებული იქნება, თუკი თავდამსხმელს შეეძლება პირის იდენტიფიცირება მონაცემთა ბაზაში („იდენტობის გამჟღავნება") ან თავდამსხმელი მოიპოვებს ინფორმაციას კონკრეტული ინდივიდის შესახებ (ელემენტების გამოხშირვა). თავდასხმის წარმატებით განხორციელებისას, თავდამსხმელს ან სწორად უნდა შეეძლოს პირის იდენტიფიცირება მონაცემთა ბაზაში (იდენტიფიკაციის გამჟღავნება), ან უნდა ჰქონდეს გარკვეული ინფორმაცია კონკრეტული ინდივიდის შესახებ („პარამეტრის დასკვნა"). აღნიშნული განმარტებიდან გამომდინარე, „დაკავშირება" ("Linkability") მიიჩნევა ზემოაღნიშნული მიზნებიდან ერთ-ერთის მიღების საშუალებად და არა თავდასხმის ტიპად, დღევანდელ ციფრულ სამყაროში არსებული ინფორმაციის სიმრავლის და პიროვნებებთან მის დასაკავშირებლად[17] საჭირო შეზღუდული შესაძლებლობებიდან გამომდინარე.[18]

---

[14] *Hern A.,* New York Taxi Details Can Be Extracted from Anonymised Data, Researchers Say, The Guardian, 2014.

[15] Article 29 Data Protection Working Party (WP29), "Opinion 05/2014 on Anonymisation Techniques", 2014.

[16] *Hu J., Bowen C.M.*, Advancing Microdata Privacy Protection: A Review of Synthetic Data, 2023.

[17] იხ.: *De Montjoye Y. A., Hidalgo C. A., Verleysen M., & Blondel V. D.*, Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports, 3(1), 2013, 1-5.

[18] *Beduschi A.*, Synthetic Data Protection: Towards a Paradigm Change in Data Regulation? *Big Data & Society*, 11(1), 2024.

3. სინთეზური მონაცემების გენერირება, როგორც ანონიმიზაციის მექანიზმი

სინთეზური მონაცემების კონცეფციას განვმარტავთ შემდეგნაირად[19]: „სინთეზური მონაცემი ("SD") არის მონაცემი, რომელიც წარმოიქმნება მიზანმიმართული მათემატიკური მოდელის ან ალგორითმის („გენერატორის") გამოყენებით, მონაცემთა მეცნიერების ამოცანების ამოსახსნელად.

ანალიზისას ჩვენ შემოვიფარგლებით მხოლოდ იმ სინთეზური მონაცემებით, რომლებიც გენერირებულია რეალურ მონაცემებზე დამყარებული მანქანური სწავლების ალგორითმის მეშვეობით (შემდგომ „სინთეზური მონაცემი" ან "SD"). იურიდიულ საზოგადოებაში სინთეზური მონაცემების სწორედ ეს ფორმა წარმოადგენს ყველაზე განხილვად საკითხს.

აღნიშნულ თავში ყურადღება გამახვილდება, თუ რატომ წარმოადგენს სინთეზური მონაცემები მე-2 თავში აღნიშნულ ანონიმიზაციის ერთ-ერთ ტექნიკას. ამ მიზნით, პირველ რიგში, განხილულ იქნება კომპიუტერული მეცნიერების პერსპექტივები, ხოლო შემდგომ ტექნოლოგიებთან დაკავშირებული დინამიკური და განვითარებადი სამართლებრივი ლანდშაფტი. ბოლოს, განიხილება შესაბამისი ტექნოლოგიების სწორი გამოყენების გზები, რისკფაქტორების გათვალისწინებით, ამასთან, თავში მოცემულია კონკრეტული რეკომენდაციებიც. ყოველივე ეს ხაზს უსვამს, რომ სინთეზური მონაცემი წარმოადგენს ანონიმიზაციის ერთ-ერთ ტექნიკას.

3.1. მეცნიერული პერსპექტივები

სინთეზურ მონაცემთა გამოყენების მთავარი მიზანი ანონიმიზაცია და კონფიდენციალობის დაცვაა.[20] აღნიშნულის მთავარი მიზეზი სინთეზური მონაცემების კონცეფციაა, რომელიც პიროვნებასა და მონაცემებს ერთმანეთს აშორებს (ვერ ხერხდება იდენტიფიცირება). ანონიმიზაციის არსებული ტექნოლოგიები ეფუძნება რეალურ მონაცემებს, მათ შორის, შემთხვევითი განაწილებისა ("randomization") და გენერაციის მეთოდებს.[21] აღნიშნულის საპირისპიროდ, სინთეზური მონაცემი გენერირდება მონაცემთა ნაკრების სტატისტიკური მოდელის შექმნით. კერძოდ, მოდელი, განსაზღვრული ალბათობის საფუძველზე, აკვირდება რეალური მონაცემების შაბლონებს ("pattern"). შემდგომ ამ მოდელის ნიმუშის აღება შესაძლებელია ახალი, სრულად ხელოვნური ჩანაწერების შესაქმნელად.

აღნიშნული პროცესის უკეთ აღქმა შესაძლებელია ექსპერიმენტის საფუძველზე.[22] დავუშვათ, პირს სურს შექმნას მონაცემთა ხელოვნური ნაკრები, რომელიც ზუსტად ასახავს ადამიანთა რეალურ მახასიათებლებს. წინასწარ ჩატარებული კვლევის საშუალებით პირს აქვს ცოდნა ადამიანთა ზოგიერთი მახასიათებლის შესახებ. მაგალითად, ცნობილია, რომ დედამიწაზე ქალისა და

[19] *Jordon J., Szpruch L., Houssiau F., Bottarelli M., Cherubin G., Maple C., Weller A.*, Synthetic Data-What, Why and How? 2022.
[20] იქვე.
[21] Article 29 Data Protection Working Party (WP29), "Opinion 05/2014 on Anonymisation Techniques", 2014.
[22] This thought experiment was previously described in a blog post, <https://aindo.com/blog/is-synthetic-good/> [30.07.2024].

მამაკაცის თანაფარდობა არის 1:1. გარდა ამისა, ცნობილია, რომ ყოველ ექვს ადამიანში მხოლოდ ერთს აქვს ცისფერი თვალები, დანარჩენს კი - ყავისფერი. „ხელოვნური ადამიანის“ მონაცემის შესაქმნელად მკვლევრები ატარებენ შემდეგ ექსპერიმენტს: მონეტის აგდებით და მიღებული შედეგით (ალბათობა ½) ადგენენ „ხელოვნური ადამიანი“ იქნება ქალი თუ კაცი. შემდგომ მკვლევრები აგორებენ კამათელს, თუკი გაგორდა 6-იანი (ალბათობა 1/6) „ხელოვნური ადამიანის“ თვალის ფერი იქნება ლურჯი, ხოლო სხვა ნებისმიერი შედეგის შემთხვევაში - ყავისფერი (ალბათობა 5/6). ამ ექსპერიმენტის მრავალჯერ გამეორებით, მკვლევრები იღებენ მთლიანად სინთეზურ მოსახლეობას, რომელსაც აქვს რეალური მოსახლეობის ყველა სტატისტიკური და მათემატიკური თვისება.[23] მიუხედავად ამისა, ჩანაწერები წარმოიქმნება ალბათობაზე დაფუძნებული ექსპერიმენტებით და არა რეალური ინდივიდების მონაცემების გამოყენებით.

გენერირებად ხელოვნურ ინტელექტზე დაფუძნებული სინთეზური მონაცემიც ანალოგიურად იქმნება, თუმცა ავტომატიზებული და შედარებით უფრო დახვეწილი მეთოდით. ზემოაღნიშნული ექსპერიმენტი მოსახლეობის შესახებ წინასწარ ცოდნას და კვლევას მოითხოვს (კაცების და ქალების თანაფარდობა და თვალის ფერები). გენერირებადი ხელოვნური ინტელექტის გამოყენებისა კი ეს არ არის სავალდებულო. ხელოვნურ ინტელექტზე დაფუძნებული გენერატორები ზუსტად იმეორებენ სახასიათო ნიმუშებს შორის თანაფარდობას. წინა აბზაცში მოცემულ ექსპერიმენტში, მაგალითად, ცისფერთვალებიანთა საერთო წილი შეიძლება იყოს 1/6, მაგრამ, შესაძლოა, ცისფერი თვალები უფრო ხშირი იყოს ქალებში, ვიდრე მამაკაცებში. გენერირებად მოდელს შეუძლია ავტომატურად დაადგინოს და გაიმეოროს თანაფარდობა.[24]

როგორც კონფიდენციალობის დაცვის, ისე სწორად გენერირებული სინთეზური მონაცემის ხარისხი ფართოდ არის აღიარებული. აშშ-ს სტანდარტებისა და ტექნოლოგიების ეროვნულმა ინსტიტუტმა (“NIST”) შეადარა დე-იდენტიფიკაციის (“de-identification”) მრავალი ალგორითმი[25]. შედეგების ცხრილი გვიჩვენებს[26], რომ სინთეზური მონაცემის გენერატორები კონფიდენციალობის დაცვის ერთ-ერთ საუკეთესო საშუალებას წარმოადგენს. აღნიშნული ემპირიული მტკიცებულების საფუძველზე შეიძლება ითქვას, რომ სინთეზური მონაცემი ანონიმიზაციის სანდო საშუალებაა, აღნიშნული დასტურდება “NeurIPS 2020 Hide-and-Seek Privacy challenge”-ის[27] ფარგლებშიც.[28]

[23] By a theorem called “The Law of Large Numbers”, see any introductory text on probability theory.

[24] *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.

[25] See: *Task C., Bhagat K., Howarth G.,* SDNist v2: DeidentifiedDataReport Tool, 2023, <https://data.nist.gov/od/id/mds2-2943> [30.07.2024].

[26] See NIST Collective Research Cycle (CRC), <https://pages.nist.gov/privacy_collaborative_research_cycle/pages/archive.html> [30.07.2024].

[27] *Jordon J., Jarrett D., Saveliev E., Yoon J., Elbers P., Thoral P., van der Schaar M.*, Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-Identification.

[28] Competition and Demonstration Track, 2021.

3.2. პოლიტიკა და მარეგულირებელი პერსპექტივები

სინთეზური მონაცემის გენერირების პროცესი, როგორც ანონიმიზაციის საშუალება სამეცნიერო საზოგადოების მსგავსად, იურიდიულ საზოგადოებაშიც დიდი პოპულარობით სარგებლობს. თავის სასემინარო ნაშრომში, სტივენ ბელოვინი აღნიშნავს, რომ[29] „სინთეზური მონაცემი გვთავაზობს პროგრესს. მიუხედავად იმისა, რომ იგი არ არის უნაკლო, აღნიშნული მეთოდი იძლევა საშუალებას ბოლო მოეღოს იდენტიფიცირების პრობლემას და ფოკუსირება მოხდეს უფრო მნიშვნელოვან საკითხებზე - სასარგებლო პერსონალურ მონაცემებზე. აქედან გამომდინარე, კონფიდენციალობის დაცვის სფეროში ჩართულ საზოგადოებას ვურჩევთ დანერგონ სინთეზური მონაცემები შესაბამისი ბაზებში კონფიდენციალობის პრობლემის გადასაჭრელად.

აღნიშნული მოსაზრება ეხმიანება Finocchiaro-ს და სხვების უახლეს იურიდიულ ანალიზს, რომელიც თანახმადაც სინთეზური მონაცემის სწორი გენერირება წარმოადგენს „ანონიმიზაციის მოწინავე ტექნოლოგიას, რომელიც შეესაბამება მონაცემთა დაცვის მოქმედ და სამომავლო რეგულაციებს“.[30]

სამეცნიერო საზოგადოების გარდა, სინთეზური მონაცემები, როგორც ანონიმიზაციის მეთოდი, უფრო და უფრო პოპულარული ხდება კანონშემოქმედებსა და სხვა სამთავრობო ორგანიზაციებს შორის.[31] ესპანეთის პერსონალურ მონაცემთა დაცვის საზედამხედველო ორგანო (“AEDP”) ცალსახად აღიარებს, რომ სინთეზური მონაცემი, გარკვეულ პირობებში, შესაძლოა გაუტოლდეს ანონიმურ მონაცემებს. „მონაცემთა მართვის აქტი“ (“The Data Governance Act”)[32] და „ხელოვნური ინტელექტის შესახებ აქტი“ (“AI Act”)[33] აღიარებს სინთეზური მონაცემის გენერაციას, როგორც მონაცემთა დამუშავებისას კონფიდენციალობის შენარჩუნების მეთოდს, რომელიც გაიგივებულია ანონიმურ მონაცემებთან. ევროკავშირის ერთობლივი კვლევითი ცენტრი (“JRC”) ასევე, მიიჩნევს, რომ სწორად შემუშავებულ სინთეზურ მონაცემი, ერთი მხრივ, ამცირებს კონფიდენციალობასთან დაკავშირებული რისკებს, ხოლო, მეორე მხრივ, ხელს უწყობს ხელოვნური ინტელექტის სფეროს სტრატეგიულ განვითარებას.[34]

ბოლოს, საინტერესოა საფრანგეთში მოქმედი კანონი "Loi pour une République numérique", რომელიც საფრანგეთის მონაცემთა დაცვის ორგანოს (“CNIL”) უფლებას აძლევს დაამტკიცოს ანონიმიზაციის ტექნიკა. 2020 წელს “CNIL”-მა გამოიყენა აღნიშნული უფლებამოსილება და დაამტკიცა სინთეზური მონაცემების გენერირების მეთოდოლოგია, კერძოდ, დაადასტურა მისი ანონიმური ბუნება, გადაწყვეტილების მიღების პროცესში კი დაეყრდნო ისეთ კრიტერიუმებს როგორიცაა

[29] *Bellovin Steven M., Dutta, Preetam K., Reitinger N.*, Privacy and Synthetic Datasets. *S*tanford Technology Law Review, Vol. 22, 2018.

[30] *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.

[31] Agencia Española de Protección de Datos (AEDP), “Approach To Data Spaces From GDPR Perspective”, 2023.

[32] See Recital 7 of the Regulation (EU) 2022/868.

[33] See art. 10.5.a and art. 59.1.b of the Regulation (EU) 2024/1689.

[34] See *Hradec J., Craglia M., Di Leo M., De Nigris S., Ostlaender N., Nicholson N.,* Multipurpose Synthetic Population for Policy Applications, JRC Technical Report, 2022, 58-60.

ინდივიდიალიზაციის, დაკავშირებისა (“Linkability”) და ვარაუდის (“Inference”) შესაძლებლობა.[35]

“CNIL”-მა დაიწყო ემპირიული შეფასების სისტემასთან დაკავშირებით საკანონმდებლო რეგულაციაზე მუშაობა, რომლის მიზანია დაადგინოს თუ რამდენად იცავს კონკრეტული სინთეზური მონაცემთა ნაკრები კონფიდენციალობას.[36] აღნიშნულის მეშვეობით, რაოდენობრივ და უფრო ობიექტური ანალიზის საფუძველზე იქნება შესაძლებელი სინთეზური მონაცემების გენერირების სხვადასხვა ტექნიკის ვალიდაცია.

4. სინთეზური მონაცემების გამოყენების რეკომენდაციები

მესამე თავში განხილულია სინთეზური მონაცემების, როგორც ანონიმიზაციის ტექნოლოგიის, მეცნიერული და საკანონმდებლო პერსპექტივები. ნაშრომში არაერთხელ აღინიშნა, რომ სწორი გამოყენების პირობებში, სინთეზური მონაცემი წარმოადგენს ანონიმურობის მიღწევის მეტად ძლიერ საშუალებას. კონფიდენციალობის პოტენციური რისკების, ასევე, სინთეზური მონაცემების გენერაციის თეორიული შედეგების საფუძველზე გაანალიზებულია თუ რას ნიშნავს „სათანადო გენერირება“. თავში მოცემულია რისკების მინიმალიზაციისთვის საჭირო კონკრეტული რეკომენდაციები. აღნიშნული ხელს უწყობს ტექნოლოგიის სწორ გამოყენებას, რათა სინთეზურ მონაცემს მიენიჭოს ანონიმიზაციის სტატუსი.

4.1. სინთეზურ მონაცემთან დაკავშირებული რისკები

ძირითადი რისკფაქტორები იდენტიფიცირებულია ჩვენს წინა ნაშრომში[37], კერძოდ: 1) გენერატორის ხარისხი; 2) სინთეზის მიდგომა; 3) რეალური მონაცემთა ნაკრების თვისებები (აღკვეთის არსებობა, სიმცირე და ა.შ.); 4) თავდამსხმელის ხელთ არსებული ინფორმაცია (საფრთხის მოდელი). ეს ფაქტორები შეესაბამება ევროკავშირის მონაცემთა დაცვის ზედამხედველის (“EDPS”)[38] მიერ გამოვლენის სინთეზურ მონაცემთან დაკავშირებულ საფრთხეებს. აღნიშნული რისკები პრაქტიკაში შეიძლება შემცირდეს კონფიდენციალობის შეფასებისას პროაქტიული და რაოდენობრივი მიდგომის მიღებით.[39/40]

---

[35] See Octopize’s communication on the CNIL evaluation in the following presentation: <https://documentation-snds.health-data-hub.fr/files/presentations/meetup-snds7/20210318_Octopize_Deck-OctopizeHdhCom_MLP-2.0.pdf> [30.07.2024].

[36] SeeCNIL’s Letterto Statice GmbH on Anonymeter, <https://www.anonos.com/hubfs/Documents/Reports/CNIL_Anonymeter.pdf> [30.07.2024].

[37] *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., Chauvenet C.R.*, Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: Hideyuki Matsumi, Paul De Hert et al. (ed) Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.

[38] *Wiewiórowski W.*, Synthetic Data: What Use Cases as Privacy Enhancing Technology? IPEN Webinar on Synthetic Data, European Data Protection Supervisor.

[39] *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L.*, Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.

[40] *Boudewijn A. T. P., Ferraris A. F., Panfilo D., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Privacy Measurements in Tabular Synthetic Data: State of the Qrt and Future Research Directions, NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI, 2023.

რეკომენდაცია: სინთეზურ მონაცემებთან მომუშავე პრაქტიკოსებმა უნდა იცოდნენ კონფიდენციალობის პოტენციური რისკები. მეტიც, მათ უნდა გამოიყენონ რაოდენობრივი და ობიექტური მეთოდები კონფიდენციალობის რისკების შესაფასებლად და შესამცირებლად. ჩვენი ბოლო გამოკითხვა კონფიდენციალობის რაოდენობრივ განსაზღვრასთან დაკავშირებით სიღრმისეულად აანალიზებს ამგვარი მეთოდებს და მათ უპირატესობებს. გარდა ამისა, წარმოგიდგენთ დამატებით პრაქტიკულ რეკომენდაციებს ამ მეთოდების ოპტიმალური გამოყენებისთვის. დამატებით გირჩევთ გამოიყენოთ სხვადასხვა ზომები, რათა დარწმუნდეთ, კონფიდენციალობასა და სარგებლიანობას შორის სწორი ბალანსის არსებობაში.

4.2. სინთეზურ მონაცემსა და გენერაციაზე დაფუძნებული კონფიდენციალობა

სინთეზური მონაცემის ტექნოლოგიებთან მიმართებით ფუნდამენტური შეკითხვაა კონფიდენციალობის დაცვა წარმოადგენს სინთეზური მონაცემის (შემდგომ - სინთეზურ მონაცემებზე დაფუძნებული) თუ გენერატორის (შემდგომ: გენერატორზე დაფუძნებული) საკუთრებას. ეს შეხედულებები გავლენას ახდენს პრაქტიკაში კონფიდენციალობის რაოდენობრივ შეფასებაზე. სხვადასხვა პერსპექტივის საილუსტრაციოდ, განვიხილოთ ე.წ. „მუდმივ მაიმუნთა თეორია“ (“infinite monkey theorem”)[41]. ეს თეორია ამტკიცებს, რომ თუკი მაიმუნი, შემთხვევითობის პრინციპით, საბეჭდი მანქანის კლავიშებს მუდმივად დააჭერს თითს, დროის გარკვეულ მომენტში აკრეფს შექსპირის ნაშრომს. ანალოგიურად მოქმედებს გენერატორიც, რომელიც თავისი შემთხვევითობრივი (“stochasticity”) ხასიათის გამო, ქმნის (თითქმის) რეალური ინდივიდის იდენტურ მონაცემებს. ჩნდება კითხვა:[42] ირღვევა თუ არა ამ პიროვნების კონფიდენციალობა? „სინთეზურ მონაცემებზე დაფუძნებული“ კონფიდენციალობის პირობებში, პასუხი არის „დიახ“, რადგან სინთეზურ მონაცემთა ნაკრები ავლენს ინდივიდს. „გენერატორზე დაფუძნებული“ კონფიდენციალობის პირობებში, პასუხი არის „არა“, რადგან ინდივიდის მონაცემების შექმნა არ ეფუძნება მოდელის მიერ სასწავლო ნაკრების ჩანაწერების დაცვის ნაკლებობას.

პრაქტიკაში უფრო გავრცელებულია „სინთეზურ მონაცემებზე დაფუძნებული“ კონფიდენციალობა: ბოლოდროინდელმა ლიტერატურულმა მიმოხილვამ როგორც სასარგებლო, ისე კონფიდენციალობის რაოდენობრივ განსაზღვრაზე სამედიცინო კონტექსტში აღმოაჩინა, რომ ნაშრომების 81.2% (22-დან 18), რომელიც სინთეზურ მონაცემთან მიმართებით კონფიდენციალობის რაოდენობრივი მიდგომას იყენებს, არის „სინთეზურ მონაცემებზე დაფუძნებული“; მხოლოდ 9.1% (22-დან 2) არის

[41] იხ. ნაშრომის საწყისი ნაწილი ალბათობის თეორიის შესახებ.

[42] “infinite monkey” თეორემისგან განსხვავებით, გენერატორი, როგორც წესი, არ გამოიმუშავებს ინდივიდის ჩანაწერს. გამომდინარე იქედან, რომ გენერატორი ზუსტად აყალიბებს შედეგებს, განსხვავებით უკვე აღნიშნულ ფარგლებში, მასში მოხვედრილი საკითხების ერთგვაროვანი განაწილებისგან. ამგვარად, გენერატორი სულაც არ წარმოქმნის ყველა შესაძლო ალბათობის კომბინაციას, უსასრულო რაოდენობის განსაზღვრით, ხოლო შემთხვევა, როდესაც გენერატორი ქმნის ჩანაწერს, რომელიც ზედმეტად ჰგავს რეალურ პირს, რომლის მონაცემები არ იყო გამოყენებული, ეს თავსდება მისი შესაძლებლობების სფეროში.

„გენერატორზე დაფუძნებული“. დანარჩენი 9.1% იყენებს ორივეს კომბინაციას[43]. „სინთეზურ მონაცემებზე დაფუძნებული“ კონფიდენციალობის პოპულარობას სავარაუდოდ ორი მიზეზი განაპირობებს. პირველ რიგში, უსასრულო მაიმუნის თეორია ავლენს მოდელზე დაფუძნებული კონფიდენციალობის ხარვეზებს. მეორეც, მონაცემთა ბაზაზე დაფუძნებული კონფიდენციალობა შეიძლება შეფასდეს ისეთი კრიტერიუმის გამოყენებით, რომელიც ადვილად გასაგებია და პირდაპირ უკავშირდება რეალურ სიტუაციებს. აღნიშნული კრიტერიუმები, როგორც წესი, ზომავს: 1. რამდენად სავარაუდოა, რომ გენერირებული ჩანაწერები დაკოპირებულია რეალურიდან, 2. სარისკო მსგავსებას სინთეზურ და რეალურ ჩანაწერებს შორის; 3. რამდენად წარმატებულია განზრახი თავდასხმები.[44]

გენერატორზე დაფუძნებული კონფიდენციალობა გთავაზობთ რამდენიმე უპირატესობას. პირველი რიგში, ის, როგორც წესი, ინტეგრირებულია გენერირების პროცესში მომხმარებლის მიერ კონტროლირებადი, პროაქტიული გზით. აღნიშნულის ყველაზე გავრცელებული მაგალითია განზოგადების კონტროლი[45] ადრეული გაჩერების მეშვეობით და დიფერენციალური კონფიდენციალობა, სადაც მომხმარებელი თავად ადგენს „კონფიდენციალობის ბიუჯეტს“. გარდა ამისა, დიფერენციალური კონფიდენციალობა იყენებს კომპოზიციის თეორიას, რომლებიც ამტკიცებს რომ მათემატიკური მანიპულირება ვერ შეუქმნის საფრთხეს კონფიდენციალობის დაცვას. აღნიშნულს დიდი მნიშვნელობა ენიჭება იმის განსასაზღვრად ანონიმიზაციის პროცესი იყო თუ არა წარმატებული. თუმცა, არის გამოწვევები: „კონფიდენციალობის ბიუჯეტის“ ინტერპრეტაცია პრაქტიკაში საკმაოდ ძნელია[46][47]. უფრო მეტიც, კომპოზიციის თეორია იცავს მათემატიკური მანიპულაციისგან, მაგრამ არა კომბინირებული დამხმარე ინფორმაციის (“compounding auxiliary information”) გამოყენებისგან.

რეკომენდაცია: ზემოაღნიშნულის გათვალისწინებით, გირჩევთ გამოიყენოთ მონაცემთა და გენერატორზე დაფუძნებული მეთოდების კომბინაცია. მონაცემთა ბაზაზე დაფუძნებული მეთოდები ფოკუსირებულია სინთეზურ მონაცემთა ნაკრებებზე, რომლებიც იყენებენ კონფიდენციალობის ემპირიულ შეფასებებს და მსგავსებების შეფასებებს, რათა უზრუნველყონ, რომ ინფორმაცია არ იყოს იდენტიფიცირებადი. ამის საწინააღმდეგოდ, მოდელებზე დაფუძნებული ტექნიკა, კონფიდენციალობის მექანიზმების უშუალოდ მონაცემთა გენერირების პროცესში ჩართვის მეშვეობით, იძლევა ძლიერ თეორიულ გარანტიებს. ეს მეთოდები უზრუნველყოფენ, რომ თავად გენერაციული მოდელები არ ააშკარავებენ ცალკეულ მონაცემებს. კომბინირებული მიდგომა მაქსიმალურად ზრდის კონფიდენციალობას, რაც ეხება როგორც მოდელს, ისე მონაცემებს.

[43] *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L*., Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.

[44] For a thorough survey on both SD-based and generator-based quantification of privacy, see: *Boudewijn A. T. P., Ferraris A. F., Panfilo D., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Privacy Measurements in Tabular Synthetic Data: State of the Qrt and Future Research Directions, NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI, 2023.

[45] *Dwork C., Roth A.,* The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science, 9(3-4), 2014, 211-407.

[46] *Lee J., Clifton C.*, How Much is Enough? Choosing for Differential Privacy.

[47] International Conference, ISC 2011, Proceedings 14, Springer Berlin Heidelberg, 2011, 325-340.

4.3. „შინაარსთან დაკავშირებული“ და სინთეზური მონაცემები

ლოპეზი და ელბი ამტკიცებენ, რომ დარღვევად უნდა ჩაითვალოს მხოლოდ იმ ინფორმაციაზე წვდომა, რომელიც „შინაარსობრივად“ დაკავშირებულია ინდივიდთან.[48] აღნიშნული თეორია საკმაოდ პრაქტიკულია, რადგან საჯაროდ არსებული სტატისტიკური მონაცემების გამოყენებით შესაძლებელია კონკრეტული ინდივიდების იდენტიფიცირება. მაგალითად რეგიონის საშუალო შემოსავლის მიხედვით, შეიძლება ამ რეგიონის ხალხის მიმართ გარკვეული მიკერძოება ან ცრურწმენა ჩამოაყალიბოს.

ამ მოსაზრების შესაბამისად სინთეზური მონაცემის გამოყენებისას ჩნდება კითხვა, აძლევს თუ არა სინთეზურ მონაცემზე წვდომა თავდამსხმელებს კონკრეტულ ინდივიდებთან დაკავშირებულ ინფორმაციას. სინთეზური მონაცემი (მათ შორის, გენერაციული) მხოლოდ იმ შემთხვევაში უქმნის რისკს კონფიდენციალობას, თუ მასზე წვდომით შესაძლებელია კონკრეტულ ინდივიდზე იმ ინფორმაციის მოპოვება, რომლის განცალკევება ზოგადი ინფორმაციიდან შეუძლებელი იქნებოდა. აღნიშნული რელევანტურია „ვარაუდის“ (“inference”) სახის თავდასხმის შემთხვევა, რადგან ასეთ დროს გამოიყენება პიროვნების დამახასიათებელი ნიშნები.

ჯიომის და სხვების[49] მიერ შემუშავებული “The Anonymeter framework” მოიცავს „ვარაუდის“ (“inference”) სახის თავდასხმის შეფასების მეთოდს, რომელიც, მათ შორის, ითვალსიწინებს ინფორმაცია არის თუ არა სპეციფიკური თუ ზოგადი. კერძოდ, აღნიშნული მეთოდი ერთმანეთს ადარებს თუ რა შემთხვევაში მოხდა „ვარაუდის“ (inference) სახის თავდასხმის იდენტიფიცირება სინთეზური მონაცემის გამოყენებით და რამდენ შემთხვევაში მაკონტროლებელი ჯგუფის მიერ. თავის მხრივ, მაკონტროლებელი ჯგუფი შედგება იმ რეალური მონაცემებისგან, რომლებიც არ გამოყენებულა გენერატორის სწავლებისთვის. თუკი აღნიშნული მაჩვენებლები მსგავსია, სინთეზური მონაცემით მოპოვებული ინფორმაცია ზოგადი ხასიათისაა. მეორეს მხრივ, თუკი სინთეზური მონაცემიდან იმაზე მეტი ინფორმაციის მიღება შეიძლება, ვიდრე მაკონტროლებელი ჯგუფისგან, აღნიშნული ინფორმაცია შინაარსობრივად დაკავშირებულია კონკრეტულ პირთან და მიიჩნევა კონფიდენციალობის დარღვევის რისკად.[50]

რეკომენდაცია: კონფიდენციალობის (კერძოდ, „ვარაუდის“ (“inference”) სახის თავდასხმის რისკის) შეფასების დახვეწისთვის, მნიშვნელოვანია გამოიყენოთ იმ რეალური მონაცემებისგან შემდგარი „მაკონტროლებელი ჯგუფი“, რომლებიც არ გამოიყენება გენერატორის სწავლებისთვის. ამით მკვლევრებს შეუძლიათ უფრო ზუსტად განსაზღვრონ, არის თუ არა სინთეზური მონაცემებიდან მიღებული ინფორმაცია ჭეშმარიტად ზოგადი, ან სპეციფიკური დეტალების არასწორი გამოვლენით, ხომ არ ექმნება საფრთხე კონფიდენციალობას.

[48] *López C.A.F., Elbi A.*, On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.

[49] *Giomi M., Boenisch F., Wehmeyer C., Tasnádi B.*, A Unified Framework for Quantifying Privacy Risk in Synthetic Data, 2022

[50] *Platzer M., Reutterer T.*, Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data, Frontiers in Big Data, 2021.

4.4. გენერატორის სწავლება, როგორც მონაცემთა დამუშავების ფორმა

მიუხედავად იმისა, რომ სწორად გენერირებული სინთეზური მონაცემი ანონიმიზებულია, შესაბამის გენერატორს სჭირდება მექანიკური სწავლება რეალური მონაცემების გამოყენებით, აღნიშნული პროცესი კი, მონაცემთა დაცვის კანონმდებლობის შესაბამისად, კლასიფიცირდება, როგორც მონაცემთა დამუშავება. გაერთიანებული სამეფოს პერსონალურ მონაცემთა დაცვის საზედამხედველო ორგანო ("ICO")[51], ცალსახად განმარტავს, რომ ზემოაღნიშნული მექანიკური სწავლების პროცესში მკაცრად უნდა იქნას დაცული შესაბამისი კანონმდებლობა.

კერძოდ, სინთეზური მონაცემების გენერატორების მიერ, საკანონმდებლო რეგულაციებიდან გამომდინარე, მხედველობაში უნდა იქნეს მიღებული შემდეგი გარემოებები:

- სამართლებრივი საფუძველი და მიზანი: დამუშავებისთვის პასუხისმგებელმა პირს უნდა ჰქონდეს მკაფიო და კანონიერი საფუძველი პერსონალური მონაცემების დასამუშავებლად, განსაკუთრებით მაშინ, როდესაც ის მოიცავს განსაკუთრებული კატეგორიის მონაცემებს, როგორიცაა ჯანმრთელობასთან დაკავშირებული მონაცემები ("GDPR" მე-6 და მე-9 მუხლები). ამასთან, მონაცემთა დამუშავების მიზანი უნდა იყოს მკაფიოდ განსაზღვრული და დოკუმენტირებული, და უნდა შეესაბამებოდეს მიზნის შეზღუდვის პრინციპებს.
- მონაცემთა დაცვაზე ზეგავლენის შეფასება ("DPIA"): "DPIA" მეტად მნიშვნელოვანია, მაშინ, როდესაც მონაცემთა დამუშავებამ, შესაძლოა, მაღალი ალბათობით გამოიწვიოს ადამიანის ძირითადი უფლებებისა და თავისუფლებების შელახვა. ეს შეფასება დეტალურად უნდა ასახავდეს მონაცემთა ნაკადებს, ასევე, უნდა აფასებდეს სინთეზის პროცესთან დაკავშირებულ რისკებს და აღწერდეს ზომებს ამ რისკების შესამცირებლად. უნდა აღინიშნოს, რომ "DPIA" არ არის ერთჯერადი პროცესი, არამედ, საჭიროა მისი პერიოდული განახლება, რათა მასში ასახულ იყოს დამუშავების აქტივობებში განხორციელებული მნიშვნელოვანი ცვლილებები და გადაიჭრას მონაცემთა უსაფრთხოებასთან დაკავშირებული სადაო საკითხები.
- ტექნიკური და ორგანიზაციული ზომები: პერსონალური მონაცემების მთლიანობისა და კონფიდენციალობის უზრუნველსაყოფად, "GDPR"-ის 32-ე მუხლი დამუშავებისთვის პასუხისმგებელი პირებისგან მოითხოვს შესაბამისი ტექნიკური და ორგანიზაციული ღონისძიებების მიღებას. აღნიშნული გულისხმობს სინთეზური მონაცემების გენერატორის სწავლების ფაზაში მონაცემთა უსაფრთხო დამუშავების პრაქტიკების მიღებას, ასევე მონაცემთა მინიმიზაციასა და ფსევდონიმიზაციის ტექნიკების შემუშავებას.
- ინფორმაციის აღრიცხვა და შესაბამისობა: მონაცემთა დამუშავების პროცესებთან დაკავშირებული ინფორმაციის აღრიცხვა უნდა მოხდეს "GDPR"-ის მოთხოვნების შესაბამისად (30-ე მუხლი)[52]. აღნიშნული მოიცავს, დამუშავების პროცესის, მიზნების, მონაცემთა კატეგორიის, სუბიექტების და მონაცემთა მიმღებების დეტალურ აღწერას.

[51] *Information Commissioner's Office (ICO)*, Anonymisation: Managing Data Protection Risk Code of Practice, 2012.

[52] As provided by the Accountability principle ex art. 5.2 GDPR.

- სერტიფიკატები და სტანდარტების შესაბამისობა: აღიარებული სტანდარტების დაცვა, როგორიცაა "ISO/IEC 27001" და სერტიფიკატების მოპოვება, როგორიცაა "Europrivacy™/®[53]", კიდევ უფრო უკეთ უზრუნველყოფს "GDPR"-თან შესაბამისობას და გააძლიერებს ანონიმიზაციის პროცესების სანდოობას. "Europrivacy™/®" აფასებს შესაბამისობას "GDPR"-თან და იმართება სერტიფიკაციისა და კონფიდენციალობის ევროპული ცენტრის მიერ ("ECCP"), რომელიც შეესაბამება "ISO/IEC 17065"-სა და "GDPR"-ის 42-ე მუხლს.
- ეთიკური მოსაზრებები: კანონთან შესაბამისობის გარდა, ყურადღება უნდა მიექცეს ეთნიკურ საკითხებსაც. კერძოდ, პროცესმა არ გააძლიეროს ან გააუარესოს თავდაპირველ მონაცემებში არსებული თავდაპირველი ბალანსი. ეს ხელს უწყობს სამართლიანობის შენარჩუნებას და მონაცემთა გამოყენებისას ეთიკური სტანდარტების დაცვას.
- გამჭვირვალობა და ანგარიშვალდებულება: "GDPR" ხაზს უსვამს გამჭვირვალობასა და ანგარიშვალდებულებას. დამუშავებისთვის პასუხისმგებელმა პირებმა მონაცემები უნდა დაამუშაონ გამჭვირვალედ, მათ შორის, სუბიექტებს უნდა მიაწოდონ შესაბამისი ინფორმაცია სინთეზური მონაცემის გამოყენების და კონფიდენციალობის დაცვის ზომების შესახებ. აღნიშნული კომუნიკაცია უნდა იყოს მკაფიო და საჭიროების შემთხვევაში უნდა მოხდეს "DPIA"-ს გამოქვეყნებაც.

რეკომენდაცია: სინთეზური მონაცემების გენერირებაში ჩართულმა პრაქტიკოსებმა უნდა შექმნან სამართლებრივი საფუძველი და განსაზღვრონ ანონიმიზაციის მკაფიო მიზანი, დაიცვან "GDPR"-ის მოთხოვნები. მიზანშეწონილია ჩატარდეს რეგულარული აუდიტი და უწყვეტი მონიტორინგი, მონაცემთა დაცვაზე ზეგავლენის შეფასების ("DPIA") განახლებებთან ერთად, ტექნოლოგიურ და მარეგულირებელ ცვლილებებთან შესანარჩუნებლად. "DPIA"-მ, რომელიც სავალდებულოა "GDPR"-ის 35-ე მუხლით, უნდა შეაფასოს რისკები და გამოკვეთოს მათი შემცირების სტრატეგია.

გარდა ამისა, სინთეზური მონაცემების პრაქტიკის საერთაშორისო სტანდარტებთან გათანაბრებას ("ISO/IEC 27001" და სერტიფიკატების მიღება, როგორიცაა "Europrivacy™/®"), შეუძლია გააძლიეროს შესაბამისობის სანდოობა. "Europrivacy™/®" აღიარებულია ევროკავშირისა და EEA-ს წევრ ქვეყნებში, აფასებს "GDPR"-ის შესაბამისობას და იმართება სერტიფიკაციისა და კონფიდენციალობის ევროპული ცენტრის მიერ ("ECCP"), იცავს "ISO/IEC 17065"-ს და "GDPR"-ის 42-ე მუხლს. ეს მიდგომა ხელს უწყობს სინთეზურ მონაცემთა გამოყენებისას სამართლებრივი და ეთიკური ნორმების დაცვას.

5. დასკვნა

მონაცემთა უსაფრთხოების და კონფიდენციალობის რისკების ზრდის პერიოდში, სინთეზური მონაცემის გენერირება წარმოადგენს ანონიმიზაციის სტრატეგიას. ნაშრომში საფუძვლიანად შეფასდა სინთეზური მონაცემები არაერთი

[53] See Aindo's entry in the Europrivacy/registry: <https://repository.europrivacy.org/en/certifications/edit/d9064da7-603a-4377-b596-b654824e365f> [30.07.2024].

თვალსაზრისით, კერძოდ იურიდიული, თეორიული და პრაქტიკული. აღნიშნული ხაზს უსვამს სინთეზური მონაცემის, როგორც კონფიდენციალობის გამაძლიერებელი ტექნოლოგიის ("PET"), ეფექტურობას. კვლევის შედეგი ადასტურებს, რომ სინთეზური მონაცემის სწორი გენერაციის პირობებში დაცულია ევროკავშირის „მონაცემთა დაცვის ძირითადი რეგულაციით" ("GDPR") და სხვა მსგავსი კანონმდებლობით აღიარებული ანონიმიზაციის მკაცრი კრიტერიუმები.

მონაცემების ისეთ მდგომარეობამდე მიყვანა, როდესაც ინდივიდის იდენტიფიცირება შეუძლებელი ხდება კრიტიკულად მნიშვნელოვანია, რადგან ანონიმიზაციის ტრადიციული ტექნიკები „ხელახალი იდენტიფიკაციის" ("re-identification") ახალ მეტად დახვეწილ საშუალებებს ვერ უმკლავდებიან. სინთეზური მონაცემები, ხელოვნური მონაცემთა ნაკრების შექმნით, გვთავაზობს პრობლემის გადაჭრის თანამედროვე გზას. მონაცემთა ნაკრები ასახავს რეალური მონაცემების სტატისტიკურ ნიმუშებს, რომელიც გამორიცხავს მათ კონკრეტულ პიროვნებასთან დაკავშირების ან/და იდენტიფიცირების საშუალებას. აღნიშნული გზით ერთის მხრივ დაცულია კონფიდენციალობა, ხოლო მეორეს მხრივ, შესაძლებელია მონაცემების გამოყენება მონაცემთა მეცნიერებისა და ხელოვნური ინტელექტის განვითარებისთვის.

სინთეზური მონაცემის, როგორც ანონიმიზაციის საშუალების, საკანონმდებლო დონეზე აღიარება სულ უფრო და უფრო პოპულარული ხდება. ესპანეთის და საფრანგეთის მონაცემთა დაცვის სამსახურებმა, ასევე ევროკავშირის გაერთიანებულმა კვლევითმა ცენტრმა სინთეზური მონაცემები ანონიმიზაციის ერთ-ერთ საშუალებად უკვე აღიარა. ამასთან სინთეზური მონაცემის მნიშვნელობას კიდე უფრო ზრდის ისეთი სერტიფიკატები, როგორიცაა "Europrivacy™/®", რომელიც ადასტურებს, რომ სინთეზური მონაცემის გამოყენება შეესაბამება "GDPR"-სა და სხვა მარეგულირებელ კანონმდებლობას.

მიუხედავად ზემოაღნიშნულისა, სინთეზურ მონაცემთა გენერირება მეტად საფრთხილოა. მნიშვნელოვანია მკაცრად შეფასდეს ისეთი ფაქტორები, როგორიცაა გენერატორის ხარისხი, სინთეზის მეთოდოლოგია, მონაცემთა წყაროს მახასიათებლები და გაბატონებული საფრთხის მოდელი. კონფიდენციალობის შესაფასებლად რაოდენობრივი მეთოდების გამოყენება და გაერთიანებული მიდგომის მიღება, რომელიც აერთიანებს მონაცემებზე და გენერატორზე დაფუძნებულ კონფიდენციალობის სტრატეგიებს, სასიცოცხლოდ მნიშვნელოვანია პოტენციური რისკების ეფექტურად შესამცირებლად. გარდა ამისა, აუცილებელია გენერატორების სწავლების დროს მონაცემთა დაცვის კანონმდებლობის დაცვა და კონფიდენციალობის ემპირიულ შეფასებებში კონტროლის ჯგუფების ჩართვა.

მოსალოდნელია, რომ მონაცემთა ანონიმიზაციის სფერო განვითარდება მოწინავე ტექნოლოგიების გამოყენებით, ასევე, მკაცრი სამართლებრივი რეგულაციებითა და დეტალური ემპირიული რისკის შეფასებებით. აღნიშნული ნაშრომი ხაზს უსვამს, სწორი და სტრატეგიული გამოყენების პირობებში სინთეზური მონაცემების დიდ პოტენციალს ციფრულ სამყაროში მონაცემთა კონფიდენციალობის გამოწვევების დასაძლევად.

ბიბლიოგრაფია:

1. European Commission, "Regulation of the European Parliament and of the Council: on European Data Governance (Data Governance Act)".
2. European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (ARTIFICIAL INTELLIGENCE ACT) and Amending Certain Union Legislative Acts".
3. 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) (Text with EEA relevance).
4. *Agencia Española de Protección de Datos (AEDP),* "Approach To Data Spaces From GDPR Perspective", 2023.
5. *Agencia Española de Protección de Datos (AEDP) and European Data Protection Supervisor (EDPS)*, 10 Misunderstanding related to Anonymization, 2021.
6. *Barth-Jones D.*, The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now, 2012.
7. *Beduschi A.*, Synthetic Data Protection: Towards a Paradigm Change in Data Regulation? *Big Data & Society*, 11(1), 2024.
8. *Bellovin Steven M., Dutta, Preetam K., Reitinger N.*, Privacy and Synthetic Datasets, Stanford Technology Law Review, Vol. 22, 2018.
9. *D'Acquisto G., Naldi M.*, Big Data e Privacy by Design, Anonimizzazione, Pseudonimizzazione, Sicurezza, Giappichelli, 2017.
10. *De Montjoye Y. A., Hidalgo C. A., Verleysen M., Blondel V. D.*, Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports, 3(1), 2013, 1-5.
11. *Dwork C., Roth A.,* The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science, 9(3–4), 2014, 211-407.
12. *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.
13. *Foglia C.*, Il Dilemma (ancora aperto) dell'Anonimizzazione e il Ruolo della Pseudonimizzazione nel GDPR, in Circolazione e Protezione dei Dati Personali, tra Libertà e Regole del Mercato. Commentario al Regolamento UE n. 2016/679 (GDPR) e al Novellato d.lgs. n. 196/2003 (Codice Privacy), 2019.
14. *Giomi M., Boenisch F., Wehmeyer C., Tasnádi B.*, A Unified Framework for Quantifying Privacy Risk in Synthetic Data, 2022.
15. *Hern A.,* New York Taxi Details Can Be Extracted from Anonymised Data, Researchers Say, The Guardian, 2014.
16. *Hradec J., Craglia M., Di Leo M., De Nigris S., Ostlaender N., Nicholson N.*, Multipurpose Synthetic Population for Policy Applications, JRC Technical Report, 2022.
17. *Hu J., Bowen C.M.*, Advancing Microdata Privacy Protection: A Review of Synthetic Data, 2023.
18. *Information Commissioner's Office (ICO),* "Anonymisation: Managing Data Protection Risk Code of Practice", 2012.

19. *Jordon J., Jarrett D., Saveliev E., Yoon J., Elbers P., Thoral P., van der Schaar M.*, Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-Identification, In NeurIPS 2020 Competition and Demonstration Track, 2021.
20. *Jordon J., Szpruch L., Houssiau F., Bottarelli M., Cherubin G., Maple C., Weller A.*, Synthetic Data-What, Why and How? 2022.
21. *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L.*, Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.
22. *Lee J., Clifton C.*, How Much is Enough? Choosing ε for Differential Privacy, In Information Security: 14th International Conference, ISC 2011, Proceedings 14, Springer Berlin Heidelberg, 2011, 325-340.
23. *López C.A.F., Elbi A.*, On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.
24. *Narayanan A., Shmatikov V.*, How to Break Anonymity of the Netflix Prize Dataset, 2006.
25. *OECD*, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches," OECD Digital Economy Papers (OECD, 2023).
26. *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: Hideyuki Matsumi, Paul De Hert et al. (ed) Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.
27. *Platzer M., Reutterer T.*, Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data, Frontiers in Big Data, 4, 2021.
28. *Stalla-Bourdillon S., Knight A.,* Anonymous Data v. Personal Data-False Debate: an EU Perspective on Anonymization, Pseudonymization and Personal Data, Wisconsin International Law Journal, 2017.
29. *Task C., Bhagat K., Howarth G.,* SDNist v2: Deidentified Data Report Tool, 2023, <https://data.nist.gov/od/id/mds2-2943> [30.07.2024].
30. *Tempestini L., D'Acquisto G.,* Il dato personale oggi tra le sfide dell'anonimizzazione e le tutele rafforzate dei dati sensibili,in Le nuove frontiere della privacy nelle tecnologie digitali, Aracne, 2016.
31. *Wiewiórowski W.*, Synthetic Data: What Use Cases as Privacy Enhancing Technology? IPEN Webinar on synthetic data, European Data Protection Supervisor.