

# The Concept of Generated Personal Data

Generated Personal Data and the GDPR

by

**Hideyuki (Yuki) MATSUMI**

Dear We Robot 2025 participants,

I very much appreciate it for willing to take your time to read my very rough draft of *The Concept of Generated Personal Data* for We Robot 2025. Basically, this Article was written to be a chapter potentially for a longer project, which aims to analyze the interplays between generated personal data and the GDPR. The following text is a very early work-in-progress with plenty left to do. As you may see, some sections are more developed than the others, some sections are less so, and thus I appreciate it for letting me know in case anybody is interested in further circulating or citing to it.

I very much welcome any feedback, critiques, questions, suggestions, thoughts, etc. I look forward to seeing you and hearing your thoughts soon!

Hideyuki (“Yuki”) MATSUMI

Draft: March 2025

# THE CONCEPT OF GENERATED PERSONAL DATA

## Generated Personal Data and the GDPR

by Hideyuki (Yuki) MATSUMI<sup>1</sup>

INTRODUCTION .....	3
I. THE CONCEPT OF GENERATED PERSONAL DATA.....	5
A. What is Generated Personal Data .....	5
1. Derived and Inferred Data .....	5
2. Data Shadow, Digital Character, or Digital Twin .....	8
3. Profiling and Profiles .....	9
B. What is Not Generated Personal Data .....	10
1. Manually generated information about individuals.....	10
2. Simple classification or restatement of existing data.....	10
3. Synthetic Data .....	11
C. Examples of Generated Personal Data.....	13
D. Who Generates Personal Data.....	16
E. Generated Personal Data and Other Areas of Law .....	17
II. GENERATED PERSONAL DATA AND THE GDPR .....	19
A. Generated Personal Data as Personal Data .....	20
1. “Predicted Face” as Biometric Data for Identification Purpose.....	20
2. Visual image as Personal Data .....	20
3. Input Data Used to Generate and Output Data Generated: Mixed Personal Data? .....	23
4. Generated Personal Data as “Non-Portable Data”? .....	27
B. Generating Personal Data as Processing of Personal Data.....	28
C. Generating Personal Data as Profiling .....	29
1. Generating Personal Data as Profiling .....	30
2. Generating Personal Data as Solely Automated Decision-Making.....	32
III. THE CHALLENGES RAISED BY GENERATED PERSONAL DATA .....	34
A. Challenges to the Concept of Personal Data .....	35
B. Challenges to the Principle of Transparency .....	36
C. Challenge to the Right to Rectification .....	37
IV. RECOGNIZING GENERATED PERSONAL DATA UNDER THE GDPR .....	39
A. Recognizing Generated Personal Data: Advancing the Legal Framework .....	39
1. Articulating Rules on How Personal Data Can Be Generated.....	40
B. Recognizing Generated Personal Data: Necessary But Not Sufficient .....	41
CONCLUSION .....	42

---

<sup>1</sup> Hideyuki (Yuki) MATSUMI. PhD researcher/candidate at the Research Group on Law Science, Technology & Society (LSTS) at the Vrije Universiteit Brussel (VUB). Visiting Scholar in Privacy and Technology Law, Center for Law and Technology at George Washington University Law School. Member of the New York Bar. I’d like to thank the participants at the Privacy Law Scholars Conference Europe 2024 for very helpful comments.

## INTRODUCTION

In 2014, to fight the littering problem, an environmental group posted portraits of potential litterbugs on public streets.<sup>2</sup> The group extracted DNA from tossed cigarettes, coffee cups, or condoms, and generated the possible likeness of individuals whose DNA was found from these items in public.

Similarly, in 2017, detectives working on a cold case sent DNA found at the crime scene of a homicide and sexual assault victim from 1990 to a company that claims it “can turn DNA into a face.” Subsequently, detectives published the “predicted face” in an attempt to solicit tips from the public. In 2020, they went further. One of the detectives asked to have the “predicted face” run through a facial recognition system.<sup>3</sup>

Depending on how and what “predicted faces” companies generate, many individuals can wrongfully be shamed in public or be identified as a suspect. Thus, it is critical that those data companies or data brokers (collectively data controllers) are held accountable under the law for their data practices, including generating data about individuals. To that end, the first set of questions that should be asked, for the purpose of data protection/privacy law, are: (1) whether data generated by companies constitutes personal data; and if so (2) is/are the data subject(s) of such generated images?

An obvious answer is the owner of the DNA because their genetic information was used to generate the face. But what about others who look similar to those faces, or who were identified or found to be like the suspect on the database of facial recognition system? Today, the issue is further complicated because generative “artificial intelligence,” or gen AI, can produce a “virtual person” that never exists on the planet but may still look like someone, while millions of actual human faces were used in the training process of that generative AI. Who is or are the data subjects of this “virtual person”?

This Article explores and analyzes how the General Data Protection Regulation (GDPR) applies to, or is challenged by, data generated by data controllers, or “generated personal data.”

---

<sup>2</sup> E.g., Justin Worland, *Hong Kong Anti-Littering Campaign Uses DNA From Trash to Shame People*, TIME (May 20, 2015), <https://time.com/3890499/hong-kong-littering-campaign/> (last visited Oct 3, 2024); Smithsonian Magazine & Emily Matchar, *DNA Testing Could Identify Litterbugs and Dog Poop Miscreants*, Smithsonian Magazine, <https://www.smithsonianmag.com/innovation/DNA-testing-could-identify-litterbugs-dog-poop-miscreants-180955178/> (last visited Oct 3, 2024); Fiona MacDonald, *These Billboards Publicly Shame Litterers Using Their Discarded DNA*, ScienceAlert (May 20, 2015), <https://www.sciencealert.com/these-billboards-publicly-shame-litterers-using-their-discarded-dna> (last visited Oct 3, 2024).

<sup>3</sup> Dhruv Mehrotra, *Cops Used DNA to Predict a Suspect’s Face—and Tried to Run Facial Recognition on It*, Wired, <https://www.wired.com/story/parabon-nanolabs-dna-face-models-police-facial-recognition/> (last visited 23rd January 2024).

The Article proceeds as follows. Part I discusses what is generated personal data, and what is not. It explores examples of generated personal data and who typically generates these data. It introduces the technical background on how personal data are generated today. Part II analyzes how generated personal data might be treated under the GDPR. Specifically, it aims to address three issues: (1) whether generated personal data falls within the definition of “personal data”; (2) whether generating personal data falls within the definition of “processing”; and (3) whether generation of personal data falls within the definition of “profiling.” Part III discusses the challenges raised by generated personal data being equated with the traditional concept of personal data, as well as the implications to the key principles built upon the concept enshrined in the current data protection/privacy law. Part IV discusses the virtues of recognizing the concept of generated personal data under the GDPR.

The Article aims to make two central contributions. First, it clarifies how the concepts of personal data and profiling apply to generated personal data. This is relevant and timely in the age of AI, including generative AI, where it is very probable that this technology, or genie, will not go back in the bottle, but will only proliferate. Second, it offers a theoretical foundation as to how and why generated personal data raises challenges to the existing data protection/privacy framework, which will be an integral part of the ongoing debate on how to regulate profiling, inferences, or generated personal data.

# I. THE CONCEPT OF GENERATED PERSONAL DATA

In essence, the concept of generated personal data refers to algorithmically generated data, such as photos, videos, audio, or texts, about individuals. The output -- i.e., generated data -- is created using input data, typically collected personal data.

The concept is nothing new. Many commentators have mentioned similar concepts. In a nutshell, what is referred to as inferred data, profiling or profiles, or credit scores. However, there are critical differences as well.

This Part illustrates the concept of generated personal data. It explains what is and what is not generated personal data, and provides examples of such data.

## A. WHAT IS GENERATED PERSONAL DATA

### 1. Derived and Inferred Data

What is referred to as derived and inferred data is equivalent to the concept of generated personal data. In the past, especially around 2011 to 2014, various taxonomies of personal data have been suggested.

A report by the World Economic Forum from 2011<sup>4</sup> categorizes personal data into three based on how organizations capture them. (1) Volunteered: Individuals voluntarily and explicitly share information about themselves through electronic media, such as someone creates a profile on social network service (SNS). (2) Observed: meaning data is capture by recording activities of users (in contrast to “Volunteered”), where examples include online behaviors or location data when using cell phones. (3) Inferred: Organisations can also discern “inferred” data from individuals, based on the analysis of personal data. Such an example is credit scores, which are calculated based on numerous factors relevant to an individual’s financial history.<sup>5</sup>

Similarly, a report by Boston Consulting Group from 2012 distinguished and compared between (1) volunteered data (unstructured), (2) required data (structured), (3) tracked data (contextual), and (4) mined data (profiled).<sup>6</sup>

---

<sup>4</sup> World Economic Forum, *Personal Data: The Emergency of a New Asset Class*, (2011), <https://www.weforum.org/publications/personal-data-emergence-new-asset-class/> [hereinafter *Personal Data: The Emergency of a New Asset Class*].

<sup>5</sup> *Id.*, at 14.

<sup>6</sup> Boston Consulting Group, *The Value of Our Digital Identity*, (2012), <https://www.bcg.com/publications/2012/digital-economy-consumer-insight-value-of-our-digital-identity>.

On 21 March 2014, the Organisation for Economic Co-operation and Development (OECD) organized an Expert Roundtable entitled “Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking.”<sup>7</sup> One of the sessions, entitled Personal data taxonomies and governance, aimed to consider data categorizations or “personal data taxonomies.”

The *OECD Roundtable* considered a proposal for a new data taxonomy. Rather than categorizing data based on its sensitivity or use model, the proposed taxonomy categorized data based on the manner in which they originated (“taxonomy based on origin”). Specifically, the proposed taxonomy made distinction between four main categories: (1) provided data; (2) observed data; (3) derived data; and (4) inferred data.<sup>8</sup> The same four categories of personal data also appear in the *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* by the Article 29 Data Protection Working Party (“Art. 29 WP”).<sup>9</sup>

The *OECD Roundtable*, however, provides more detailed explanations as to their meanings and the differences. The *provided data* originates from direct actions taken by an individual who is fully aware of the actions that lead to the data's creation. *Observed data* are data that have been observed by others and recorded in a digital format. This data can be recorded either at the moment of its creation or transmitted to a digital carrier after observation.

The derived data and inferred data, which are most relevant to the concept of personal data, are explained as below:

*Derived data* are data generated from other data, after which they become new data elements related to a particular individual. Derived data are said to be created in a fairly “mechanical” fashion using simple reasoning and basic mathematics to detect patterns within a data set and create classifications. While these classifications may later be used for predictive purposes, they are not themselves based on probabilistic reasoning.

*Inferred data* are the product of probability-based analytic processes. They are a result of the detection of correlations which are used to create predictions of behaviour. These predictions are then used to categorise individuals.<sup>10</sup>

The examples of *derived data* are “computational data (e.g. a calculation of customer profitability based on the ratio between number of visits and the items

---

<sup>7</sup> See OECD, *Working Party on Security and Privacy in the Digital Economy: Summary of the OECD Privacy Expert Roundtable*, (2014), [https://one.oecd.org/document/DSTI/ICCP/REG\(2014\)3/en/pdf](https://one.oecd.org/document/DSTI/ICCP/REG(2014)3/en/pdf) [hereinafter *OECD Roundtable Summary*]. A week before, on 14th March 2014, one of the discussion leaders of the OECD roundtable (Martin Abrams, The Information Accountability Foundation) made a presentation on the taxonomy, which the slides are available at <https://www.slideshare.net/slideshow/a-taxonomy-of-personal-data-by-origin-rc/41559737>.

<sup>8</sup> *Id.*, at 5.

<sup>9</sup> Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, (2018), [https://ec.europa.eu/newsroom/article29/item-detail.cfm?item\\_id=612053](https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053) [hereinafter *Profiling Guidelines*].

<sup>10</sup> *OECD Roundtable Summary*, *supra* note \_\_, at 5 (emphasis in original).

bought) and notational data (e.g. the detection of common attributes among ‘profitable’ customers which are then used to classify potential customers)”<sup>11</sup> while “statistical data (e.g., credit risk scores, life expectancy scores), and advanced analytical data (e.g., likelihood of future health outcomes based on an analysis of large and diverse medical data sets)” are examples of *inferred data*.<sup>12</sup>

In *Property and (Intellectual) Ownership of Consumers’ Information*,<sup>13</sup> Gianclaudio Malgieri expanded this personal data taxonomy and offered a different perspective by categorizing based on the degree of “ownership” (and of “relationship”) of a specific piece of information to subjects and to enterprises: strong relationship data (data provided directly by customers), intermediate relationship data (data observed or inferred and related to the present life of consumers), and weak relationship data (predictive data).<sup>14</sup>

Frederike Zufall and Raphael Zingg builds on the categorization model of data that was introduced by the OECD Roundtable and expanded by Malgieri. In a book chapter entitled *Data Portability in a Data-Driven World*<sup>15</sup> -- Chapter 11 of the book entitled *Artificial Intelligence And International Economic Law: Disruption, Regulation, And Reconfiguration*<sup>16</sup> -- they discuss generated data (“data controller-generated data”) contrasting with raw data (“user-generated data”). Raw data, or user-generated data, encompasses provided and observed data, while generated data, or data controller-generated data, consists of derived and inferred data.<sup>17</sup> They explain:

*Derived data* is data generated from other data, created in a “mechanical” manner using simple, non-probabilistic reasoning and basic mathematics for pattern recognition and classification creation (e.g., customer profitability as a ratio of visits and purchases, common attributes among profitable customers). *Inferred data* is data generated from other data either by using probabilistic statistical models for testing causal explanation (“causal inferences”) or by using machine learning models for predicting output values for new observations given

<sup>11</sup> *Id.*

<sup>12</sup> *Id.*

<sup>13</sup> Gianclaudio Malgieri, *Property and (Intellectual) Ownership of Consumers’ Information: A New Taxonomy for Personal Data*, (2016), <https://papers.ssrn.com/abstract=2916058>. See also Gianclaudio Malgieri & Giovanni Comandé, *Sensitive-By-Distance: Quasi-Health Data in the Algorithmic Era*, (2017), <https://papers.ssrn.com/abstract=3020628>.

<sup>14</sup> *Id.*, at 1.

<sup>15</sup> Frederike Zufall & Raphael Zingg, *Data Portability in a Data-Driven World*, in *Artificial Intelligence and International Economic Law: Disruption, Regulation, and Reconfiguration* 215 (Ching-Fu Lin, Shin-yi Peng, & Thomas Streinz eds., 2021), <https://www.cambridge.org/core/books/artificial-intelligence-and-international-economic-law/data-portability-in-a-datadriven-world/F445EC4A9E9665A05E773A88E8840027> [hereinafter *Data Portability in a Data-Driven World*].

<sup>16</sup> Shin-Yi Peng, Ching-Fu Lin & Thomas Streinz, *Artificial Intelligence and International Economic Law* (2021).

<sup>17</sup> *Data Portability in a Data-Driven World*, *supra* note \_\_, at 216.

their input values (“*predictive inferences*”).<sup>18</sup>

## 2. Data Shadow, Digital Character, or Digital Twin

Commentators have been making distinctions between generated data or profiles from other forms of data to address issues specific to generated data or profiles.

In *Automated Profiling*<sup>19</sup>, Lee A. Bygrave explains the difference between “profile generation” and “profile application” to analyze Art. 15 of the 1995 Data Protection Directive. In *Forgetting Footprints, Shunning Shadows*,<sup>20</sup> Bert-Jaap Koops pays particular attention to the differences between individuals’ digital footprints -- data they themselves leave behind -- and individuals’ *data shadows* -- information about them generated by others -- when analyzing a right to be forgotten. Personal data generated by others is also referred to as “*data shadows*” or “*data body*.”<sup>21</sup>

In her book chapter Digital Character in “The Scored Society,”<sup>22</sup> Sociologist Tamara K. Nopper discusses about digital character produced by new credit-scoring models “could maintain the racial wealth gap and expose people to more

---

<sup>18</sup> *Id.*

<sup>19</sup> Bygrave LA, *Automated Profiling* (2001) 17 Computer Law & Security Review 17, <http://classic.austlii.edu.au/au/journals/PrivLawPRpr/2000/40.html>. “[T]he profiling process has two main components:(i) profile generation - the process of inferring a profile; (ii) profile application - the process of treating persons/entities in light of this profile. The first component typically consists of analysing personal data in search of patterns, sequences and relationships, in order to arrive at a set of assumptions (the profile) based on probabilistic reasoning. The second component involves using the generated profile to help make a search for, and/or decision about, a person/entity”). Moreover, he states “[o]n its face, Art.15(1) only lays restrictions on the process of profile application. The same applies with earlier versions of the provision as contained in the first and amended proposals for the data protection Directive,” and contrasts it with “the original proposal for the Directive on telecommunications privacy which specifically restricted the creation of electronic subscriber profiles.” [hereinafter *Automated Profiling*]

<sup>20</sup> Koops B-J, *Forgetting Footprints, Shunning Shadows: A Critical Analysis of the “Right to Be Forgotten” in Big Data Practice* (2011), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1986719](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1986719) (“Big Data involves not only individuals’ digital footprints (data they themselves leave behind) but, perhaps more importantly, also individuals’ *data shadows* (information about them generated by others)”) [hereinafter *Forgetting Footprints, Shunning Shadows*].

<sup>21</sup> E.g., *Forgetting Footprints, Shunning Shadows*; Rob Kitchin, *The Data Revolution Big Data, Open Data, Data Infrastructures and Their Consequences* (2014), <https://uk.sagepub.com/en-gb/eur/the-data-revolution/book242780> (“There are two primary ways in which data can be generated. . . Captured and exhaust data are considered ‘raw’ in the sense that they have not been converted or combined with other data. In contrast, *derived data* are *produced* through additional processing or analysis of captured data. For example, captured data might be individual traffic counts through an intersection and derived data the total number of counts or counts per hour. The latter have been derived from the former. Captured data are often the input into a model, with derived data the output.”); Philip N Howard, *New Media Campaigns and the Managed Citizen* (2006) (“. . . I described the new mechanisms of representation that work in hypermedia campaigns, a system of shadow citizenship in which lobbyists represent public interests but rely on our *data shadows* to model and predict our opinions. . .”);

<sup>22</sup> Tamara K. Nopper, *Digital Character in “The Scored Society”: FICO, Social Networks, and Competing Measurements of Creditworthiness* 170 in *Captivating technology: race, carceral technoscience, and liberatory imagination in everyday life*, (Ruha Benjamin ed., 2019) [hereinafter *Digital Character in “The Scored Society”*].

pernicious forms of financial control.”<sup>23</sup> Digital character refers to “a digital profile assessed to make inferences regarding character in terms of credibility, reliability, industriousness, responsibility, morality, and relationship choices.”

Some commentators have also referred to as “digital twin,” “digital copy,” “digital doppelgangers.”

### 3. Profiling and Profiles

“Profiling”<sup>24</sup> is very closely related to, but has a different perimeter compared to, the concept of generated data. In essence, data generated as the result of “profiling” -- i.e., “profile”<sup>25</sup> -- is generated personal data. Profiling under the GDPR, however, is a narrow concept than that of generated personal data. Thus, the two concepts are not identical.

The term “profiling” can have different meanings based on the context in which it is used.<sup>26</sup> Broadly speaking, “profiling”<sup>27</sup> in general involves making “inferences”<sup>28</sup> about people. *Data Brokers in an Open Society*,<sup>29</sup> for example, refers to profiling as “the creation or use of inferences about people.”<sup>30</sup>

Profiling under the GDPR is narrower, as discussed later. It requires some form of “automated processing of personal data,” and it has to be carried out with the objective to “evaluate certain personal aspects” of individuals. *Profiling* under the GDPR (simply referred to as *profiling*, without double quotation mark, in this Article) is defined as “any form of *automated* processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person.”<sup>31</sup> Profiling consists of three elements: (1) it has to be an *automated* form of processing; (2) it has to be carried out *on personal data*; and (3) the objective of the profiling must be to evaluate personal aspects about a natural person.<sup>32</sup>

---

<sup>23</sup> Ruha Benjamin, *Imagination: A Manifesto* (2024), at 54.

<sup>24</sup> According to Merriam-Webster’s dictionary, profiling is “the act or process of extrapolating information about a person based on known traits or tendencies.” <https://www.merriam-webster.com/dictionary/profiling>.

<sup>25</sup> This Article uses the term profile as data generated by the profiling.

<sup>26</sup> See generally *Profiling the European citizen: cross-disciplinary perspectives*, (Mireille Hildebrandt & Serge Gutwirth eds., Springer 2008) [hereinafter *Profiling the European citizen*].

<sup>27</sup> The term of “profiling” here is different from how the GDPR defines profiling. When I simply write *profiling*, it is the concept of *profiling* under the GDPR defined in Article 4(4), GDPR, while “profiling” with double quotation refers to a broader and commonly used term of profiling. Please note that the CCPA also defines *profiling*, but the definition is basically the mirror of the GDPR definition. CAL. CIV. CODE § 1798.140(z).

<sup>28</sup> The term “inference” with double quotation mark refers to a broader and commonly used term of “inference.” When I simply write *inferences*, it is the concept of profiling under the California Privacy Rights Act (CCPA). *Infer* and *inference* are defined in § 1798.140(r).

<sup>29</sup> Aaron Rieke et al., *Data Brokers in an Open Society*, (2016), <https://www.opensocietyfoundations.org/publications/data-brokers-open-society> [hereinafter *Data Brokers in an Open Society*].

<sup>30</sup> *Id.*, at 5.

<sup>31</sup> Article 4(4) GDPR (emphasis added).

<sup>32</sup> *Profiling Guidelines*, at 6 (emphasis in original).

Thus, human formed opinions or inferences, for example, are excluded from the definition because it lacks the automation element. Similarly, automated processing of personal data, but without an objective to evaluate individuals, will fall outside the definition of profiling.

Importantly, because of how profiling is defined under the GDPR, generated personal data has not all generation of personal data constitutes profiling, and that difference will impact whether the protections offered by the profiling regulation under GDPR applies to generation of personal data, as discussed later.<sup>33</sup>

## B. WHAT IS NOT GENERATED PERSONAL DATA

Some things are not considered personal data, and certain issues fall outside the scope of this Article.

### 1. Manually generated information about individuals

Information about individually that is manually generated is not generated personal data. What sparks my concerns is personal data that is automatically or algorithmically generated by data controllers or data brokers.

As discussed later, the Court of Justice of the European Union (“CJEU”) has decided on the concept of personal data on a numerous occasion. *Nowak*<sup>34</sup> as well as *YS and Others*<sup>35</sup> are such cases. The facts of *Nowak* as well as *YS and Others*, however, do not suggest that the information about individuals at issue (i.e., traditional examination in *Nowak*, and legal analysis in *YS and Others*) were produced automatically or algorithmically.<sup>36</sup>

### 2. Simple classification or restatement of existing data

One relevant issue to ask is, when is new personal data generated, and when is it merely a restatement of already collected data? Consider a case where a data controller classifies individuals into categories, such as “tall” or “slim,” based on height and weight data they collected and compared it with the average. This does not, arguably, constitute the generation of new personal data.

---

<sup>33</sup> See *Generating Personal Data as Profiling, infra*.

<sup>34</sup> Case C-434/16 - *Peter Nowak v Data Protection Commissioner*, judgment of 20 December 2017 (ECLI:EU:C:2017:994) [hereinafter *Nowak*]

<sup>35</sup> Joined Cases C-141/12 and C-372/12, *YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S*, judgment of 17 July 2014 (ECLI:EU:C:2014:2081) [hereinafter *YS and Others*].

<sup>36</sup> See Ausloos, Jef and Mahieu, Rene and Veale, Michael, *Getting Data Subject Rights Right* (December 2019). (2019) 10 JIPITEC 283, <https://ssrn.com/abstract=3544173>, para 45 (“Neither *Nowak* or *YS and Others* can be easily construed as profiling, as both were cases of manual, rather than automated, processing. Legal analysis of the type in *YS and Others* would not fall under the profiling definition. It also seems doubtful that a traditional examination, such as that in *Nowak*, would fall under the concept of profiling (unless it was marked automatically).”) [hereinafter *Getting Data Subject Rights Right*].

On the issue, the Article 29 Data Protection Working Party (“Art. 29 WP”)<sup>37</sup> takes a similar view. In the *Profiling Guidelines*, they observe that “[a] simple classification of individuals based on known characteristics such as their age, sex, and height does not necessarily lead to profiling.”<sup>38</sup>

The Advocate General’s Opinion in the *YS and Others* also touches on a similar issue. In addressing the issue of whether or not the analysis made by the case officer of the Immigration and Naturalisation Service constitutes personal data,<sup>39</sup> the Advocate General raised an example of where “a person’s weight might be expressed objectively in kilos or in subjective terms such as ‘underweight’ or ‘obese’.”<sup>40</sup>

The line between generation of personal data and restatement of collected data, however, becomes blurry when individuals are categorized into groups, such as “interested in gym subscription,” “needs diet,” or “higher health risk.”

Eventually, it hinges upon the purpose of the classification, as the *Profiling Guidelines* explains.<sup>41</sup> This is understandable because the third element (the intent element) of profiling<sup>42</sup> is “to evaluate personal aspects about a natural person,” mirroring the text of the definition.<sup>43</sup>

### 3. Synthetic Data

Synthetic data<sup>44</sup> is closely related to, but has different qualities from, the concept of generated data. In this Article, however, it is distinguished from the concept of generated personal data due to how the concept of synthetic data is used in various contexts. In much of the literature, in short, synthetic data is considered anonymous data, and is viewed as one of privacy enhancing technologies (PET).

---

<sup>37</sup> Now the European Data Protection Board (“EDPB”) under the GDPR.

<sup>38</sup> *Profiling Guidelines*, at 7.

<sup>39</sup> *YS and Others*, at para 56 (“In my opinion, only information relating to facts about an individual can be personal data. Except for the fact that it exists, a legal analysis is not such a fact.”).

<sup>40</sup> *YS and Others*, para 57 (“... I do not find it helpful to distinguish between ‘objective’ facts and ‘subjective’ analysis. Facts can be expressed in different forms, some of which will result from assessing whatever is identifiable. For example, a person’s weight might be expressed objectively in kilos or in subjective terms such as ‘underweight’ or ‘obese’. Thus, I do not exclude the possibility that assessments and opinions may sometimes fall to be classified as data.”).

<sup>41</sup> See also the Council of Europe Recommendation CM/Rec (2010)132 [*CoE Recommendation*].

<sup>42</sup> “(3) the objective of the profiling must be to evaluate personal aspects about a natural person,” *supra* note \_\_.

<sup>43</sup> Article 4(4) GDPR.

<sup>44</sup> See generally Sergey I. Nikolenko, *Synthetic Data for Deep Learning* (2021) [hereinafter *Synthetic Data for Deep Learning*]; James Jordon et al., *Synthetic Data -- What, Why and How?*, (2022), <https://arxiv.org/abs/2205.03257v1> [hereinafter *Synthetic Data -- What, Why and How?*]; Colin Mitchell & Elizabeth Redrup Hill, *Are Synthetic Health Data “Personal Data”?*, PHG Foundation, <https://www.phgfoundation.org/report/are-synthetic-health-data-personal-data> [hereinafter *Are Synthetic Health Data “Personal Data”?*]; Michal Gal & Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, (2023), <https://papers.ssrn.com/abstract=4414385> [hereinafter *Synthetic Data: Legal Implications of the Data-Generation Revolution*].

Although there seems no widely accepted definition of synthetic data,<sup>45</sup> the following define and/or explain the concept well:

- “Data that has been generated using a purpose built-mathematical model or algorithm, with the aim of solving a (set of) data science task(s).”<sup>46</sup>
- “Synthetic data can be thought of as artificial data that closely mimic the properties and relationships of real or source data.”<sup>47</sup>
- “Synthetic data is artificially generated data, created using generative AI, that has analytical value.”<sup>48</sup>

In short, the concept of generated personal data and synthetic data are very similar in the sense that a new data (output) is generated from the source data (input). Nevertheless, for the purpose of this Article, the two concepts differ primarily because of how the latter, i.e., synthetic data, is considered and viewed.

The recently enacted AI Act refers to synthetic data, but next to “anonymised data.”<sup>49</sup> The term “synthetic” appears a few times, and is used to describe AI-generated contents.<sup>50</sup> The term synthetic *data*, however, is coupled with “anonymised data” implying they share similar qualities.

In explaining synthetic data, the European Data Protection Supervisor (EDPS) advises “[a] privacy assurance assessment should be performed to ensure that the resulting synthetic data is not actual personal data.”<sup>51</sup>

The Spanish Data Protection Agency (AEDP) equates synthetic data with

---

<sup>45</sup> See Daniele Panfilo et al., *Measuring Privacy Protection in Structured Synthetic Datasets: A Survey*, in CPDP Book (Hideyuki MATSUMI et al., eds).

<sup>46</sup> *Synthetic Data -- What, Why and How?*, *supra* note \_\_, at 5.

<sup>47</sup> *Are Synthetic Health Data “Personal Data”?*, *supra* note \_\_, at 13 (citing Puja Myles, Johan Ordish, Richard Branson, *Synthetic data and the innovation, assessment, and regulation of AI medical devices*, RF Q. 2021;1:48-53, [https://www.cprd.com/sites/default/files/2022-12/Myles%20et%20al.%20preprint\\_2022.pdf](https://www.cprd.com/sites/default/files/2022-12/Myles%20et%20al.%20preprint_2022.pdf)).

<sup>48</sup> *Synthetic Data: Legal Implications of the Data-Generation Revolution*, *supra* note \_\_, at 2 (Emphasizing “the importance of maintaining analytical value see Donald B. Rubin, *Statistical Disclosure Limitation*, 9 J. OFF. STAT. 461, 462 (1993).”).

<sup>49</sup> Article 10(5)(a), AI Act (“the bias detection and correction cannot be effectively fulfilled by processing other data, including *synthetic* or anonymised data;”).

<sup>50</sup> Article 50, AI Act (“Providers of AI systems, including general-purpose AI systems, generating *synthetic* audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated.”); Recital 133, AI Act (“A variety of AI systems can generate large quantities of *synthetic* content that becomes increasingly hard for humans to distinguish from human-generated and authentic content.”).

<sup>51</sup> European Data Protection Supervisor (EDPS), *Synthetic Data*, European Data Protection Supervisor, <https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data> (“A privacy assurance assessment should be performed to ensure that the resulting synthetic data is not actual personal data. This privacy assurance evaluates the extent to which data subjects can be identified in the synthetic data and how much new data about those data subjects would be revealed upon successful identification.”) (last visited Oct 10, 2024).

anonymized data, asserting its feasibility for personal data anonymization.<sup>52</sup>

In the *Opinion of the Data Ethics Commission*, the German Federal Data Ethics Commission found that “the field of synthetic data shows enormous promise” and recommended increasing funding for this area.<sup>53</sup>

In the US, a report, issued as part of an effort by the White House Office of Science and Technology Policy (OSTP) to advance privacy-preserving data sharing and analytics (PPDSA) technologies,<sup>54</sup> refers to synthetic data as one of the PETs.

Because synthetic data -- artificial data that closely mimic the properties and relationships of real or source data -- is considered anonymous data or at least is viewed as one of privacy enhancing technologies (PET), it is distinguished from the concept of generated personal data in this Article.

### C. EXAMPLES OF GENERATED PERSONAL DATA

At the 2014 OECD Roundtable, a commentator noted that inferred data was believed to have more recent origins than other data (i.e., provided, observed, and derived), dating back to the early 1980s.<sup>55</sup> Unsurprisingly, there has been a remarkable increase in both the volume and variety of available data sets.

Today, the generation of personal data has become more affordable and accessible in terms of time and cost due to technological advancements. One such representative technology is generative artificial intelligence, or GenAI.<sup>56</sup>

---

<sup>52</sup> Agencia Española de Protección de Datos (AEPD), *Approach To Data Spaces From GDPR Perspective*, (2023), <https://www.aepd.es/documento/approach-to-data-spaces-from-gdpr-perspective.pdf>, at 50 (“Another data minimisation strategy is the use of synthetic data. Synthetic data are not random data, but data that meet the same requirements as real data for a specific purpose. . .”).

<sup>53</sup> German Federal Data Ethics Commission, *Opinion of the Data Ethics Commission*, (2019.10.23), [https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission-Opinion.pdf?\\_\\_blob=publicationFile&v=2](https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission-Opinion.pdf?__blob=publicationFile&v=2) (“The development of procedures and standards for data anonymisation and pseudonymisation is central to any efforts to improve controlled access to (formerly) personal data. A legal presumption that, if compliance with the standard has been achieved, data no longer qualify as personal, or that “appropriate safeguards” have been provided in respect of the data subject’s rights, would improve legal certainty by a long way. . . Also research in the field of synthetic data shows enormous promise, and more funding should be funnelled into this area.”).

<sup>54</sup> Fast-Track Action Committee on Advancing Privacy-Preserving Data Sharing and Analytics, *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics*, (2023), <https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf> (“. . . PETs that are essential for enabling data sharing and analytics in a privacy preserving manner, such as secure multiparty computation, homomorphic encryption, differential privacy, zero knowledge proofs, synthetic data, federated learning, and trusted execution environments, which are discussed further in the document”).

<sup>55</sup> *OECD Roundtable Summary*, *supra* note \_\_, at 5.

<sup>56</sup> Walter H. L. Pinaya et al., *Generative AI for Medical Imaging: Extending the MONAI Framework* (27 July 2023), <https://arxiv.org/abs/2307.15208> (“Generative AI refers to a set of artificial intelligence techniques and models designed to learn the underlying patterns and structure of a dataset and generate new data points that plausibly could be part of the original

Generative AI refers to technology that is capable of creating content, such as images, video, audio, text, and computer code, by identifying patterns in large quantities of training data (input data), and then creating original material with similar characteristics (output data).<sup>57</sup>

The “predicted face” example given in the *Introduction* is generated with the aim of resembling the DNA sample owner as accurately as possible. At the same time, some generated images do not depict any specific person, or they are said to depict non-existent people. There are various generative artificial intelligence (GenAI) tools<sup>58</sup> that enable users to quickly generate images. For illustration, these images below were created using Stable Diffusion. To create images, users will provide descriptive text prompt. This prompt works as an instruction to the AI tool in generating an image, so it matches the user’s intent or wish.



Figure 1: “Albert Einstein presenting. . .”

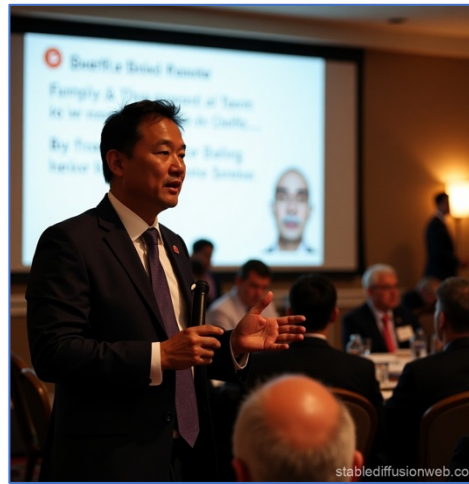


Figure 2: “[Name of the Author]. . .”

These images are two of many generated using one of popular AI image generators, Stable Diffusion Online.<sup>59</sup> The prompts used are: (1) “An image of Albert Einstein presenting an academic paper at a conference in Canada.”; and (2) “An image of [Name of the Author] presenting an academic paper at a conference in Canada.” For the record, the image on the right, which was generated with the author’s actual full name, does not resemble the author. Images generated with “Albert Einstein,” however, resembled the actual likeness of Albert Einstein, a theoretical physicist who is best known for developing the theory of relativity. The author also tried to generate various images using

dataset.”) [hereinafter *Generative AI for Medical Imaging*].

<sup>57</sup> Adam Pasick, *Artificial Intelligence Glossary: Neural Networks and Other Terms Explained*, The New York Times (27th March 2023), <https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html>.

<sup>58</sup> For example, Stable Diffusion, Midjourney, or DALL-E for text-to-image gen AI tools, and Sora for text-to-video gen AI tools.

<sup>59</sup> Stable Diffusion Online, <https://stablediffusionweb.com/>.

prominent privacy law scholars, but none of them resembled the actual person.

Additionally, one can generate more creative images by providing prompts that depict the person performing actions they may not have performed in real life. The image on the right is generated with the prompt, “Albert Einstein is playing table tennis.”

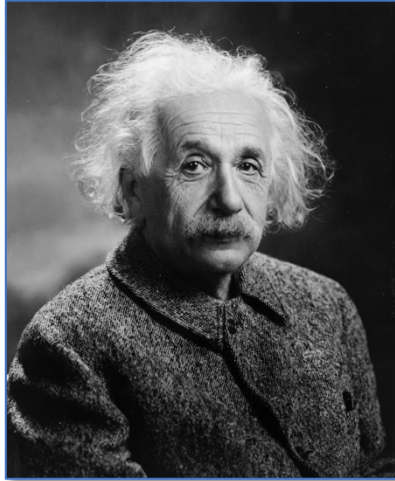


Figure 3: Picture of Albert Einstein from Wikipedia



Figure 4: “Albert Einstein is playing table tennis.”

The image on the left is a picture of Audrey Hepburn, a famous actor who was born in Brussels, Belgium. The image on the right is generated using the prompt, “Audrey Hepburn likes playing soccer.”



Figure 5: Picture of Audrey Hepburn from Wikipedia



Figure 6: “Audrey Hepburn is playing soccer.”

Misuse of GenAI is commonly referred to as deepfakes,<sup>60</sup> i.e., images, videos,

<sup>60</sup> See Chesney, Robert and Citron, Danielle Keats, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 California Law Review 1753 (2019), U of Texas Law, Public Law Research Paper No. 692, U of Maryland Legal Studies Research Paper No. 2018-21,

audio that are edited or generated using AI tools, and the outputs also constitute generated personal data. Some cases involving deepfakes have drawn public attention because they caused harm or were problematic from an ethical, legal, or social perspective. A Hong Kong company lost \$25 million due to a deepfake impersonating its CFO during a video call.<sup>61</sup> Fake, sexually explicit images of celebrity Taylor Swift, likely generated by artificial intelligence, spread rapidly across social media platforms.<sup>62</sup> Sexually explicit AI-generated deepfake images or videos are specifically referred to as deepfake pornography.<sup>63</sup>

In addition to algorithmically generated photos, videos, or audio, various scores<sup>64</sup> assigned to individuals are also examples of generated personal data.<sup>65</sup> One of the most well-known and used is the credit score.<sup>66</sup> Other scores include calculation of customer profitability based on the ratio between number of visits and the items bought,<sup>67</sup> life expectancy scores,<sup>68</sup> likelihood of future health outcomes,<sup>69</sup> recidivism risk score,<sup>70</sup> or likelihood of students to graduate from high school on time.<sup>71</sup>

## D. WHO GENERATES PERSONAL DATA

One of the often-overlooked entities that constantly generate, and sell, personal data are “data brokers.”<sup>72</sup> Data brokers are also known as “information

---

<http://dx.doi.org/10.2139/ssrn.3213954>. [hereinafter *Deep Fakes*].

<sup>61</sup> Heather Chen and Kathleen Magramo, *Finance worker pays out \$25 million after video call with deepfake ‘chief financial officer,’* CNN (4th February 2024), <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>; Nick Robins-Early, *CEO of world’s biggest ad firm targeted by deepfake scam*, Guardian (10th May 2024), <https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam>.

<sup>62</sup> Kate Conger and John Yoon, *Fake Explicit Taylor Swift Images Swamp Social Media*, The New York Times, <https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html>.

<sup>63</sup> See Danielle Keats Citron, *Sexual Privacy*, 92 (2018), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3233805](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233805); *Deep Fakes*, *supra* note \_\_; E.g., Roger Segarra et al., *Spain seeks to adapt its regulations to the artificial intelligence era* (27th Nov 2023), <https://www.osborneclarke.com/insights/spain-seeks-adapt-its-regulations-artificial-intelligence-era> (last visited 30 September 2024).

<sup>64</sup> See generally Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1, 9 (2014).

<sup>65</sup> See generally Hideyuki Matsumi & Daniel J. Solove, *The Prediction Society: AI and the Problems of Forecasting the Future*, forthcoming in U. Illinois L. Rev. (2025) [hereinafter *The Prediction Society*].

<sup>66</sup> *Personal Data: The Emergency of a New Asset Class*, *supra* note \_\_, at 14.

<sup>67</sup> *OECD Roundtable Summary*, *supra* note \_\_, at 5.

<sup>68</sup> *Id.*

<sup>69</sup> *Id.*

<sup>70</sup> Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk score. See *The Prediction Society*, *supra* note \_\_, at 14.

<sup>71</sup> Dropout Early Warning System (DEWS) score. See *The Prediction Society*, *supra* note \_\_, at 18.

<sup>72</sup> See Aaron Rieke et al., *Data Brokers in an Open Society*, (2016), <https://www.opensocietyfoundations.org/publications/data-brokers-open-society> [hereinafter *Data Brokers in an Open Society*].

resellers,” “data vendors,” “information brokers,” “consumer data analytics,” or “data warehousing.”<sup>73</sup> While the definition of data broker is yet to be settled,<sup>74</sup> *Data Brokers in an Open Society* used the definition: (1) A company or business unit (2) that earns its primary revenue (3) by supplying data or inferences about people (4) gathered mainly from sources other than the data subjects themselves.<sup>75</sup> The fourth element distinguishes data brokers from other tech giants that often draw public attention regarding data protection and privacy concerns. They do not interact directly with individuals; the sources of their data are not directly from individuals.

Yet, data brokers not only collect data, but also generate and sell data about individuals. Their “products consist of both ‘actual’ and ‘modeled’ data.”<sup>76</sup> Modeled data is the result of profiling, which means inferences made about individuals.<sup>77</sup>

## E. GENERATED PERSONAL DATA AND OTHER AREAS OF LAW

Generated personal data raises numerous issues in various areas of law.

One such a law is EU’s Artificial Intelligence Act (AI Act).<sup>78</sup> Unsurprisingly, the AI Act is relevant to generated personal data. In a nutshell, the AI Act regulates “AI systems.” By adopting a risk-based approach, it classifies AI systems into different risk levels, with tailored requirements and obligations for each level. AI practices that pose “unacceptable risks,” the highest risk category of all four, are prohibited.<sup>79</sup> The AI Act imposes various obligations<sup>80</sup> on “high-risk AI systems,” the highest risk category of the four, which is the key focus of the AI Act.<sup>81</sup> The AI Act imposes information and transparency requirements on AI systems that pose limited risks due to their lack of transparency.<sup>82</sup> The AI Act

---

<sup>73</sup> *Id.*, at 4 (“In Europe, the term data broker is less common than in the US. European commentators use a variety of different terms to refer to data brokers, including ‘information resellers,’ ‘data vendors,’ ‘information brokers,’ ‘consumer data analytics,’ ‘data warehousing,’ ‘Datenhändler’ (German), and ‘traders de données’ (French).”).

<sup>74</sup> *Id.* The FTC defines as “companies that collect consumers’ personal information and resell or share that information with others,” and the European Data Protection Supervisor (EDPS) defines as entities that “collect personal information about consumers and sell that information to other organisations.”

<sup>75</sup> *Id.*

<sup>76</sup> *Id.*, at 11 (“Actual data is factual information about people, such as their name, contact information, demographics and other behavioral data. Modeled data is the result of profiling (i.e., inferences or guesses about people based on actual data). For example, a data broker might infer a person is a woman based on her shopping habits. Or, a data broker might infer that a person is likely to default on a loan based on her past financial behavior.”).

<sup>77</sup> *Id.*, at 5.

<sup>78</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.

<sup>79</sup> Article 5 (Prohibited AI Practices), AI Act.

<sup>80</sup> Chapter III (High-Risk AI Systems), AI Act.

<sup>81</sup> The entire Chapter III of the AI Act is on high-risk AI systems, and many of the obligations are on the “provider” (i.e., developers) of such AI systems.

<sup>82</sup> See European Commission, *AI Act | Shaping Europe’s digital future*, <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (last visited 31 May 2024).

does not impose any obligations on AI systems classified as minimal-risk, which is the lowest risk category.<sup>83</sup>

The key element in all four risk classifications is the term “AI systems.” While definitions are crucial in any law, this is especially true in the AI Act because whether a particular technology amounts to an “AI system” or whether it only amounts to a traditional software system, for example, determines whether the AI Act applies in the first place. The importance of this threshold question is similar to determining whether a particular piece of data constitutes personal data, as it determines whether the GDPR applies.

The term “AI systems” is defined as:

a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, *infers*, from the input it receives, how to *generate outputs such as predictions, content, recommendations, or decisions* that can influence physical or virtual environments;<sup>84</sup>

The text of definition suggests<sup>85</sup> that “AI system,” or “a machine-based system,” comprises at least these four elements:

1. *Autonomy*. The wording of “is designed to operate with varying levels of autonomy” suggests there should be some level of autonomy. I.e., a machine capable of completing tasks autonomously, at least to some extent.
2. *Adaptability*. The wording “may exhibit adaptiveness after deployment” indicates that, although not necessarily, it is capable of adapting itself to its context, environment, or situation in which it is deployed. This means that AI systems are capable of learning their context and performing differently depending on where they are deployed. An AI system deployed in setting X can learn its context and perform differently from AI systems deployed in setting Y.
3. *Inference*. While the objectives can be explicit or implicit, “*infers* how to *generate outputs* from the input it receives” signals that AI systems should be capable of *making inferences and generating outputs*. The definition enumerates likely non-exhaustive examples, such as predictions, content, recommendations, or decisions.
4. *Influence*. The wording of “can influence physical or virtual environments” suggests that, although not necessarily, it is capable of exerting influence on its physical surroundings (e.g., in the case of robots) or on the virtual environment.

The third element -- Inference -- is particularly important for the purpose of this Article because systems capable of making inferences and generating outputs, including personal data, can constitute AI Systems depending on other

---

<sup>83</sup> AI-enabled video games or spam filters are examples of such minimal-risk AI systems.

<sup>84</sup> Article 3(1), AI Act (emphasis added).

<sup>85</sup> As the Commission is charged with the mission to issue guidelines on how the AI Act applies in practice, it is certainly too early to conclude what constitutes “AI system.”

elements, which the AI Act will be applicable, which triggers various provisions, such as Article 10 that articulates about data and data governance. The AI Act also refers to deepfake<sup>86</sup> and provides certain obligations.<sup>87</sup>

The interplay between generated personal data and the AI Act, however, is not discussed in this Article/Chapter, partly because this Article/Chapter focuses primarily on the interplay between generated personal data and the GDPR.<sup>88</sup> The AI Act's potential applicability to generated personal data does not necessarily resolve all issues related to such data under the GDPR.

Similarly, issues under other area of laws, such as copyright law and image rights,<sup>89</sup> tort and defamation law,<sup>90</sup> and the Digital Services Act and other EU directives,<sup>91</sup> as well as other measures and initiatives,<sup>92</sup> are beyond the scope of this Article/Chapter.

## II. GENERATED PERSONAL DATA AND THE GDPR

This Part explores how generated personal data might be treated under the General Data Protection Regulation (GDPR). The GDPR “applies to the processing of personal data.”<sup>93</sup> Thus, to analyze how the GDPR might apply to generated personal data, the first two questions that must be asked are: (1) whether generated personal data falls within the definition of *personal data*; and (2) whether generating personal data falls within the definition of *processing*. Additionally, because generation of personal data may be regarded as profiling, the third question would be to ask: (3) whether generation of

<sup>86</sup> Cf. Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on combating violence against women and domestic violence (COM/2022/105 final) has been

<sup>87</sup> Article 3(60), Article 50, Recital 134, AI Act.

<sup>88</sup> Generated personal data and other areas of law, including the AI Act, would be explored in other Chapters of this project.

<sup>89</sup> E.g., Frederick Mostert & Sheyna Cruz, *Image Rights in the Digital Universe*, 13 (2022), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4026437](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4026437) (“Synthetic media refers to content generated or modified by artificial intelligence (AI). . . In the right of publicity case of In re NCAA Student-Athlete Name & Likeness Licensing Litigation concerning video game depictions of college athletes, the defendant could not rely on the ‘transformative use’ defence as they had sought to portray the plaintiff as realistically as possible. Claimants invoking their image rights against deepfake uses of their image may be able to rely by analogy on the treatment of video games as expressive works which are subject to image rights. . .”).

<sup>90</sup> E.g., Hideyuki MATSUMI, *Prediction as Defamation*, presented at *We Robot 2022* in Seattle, Washington (Unpublished manuscript; On file with author) (discussing issues under US defamation law, such as whether machine-generated personal data constitutes an opinion or statement of fact).

<sup>91</sup> E.g., e-Commerce Directive 2000/31/EC; Audiovisual Media Services Directive 2018/1808 (AVMSD); or the Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on combating violence against women and domestic violence (COM/2022/105 final).

<sup>92</sup> E.g., Code of Practice on Disinformation (2018). See generally European Parliamentary Research Service, *Tackling Deepfakes in European Policy*, European Parliament, July 2021, [https://www.europarl.europa.eu/thinktank/en/document/EPRS\\_STU\(2021\)690039](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039).

<sup>93</sup> Article 2(1), GDPR.

personal data falls within the definition of *profiling*.

Note that, however, one cannot categorically determine if generated personal data, as a class of data, constitutes personal data because the concept is broad, and some elements may be lacking depending on facts. Thus, this Article aims to identify some of the relevant issues concerning the interplay between the GDPR and generated personal data.

## A. GENERATED PERSONAL DATA AS PERSONAL DATA

### 1. “Predicted Face” as Biometric Data for Identification Purpose

Consider the examples introduced in the *Introduction*. In the second example, detectives asked a third-party company to generate “predicted face” of a potential suspect by using DNA sample found at the crime scene of a homicide and sexual assault victim.<sup>94</sup> Subsequently, they asked to have the “predicted face” run through a facial recognition system to identify the suspect.

Under these facts, it seems quite straightforward that the “predicted face,” for identification purpose,<sup>95</sup> constitutes biometric data under the GDPR.<sup>96</sup> Because biometric data as defined in Article 4(14) is a subcategory of personal data, the same questions must be asked to determine if it is personal.<sup>97</sup> In the context of biometric identification, the person is generally identifiable, as biometric data are used to identify or authenticate/verify that the data subject is distinct from any other individual.<sup>98</sup>

In the present case, algorithmically generated data, i.e., “generated face,” is considered personal so long as it is sufficiently accurate that <sup>99</sup>

### 2. Visual image as Personal Data

Consider examples given in *Examples of Generated Personal Data*. These were

---

<sup>94</sup> Here, issues concerning use of genetic data and for the purpose of law enforcement are not within the scope of this Chapter.

<sup>95</sup> Identification of a person refers to “establishing who a person is relative to other persons.” *GDPR Commentary*, *supra* note \_\_, at 213.

<sup>96</sup> Article 4(14), GDPR (“personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;”).

<sup>97</sup> For the detailed analysis on biometric data as personal data, see *GDPR Commentary*, *supra* note \_\_, pp 214 - 215.

<sup>98</sup> Article 29 Data Protection Working Party, *Working Document on Biometrics (WP 80)*, (2003), [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2003/wp80\\_en.pdf](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2003/wp80_en.pdf), at 5 [hereinafter *Working Document on Biometrics*].

<sup>99</sup> *Working Document on Biometrics*, *supra* note \_\_, at 5 (“The identifiability of the person also depends on the availability of other data which -- jointly or separately -- allows the person in question to be identified.”).

images generated using widely available generative AI tools with slightly different prompts: images created using the real name of the author; using the real name of prominent privacy law scholars; and images generated with public figures, such as “Barack Obama,” including misspells.

The GDPR defines “personal data” as “any information relating to an identified or identifiable natural person (‘data subject’).<sup>100</sup> An identifiable natural person means “one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.”<sup>101</sup> This definition derives four elements: (1) “any information”; (2) “relating to”; (3) “identified or identifiable”; (4) “natural person.”<sup>102</sup>

**First element: “any information.”** Any information, including untrue<sup>103</sup> information as well as visual images, can qualify as personal data. In *Ryneš*,<sup>104</sup> the CJEU held that the image of a person recorded by a camera constitutes “personal data” to the extent that it makes it possible to identify the person concerned.<sup>105</sup>

Similarly, in *Buivids*,<sup>106</sup> the CJEU found that recorded images of persons in the video constitute personal data because it is possible to see and hear the police officers in the video in question,<sup>107</sup> applying the same test.

These cases concern videos or photos in which individuals were actually recorded (collected personal data), rather than generated. So far, the author was unable to find cases in which involves drawing, illustration, or painting -- manually generated data or information about individuals.

Nevertheless, there is no apparent reason to deny that algorithmically generated personal data meets this “any information” element, considering how the CJEU has found an array of information to give rise to personal data.

**Second element: “relating to.”** The second element asks whether information at issue is sufficiently linked to a natural person to amount to their personal data. This element “is satisfied where the information, by reason of its

---

<sup>100</sup> Article 4(1), GDPR.

<sup>101</sup> *Id.*

<sup>102</sup> See Lee A. Bygrave, *Article 4(1). Personal data* in *The EU General Data Protection Regulation (GDPR): A Commentary*, (Christopher Kuner, Lee A. Bygrave, & Christopher Docksey eds., 2019) [hereinafter *GDPR Commentary*].

<sup>103</sup> Capturing fake images.

<sup>104</sup> Case C-212/13, *František Ryneš v Úřad pro ochranu osobních údajů* (ECLI:EU:C:2014:2428) [hereinafter *Ryneš*].

<sup>105</sup> *Ryneš*, *supra* note \_\_, at para 22.

<sup>106</sup> Case C-345/17, *Proceedings brought by Sergejs Buivids* (ECLI:EU:C:2019:122) [hereinafter *Buivids*].

<sup>107</sup> *Buivids*, *supra* note \_\_, at para 32.

content, purpose or effect, is linked to a particular person.” *Nowak*.<sup>108</sup>

The Article 29 Data Protection Working Party (“Art. 29 WP”) takes the same view in the *Opinion 04/2007 on the Concept of Personal Data*.<sup>109</sup> The element of “relating to” is sufficed if one of those three elements -- the “content” element or the “purpose” element or the “result”<sup>110</sup> element is sufficed.<sup>111</sup> These conditions can be met simultaneously, which makes information relating to multiple individuals.<sup>112</sup>

The “content” element is sufficed if the content of information is given about a particular person, irrespective of “any purpose on the side of the data controller or of a third party, or the impact of that information on the data subject.”<sup>113</sup>

Thus, a straightforward interpretation suggests that, even if the information in question -- e.g., a generated image depicting a particular person -- is not genuinely real, it may be nevertheless found to be “about a particular person” so long as it is sufficiently accurate. If an average person sees the image and describes it as “it’s an image of Obama presenting a paper,” “making a speech,” or “playing soccer,” then it is reasonable to consider it as an image “about that person.” The other example, however, unlikely suffices the element, as the person depicted in the image did not resemble the author. The same for prominent law scholars.

The “purpose” element focuses on the purpose for which information is used or likely to be used. This element is sufficed if the information is used or expected to be used to assess, impact, or shape an individual’s circumstances or actions.<sup>114</sup> This element is likely satisfied in the cases of “predicted face,” where the purpose is to identify suspects or litterbugs. However, it would be unclear if this element is satisfied if an image is generated just for the sake of generating it.

The “result” element is satisfied if the use of information is likely to have an impact on an individual’s rights and interests, considering all the surrounding circumstances. The impact need not be major; it is sufficient if the individual might be treated differently from other people as a result of the information.

This test may have unintended consequences due to its expansive nature. Because the test simply asks if there is an impact or effect on a person, anyone who appears similar, looks alike, or resembles the individual in question may be the subject of “information relating to.”

---

<sup>108</sup> *Nowak*, *supra* note \_\_, at para 35.

<sup>109</sup> Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, (2007), [https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/index\\_en.htm](https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/index_en.htm) [hereinafter *Opinion on Personal Data*].

<sup>110</sup> Note that Art. 29 WP uses the term “result,” while CJEU uses the term “effect.”

<sup>111</sup> *Opinion on Personal Data*, *supra* note \_\_, at 10.

<sup>112</sup> *Id.*, at 11 (“The same information may relate to individual Titius because of the ‘content’ element (the data is clearly about Titius), AND to Gaius because of the ‘purpose’ element (it will be used in order to treat Gaius in a certain way) AND to Sempronius because of the ‘result’ element (it is likely to have an impact on the rights and interests of Sempronius”).

<sup>113</sup> *Id.*, at 10.

<sup>114</sup> *Id.*, at 10.

Contrast this with the cases involving actual pictures taken by camera. Similar to the case of generated personal data, unrelated people who happen to look like the person in the picture could be considered the subject of “information relating to.” In these cases, however, the pictures have to be actually taken. Whereas the advent and advancement of generative AI enable the generation of realistic photos of a person under specific conditions in infinite ways.

**Third element: “identified or identifiable.”** This element asks if a person within a group can be *distinguished* from all other members of the group.<sup>115</sup> One way how this element may be relevant in the context of generated personal data is, how accurate it is. The European Data Protection Board (EDPB)’s Guidelines on Processing of Personal Data through Video Devices,<sup>116</sup> notes that, generally, visual information or video footage of individuals are “identifiable on basis of their looks.”<sup>117</sup> Conversely, video footage is not deemed personal data if it cannot be related to a specific person because, for example, recordings were made from a high altitude.<sup>118</sup>

**Fourth elements: “natural person.”** This element excludes information relating to deceased individuals from personal data. There are, however, issues involving generated photos, videos, or audio relating to deceased individuals. . .

### 3. Input Data Used to Generate and Output Data Generated: Mixed Personal Data?

One of the challenges in analyzing whether generated personal data constitutes personal data, or where it fits within the data protection regime, arises from how GenAI functions. In short, to generate an image, millions of actual human faces would have to be used in the training process of that generative AI.

Generative AI, which refers to a set of artificial intelligence techniques and models, is designed to learn the underlying patterns and structure of a dataset (input data) and generate new data (output data) that plausibly could be part of the original dataset.<sup>119</sup> In *What Makes Data Personal?*,<sup>120</sup> Montagnani and Verstraete explain it as:

There are a few steps that companies use to *create inferences*. To start, the *underlying data (source data)* that is used to *generate an inference* is *collected*. Next, the *source data* is prepared to be inputted into an AI

<sup>115</sup> Recital 26, GDPR. *Opinion on Personal Data*, *supra* note \_\_, at 12.

<sup>116</sup> European Data Protection Board, *Guidelines 3/2019 on Processing of Personal Data through Video Devices*, (2020), [https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32019-processing-personal-data-through-video\\_en](https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-32019-processing-personal-data-through-video_en) [hereinafter ]

<sup>117</sup> *Id.*, para 7, at 7.

<sup>118</sup> *Id.*

<sup>119</sup> *Generative AI for Medical Imaging*, *supra* note \_\_, at 2.

<sup>120</sup> Maria Lilla Montagnani and Mark Verstraete, *What Makes Data Personal?* (June 4, 2022). UC Davis Law Review, Vol. 56, No. 3, 2023 Bocconi Legal Studies Research Paper No. 4128080, <http://dx.doi.org/10.2139/ssrn.4128080> (hereinafter *What Makes Data Personal?*).

tool or analytical model. After that, the *source data* trains the AI tool or analytical model. Finally, the *source data* is inputted into an AI tool or analytical model to *create an inference*.<sup>121</sup>

Thus, the two tiers of data must be analyzed: the input data used to train the model, and the output data generated by GenAI tools.

This means that, for example, building a generative AI model capable of producing pictures of Barack Obama -- some of which depict him in places he hasn't been or doing things he hasn't done -- would have to be built with actual pictures of Barack Obama. In other words, if he is making a speech on the moon, a generative AI model needs to be trained with pictures of the moon and Barack Obama.

Considering the following, as more of a thought experiment. Because a generative AI model is trained with pictures of Barack Obama, along with millions of other individuals (input data), it is capable of generating images of Barack Obama (output data) where he is doing things he hasn't done, but others have done: E.g., playing soccer,<sup>122</sup> climbing, or paragliding. When an image generated with a prompt, "Barack Obama is playing soccer," the AI model is technically remixing,<sup>123</sup> rather than "creating" like humans draw or paint. During the training phase, pictures of individuals playing soccer are used, and many of these images constitute personal data because the players can be identified through various means, such as their faces, body parts, including tattoos, or uniforms.

---

<sup>121</sup> *Id.*, at 1221 (citing Joe O'Callaghan, *Inferential Privacy and Artificial Intelligence - A New Frontier?*, 11 J.L. & ECON. REGUL. 72) (emphasis added).

<sup>122</sup> The author is not sure if he's been pictured when playing soccer, but I chose this as an example because soccer is a very popular sport in Europe, and there can be many well-known professional soccer players, which may have been used to train the generative AI model.

<sup>123</sup> Clément Stenac, Council Post: AI Remix: Generative AI Is Incredible—And A Con, AI Remix: Generative AI Is Incredible—And A Con (Jan. 13, 2023), <https://www.forbes.com/councils/forbestechcouncil/2023/01/13/ai-remix-generative-ai-is-incredible-and-a-con/> (last visited Oct 11, 2024).



Figure 7: “Barack Obama is playing soccer”

Pictures of Barack Obama (input data) constitute personal data. Naturally, the generated image (output data), even if it depicts him making a speech on the moon or playing soccer, will still constitute personal data.

A question: Does, or should, the same apply to the professional soccer player, or players, who have been used to train the AI model and been remixed to “generate” the image? Or, do images of these soccer players lose their status as personal data once they are remixed and become part of the output image (i.e., Barack Obama playing soccer), which happens to be more well-known and more *identifiable* than the original players?

Or, theoretically, are such images excluded from the definition of personal data because it is not “linked to a particular person,”<sup>124</sup> rather linked to an “aggregate of persons”?<sup>125</sup>

A similar concern exists, or is accentuated, in deepfake pornography cases. Often, deepfake pornography is created with the faces of targeted individuals, combined with body parts from many who are sexual workers. While attention tends to be drawn to issues concerning individuals whose faces have been used<sup>126</sup> to generate deepfake pornography, the issues concerning individuals whose body parts have been misused or exploited also deserve attention.

One way to approach this issue is to follow the tests articulated in *Nowak*. That is, to focus on the output data, and to ask whether the generated data at issue is linked to a particular person “by reason of its content, purpose or effect.”<sup>127</sup>

---

<sup>124</sup> *Nowak*, *supra* note \_\_, at para 35.

<sup>125</sup> *GDPR Commentary*, *supra* note \_\_, at 110.

<sup>126</sup> In the Taylor Swift deepfake pornography case, controversies tended to focus predominantly on the celebrity Taylor Swift, but virtually no attention was given to the individuals whose body parts were used to generate such deepfakes.

<sup>127</sup> *Nowak*, *supra* note \_\_, at para 35.

This approach, however, overlooks the technical issues regarding how generative AI works, including the relationship between training (input) data and generated (output) data in the context of data protection and privacy.

DPAAs seem to be struggling with similar issues. Shortly after ChatGPT, one of well-known services by OpenAI, which is a generative AI chatbot, introduced into the EU market, the Italian DPA ordered a temporary ban on ChatGPT due to concerns that it violates the GDPR.<sup>128</sup> Subsequently, the ban was lifted because the company, *inter alia*, implemented a mechanism for users and non-users to remove their personal data to be used to train the algorithm of ChatGPT. Measures taken by the company may solve (some) problems with training data (input data), but the question remains open as to whose personal data is the generated data (output data).

When ChatGPT (GPT stands for Generative Pre-trained Transformer, another GenAI) was introduced into the EU market, the Italian DPA ordered a temporary ban on ChatGPT due to concerns that it violates the GDPR. Subsequently, the ban was lifted<sup>129</sup> because the company, *inter alia*, implemented a mechanism for users and non-users to remove their personal data to be used to train the algorithm of ChatGPT.<sup>130</sup> Measures taken by the company may solve (some) problems with training data (input data), but the question remains open concerning generated data (output data).

Another approach is to draw from the concept of mixed personal data and rules concerning such data. But this is only a starting point. For clarification, mixed datasets and mixed personal data are a bit confusing. Mixed datasets refer to datasets consisting of both personal and non-personal data,<sup>131</sup> and the Regulation 2018/1807 on a framework for the free flow of non-personal data<sup>132</sup> governs such data. The Regulation 2018/1807 applies to non-personal data part of the data set, while the GDPR applies to if personal and non-personal data in a data set are inextricably linked.

Whereas mixed personal data refers to data that relates to more than one person.<sup>133</sup> Issues concerning personal data that relates to more than one person

---

<sup>128</sup> Shiona McCallum, *ChatGPT banned in Italy over privacy concerns*, BBC (1 April 2023), <https://www.bbc.com/news/technology-65139406>.

<sup>129</sup> Shiona McCallum, *ChatGPT accessible again in Italy*, BBC (28 April 2023), <https://www.bbc.com/news/technology-65431914>.

<sup>130</sup> ChatGPT: OpenAI riapre la piattaforma in Italia garantendo più... - Garante Privacy, <https://www.gdpd.it/home/docweb/-/docweb-display/docweb/9881490>.

<sup>131</sup> *GDPR Commentary*, *supra* note \_\_, at 112.

<sup>132</sup> Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, 303 OJ L (2018), <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1807>.

<sup>133</sup> *Getting Data Subject Rights Right*, *supra* note \_\_, at para 51.

are not new.<sup>134</sup> Outside the EU, the European Court of Human Rights (ECtHR) has addressed the issue in relation to the right to access.<sup>135</sup> The ICO has also provided a guidance on the right of access, including if the request involves information about other individuals.<sup>136</sup> In a case where a dataset is composed of both personal and non-personal data, the Regulation 2018/1807 on a framework for the free flow of non-personal data governs. Apparently, there are no cases from the CJEU or DPAs on this issue yet.<sup>137</sup>

#### 4. Generated Personal Data as “Non-Portable Data”?

Last, it should be noted that, even if generated personal data constitutes personal data, it may be considered as “second-class personal data” because it does not receive the same protection as collected personal data.

Generated personal data is different from other forms of personal data, such as collected ones, under the right to “data portability” under the GDPR.

Article 20 of the GDPR creates a new right to data portability -- the right to receive personal data, in a structured, commonly used, and machine-readable format. By exercising this right, individuals can oblige data controllers to furnish personal data that they have provided to the controller, or to transmit those data to another data controller.

The GDPR defines the right of data portability in Article 20 (1) as follows:

The data subject shall have the right to receive the personal data concerning him or her, which he or she has *provided* to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the data have been *provided*. . .<sup>138</sup>

The key text of the Article is “which he or she has *provided* to a controller.” Because of this text, the right to data portability does not treat “ordinary” personal data and generated personal data equally. The *Guidelines on the right to “data portability”*<sup>139</sup> explains this by breaking down to two conditions: (1) personal data concerning him or her; and (2) which he or she has *provided* to a data controller.<sup>140</sup> Hence, the Art. 29 WP states that the right to data portability does not apply to “‘inferred data’ and ‘derived data’, which include personal data

---

<sup>134</sup> See *Getting Data Subject Rights Right*, *supra* note \_\_, at paras 52 - 53.

<sup>135</sup> *Gaskin v United Kingdom* [1990] EHRR 36; *Társaság a Szabadságjogokért v Hungary* App no 37374/05 (2009).

<sup>136</sup> ICO, *What should we do if the request involves information about other individuals?*, The Right of Access, <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/right-of-access/information-about-other-individuals/>.

<sup>137</sup> Cf. Spanish case, AEPD (Spain) - PS/00281/2022, involves mixed datasets (personal and non-personal) was the case found in the search result at GDPRhub.

<sup>138</sup> Article 20, GDPR (emphasis added).

<sup>139</sup> Article 29 Data Protection Working Party (Art. 29 WP), *Guidelines on the Right to “Data Portability”* (Wp242rev.01), (2017), <https://ec.europa.eu/newsroom/article29/items/611233> [hereinafter *Right to Data Portability Guidelines*].

<sup>140</sup> *Id.*, at 9.

that are created by a service provider (for example, algorithmic results).”

## **B. GENERATING PERSONAL DATA AS PROCESSING OF PERSONAL DATA**

Generating personal data falls within that broad definition of “processing.” Processing under the GDPR is a broad concept. Virtually, “any operation” performed on personal data constitutes “processing.”<sup>141</sup> The definition enumerates a non-exhaustive<sup>142</sup> list of examples:

collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction<sup>143</sup>

In addition to this list, the Court of Justice of the European Union (CJEU) has found numerous data operations to be “processing.”

In the case of *Google Spain*,<sup>144</sup> the CJEU found that “in exploring the internet automatically, constantly and systematically in search of the information. . . , the operator of a search engine ‘collects’ such data which it subsequently ‘retrieves’, ‘records’ and ‘organises’ within the framework of its indexing programmes, ‘stores’ on its servers and, as the case may be, ‘discloses’ and ‘makes available’ to its users in the form of lists of search results.”<sup>145</sup> Here, the operator of a search engine (i.e., Google Inc.) is collecting information, but also generating “lists of search results,” which makes it relevant to generation of personal data.

In *Ryneš*,<sup>146</sup> the CJEU affirmed that “surveillance in the form of a video recording of persons. . . which is stored on a continuous recording device -- the hard disk drive -- constitutes, pursuant to Article 3(1) of Directive 95/46, the automatic processing of personal data.”<sup>147</sup> In *Buivids*,<sup>148</sup> the CJEU found that “a video recording of persons which is stored on a continuous recording device, namely the memory of that camera. . . constitutes a processing of personal data

---

<sup>141</sup> See generally *GDPR Commentary*, *supra* note \_\_, at 116.

<sup>142</sup> *Opinion of Adv. Gen. in YS and Others*.

<sup>143</sup> Article 4(2), GDPR (“any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;”). See also Recital 15, GDPR.

<sup>144</sup> Case C-131/12, *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* (ECLI:EU:C:2014:317), paras. 26 - 31.

<sup>145</sup> *Id.*, at para 28.

<sup>146</sup> Case C-212/13, *František Ryneš v Úřad pro ochranu osobních údajů* (ECLI:EU:C:2014:2428) [hereinafter *Ryneš*].

<sup>147</sup> *Id.*, at paras 25, 35.

<sup>148</sup> Case C-345/17, *Proceedings brought by Sergejs Buivids* (ECLI:EU:C:2019:122) [hereinafter *Buivids*].

by automatic means. . .”<sup>149</sup> In *Pušár*,<sup>150</sup> the CJEU held that drawing up of a list -- i.e., creating a list -- which contains the names of certain natural persons, constitutes processing of personal data.<sup>151</sup> In *Buivids*,<sup>152</sup> the CJEU held that the “act of publishing a video recording. . . which contains personal data, on a video website on which users can send, watch and share videos” constitutes processing.<sup>153</sup>

The Art. 29 WP also follows the view of expansive definition of “processing.”<sup>154</sup> In *Approach To Data Spaces From GDPR Perspective*, the AEPD (the Spanish DPA) provides a non-exhaustive list of various processing operations. AEPD enumerates, *inter alia*, “extraction of data from datasets for the *creation of new datasets*” and “*generation of synthetic data*.”<sup>155</sup>

\* \* \*

A very expansive definition of “processing” has its pros and cons. The advantage is that relatively new forms of operation, such as generation of personal data, can be captured within its long arm. The disadvantage is, however, it will be difficult to articulate different rules for different types of operations, such as generation of personal data.

As discussed in the next Part, the generation of personal data has different qualities and consequences compared to other forms of operations, such as the traditional collection of personal data. Conflating the two will render the GDPR more difficult and less effective in addressing the challenges raised by the generation of personal data.

### C. GENERATING PERSONAL DATA AS PROFILING

Profiling under the GDPR is very closely related to, but has a different perimeter compared to, the concept of generated data. Importantly, however, not all profiling constitutes generation of personal data. As discussed below, not all generation of personal data falls within the definition of profiling. Consequently, some types of personal data generation will be covered by the profiling provisions under the GDPR, while some types will not.

---

<sup>149</sup> *Id.*, at para. 35.

<sup>150</sup> Case C-73/16, *Peter Pušár v Finančné riaditeľstvo Slovenskej republiky and Kriminálny úrad finančnej správy* (ECLI:EU:C:2017:725).

<sup>151</sup> *Id.*, at para. 103.

<sup>152</sup> Case C-345/17, *Proceedings brought by Sergejs Buivids* (ECLI:EU:C:2019:122).

<sup>153</sup> *Id.*, para. 39.

<sup>154</sup> See Article 29 Data Protection Working Party (Art. 29 WP), Opinion 01/2015 on Privacy and Data Protection Issues relating to the Utilisation of Drones, <https://ec.europa.eu/newsroom/article29/items/640602/en>.

<sup>155</sup> *Approach To Data Spaces From GDPR Perspective*, *supra* note \_\_, at 28 (emphasis added).

## 1. Generating Personal Data as Profiling

Profiling under the GDPR is:

any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.<sup>156</sup>

It is composed of three elements: (1) *automated* form of processing; (2) carried out on personal data; and (3) the objective is *to evaluate personal aspects* about a natural person.<sup>157</sup>

Since generation of personal data is an automated form of processing carried out on personal data, the third element -- the objective or evaluation condition -- will be the issue. If this third element is not sufficed, then generating personal data will not constitute profiling, and consequently, the profiling rules under the GDPR will not apply. Therefore, what amounts to "*to evaluate personal aspects*" is a material question.

The objective condition seems to encompass both the process of profile generation and profile application.<sup>158</sup> Because "the verb "to evaluate" ordinarily connotes the act of judging, calculating or assessing,"<sup>159</sup> there must be some degree of judging, calculating, or assessing to satisfy this condition.

Because the definition under the GDPR does not include the criterion of "intention," unlike Article 15(1) of the Data Protection Directive (DPD),<sup>160</sup> the *GDPR Commentary* notes that the definition under the GDPR limits or reduces the possibility that profiles created merely as a side effect of automated processing are excluded from the definition, which was a limitation that arguably weakened Article 15(1) of the DPD,<sup>161</sup> by citing to *Profiling in the Present and New EU Data Protection Frameworks*.<sup>162</sup>

<sup>156</sup> Article 4(4), GDPR.

<sup>157</sup> *Profiling Guidelines*, at 6 (emphasis added).

<sup>158</sup> *GDPR Commentary*, *supra* note \_\_, at 130 ("During the legislative process, however, the evaluation component of the definition of 'profiling' was understood, at least by some Member States, as pertaining to the purpose for which a profile is created, rather than the means for profiling: see Council Report 2014, p. 7, fn. 2.").

<sup>159</sup> *Id.*, at 130.

<sup>160</sup> Which covered "processing of data intended to evaluate certain personal aspects." *GDPR Commentary*, *supra* note \_\_, at 130.

<sup>161</sup> *Id.*, at 130 ("This curtails the possibility whereby profiles that arise only as an ancillary effect of automated processing fall outside the scope of the definition—a possibility that arguably hobbled Article 15(1) DPD.").

<sup>162</sup> Andrej Savin, *Profiling in the Present and New EU Data Protection Frameworks*, (2015), <https://papers.ssrn.com/abstract=2697531> ("If the intention is not to evaluate personal aspects, the article [Article 15] does not apply. The key to application of the article is the intention understood as the data processor's awareness of and desire to analyse personal information. If personal information analysis is not the intended but ancillary effect, the article would not apply.

This does not necessarily mean, however, that the omission of the “intent” condition in the GDPR leads to the conclusion that profiles generated as “an ancillary effect of automated processing” will nevertheless fall within the definition of profiling absent “objective is to evaluate personal aspects” element.

On this issue, the *Profiling Guidelines* states that the term “evaluating” implies that profiling entails making some kind of assessment or judgment about an individual.<sup>163</sup>

The CJEU in *Schufa* case addressed the issue of whether generating credit scores constitute automated decision-making, as discussed in the next section. The concept of profiling, however, especially as to the meaning of “to evaluate certain personal aspects,” has not yet been discussed squarely by the CJEU.<sup>164</sup>

Thus, it is yet to be clarified as to what suffices the condition derived from the text of “to evaluate certain personal aspects.” But presuming that this “objective” condition requires some degree of evaluation for generation of personal data to constitute profiling, then some of them will fall outside the definition.

When companies or data brokers generate personal data with an objective to evaluate personal aspects of those individuals, their acts constitute profiling, which triggers various regulations under the GDPR.

Various scoring, including credit scoring, suffices the objective condition because scoring entails evaluating certain aspects of individuals, including person’s performance at work, economic situation, health, personal preferences, interests, reliability, behavior, and the like. In fact, neither of the parties in *Schufa* case contested that scoring by SCHUFA, i.e., generating personal data, constitutes profiling.<sup>165</sup> In addition to credit scores, various scoring, including recidivism risk score<sup>166</sup> or likelihood of students to graduate from high school on time,<sup>167</sup> can fall within the definition of profiling.

---

This is not an ideal solution. It is submitted here that intention should not form part of the provision. The individual should be able to object to automatic decision making based on personal data whether the data controller’s intention had originally been to analyse such data or not.”).

<sup>163</sup> *Profiling Guidelines*, at 7 (“The use of the word ‘evaluating’ suggests that profiling involves some form of assessment or judgement about a person.”).

<sup>164</sup> *B. What is Not Generated Personal Data, infra*. See also *GDPR Commentary, supra* note \_\_, at 130 (“The CJEU has not yet interpreted the definition of ‘profiling’ in the GDPR. . .”); *Getting Data Subject Rights Right, supra* note \_\_, at 45 (“Consequently, we have not seen the Court provide judgements clearly analogous to profiling.”).

<sup>165</sup> Opinion of Advocate General (ECLI:EU:C:2023:220) in *Schufa, supra* note \_\_, at para 33 (“Lastly, I would stress that none of the interested parties contests the classification of the procedure at issue as ‘profiling’, and this condition can therefore be considered to be satisfied in the present case.”) [hereinafter *Opinion of Adv. Gen. in Schufa*].

<sup>166</sup> E.g., Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk score. See *The Prediction Society, supra* note \_\_, at 14.

<sup>167</sup> E.g., Dropout Early Warning System (DEWS) score. See *The Prediction Society, supra* note \_\_,

The question becomes difficult to answer when it involves generated photos, videos, or audio of individuals, as they do not necessarily entail evaluating personal aspects about those individuals. It becomes especially difficult if those photos, videos, or audio are produced solely for the sake of production, absent evaluation aspect.

One way to avoid such a conclusion is to interpret “to evaluate certain personal aspects” as including predicting -- or inferring in this context -- physical or physiological characteristics,<sup>168</sup> such as one’s “predicted face.” The subsequent criterion of “in particular” indicates that evaluation does not necessarily have “to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.” Admittedly, this interpretation is a far stretch considering the history of legislative discussions concerning profiling, and compared to the enumerated items, even if it is a non-exhaustive list. That said, it is a way forward.

\* \* \*

The key points of this section are that: (1) the concept of profiling is narrower than that of generated personal data due to the “to evaluate personal aspects” element; and consequently, (2) some generated personal data, such as data generated without an evaluative aspect, may fall outside the definition of profiling, as well as the protection it provides.

## **2. Generating Personal Data as Solely Automated Decision-Making**

Generated personal data can constitute “solely automated decision-making or profiling” under the GDPR<sup>169</sup> if these generated personal data is relied on heavily that is equivalent of making a “decision,” like in the *Schufa* case.<sup>170</sup>

Other conditions must be met to trigger this rule.<sup>171</sup> Article 22 provides that the data subject is to have the right not to be subject to (1) a decision (2) based solely on automated processing, including profiling, which (3) produces legal effects concerning him or her or similarly significantly affects him or her.<sup>172</sup>

---

at 18.

<sup>168</sup> Cf. voiceprint.

<sup>169</sup> Article 22(1), GDPR.

<sup>170</sup> Case C-634/21 - *OQ v Land Hessen, SCHUFA Holding AG* (ECLI:EU:C:2023:957).

<sup>171</sup> The *Schufa* ruling, as well as the *Opinion of Adv. Gen.*, confirms that this is a general prohibition, rather than the right individuals must invoke. *Id.*, *supra* note \_\_, at para 52 (“That provision lays down a prohibition in principle, the infringement of which does not need to be invoked individually by such a person.”).

<sup>172</sup> *E.g.*, *Schufa*, *supra* note \_\_, at para 42.

While the concept of “decision” is not defined under the GDPR,<sup>173</sup> the *Schufa* case found that the concept is broad enough that it can include “calculating a person’s creditworthiness in the form of a probability value concerning that person’s ability to meet payment commitments in the future.”<sup>174</sup> Because facts in *Schufa* suggest that, alongside the mathematical and statistical procedure carried out by SCHUFA, no individual evaluation and assessment by a human was made when establishing the score,<sup>175</sup> and because “automated establishment of a probability value concerning person’s ability to repay a loan in the future” constitutes a “decision,” the Court found it is also a “solely automated decision.”<sup>176</sup>

Whether the second and third elements are sufficed, and therefore whether a decision at issue is general in nature, or turns out to be a “solely automated one with legal effects or similarly significant effects on an individual” triggering the protections offered under the GDPR, depends on the facts.

One point to note here is that, as the scope of “solely automated decision-making or profiling” is narrower than the concept of generated personal data discussed here, some forms of generating personal data will not trigger this rule.

As introduced in the *Introduction*, generating “predicted faces” from DNA samples, and publicly “naming and shaming” individuals likely satisfies the third element. Having “predicted face” run through a facial recognition system to identify potential suspects likely suffice as well.

In some other examples of generated personal data, however, are less likely to qualify as “solely automated decision-making or profiling,” as they lack either the second and/or the third element.

\* \* \*

For the purpose of this Article, there are three important points to take away from this Part. First, there is nothing that negates that output data of generative AI can fall within the definition of personal data. In the context of generated personal data, however, it is unclear on several accounts. What is the relationship between input data (source or training data) and output (generated) data? When the generated image (of an individual) is a remix or composite of several images (of individuals), does that still constitute “data relating to a particular person”?

Second, the very expansive concept of processing will encompass the act of generating personal data within its purview. However, this also introduces

---

<sup>173</sup> Cf. Recital 71, GDPR.

<sup>174</sup> *Schufa*, *supra* note \_\_, at para 46.

<sup>175</sup> *Opinion of Adv. Gen. in Schufa*, *supra* note \_\_, at 36.

<sup>176</sup> *Schufa*, *supra* note \_\_, at para 47.

another problem. By equating the generation of personal data with other forms of data operations, such as traditional data collection, it becomes more challenging to distinguish and specifically address the unique challenges raised by the generation of personal data.

Third, generating personal data can sometimes be considered as profiling under the GDPR, but this is not always true. Since profiling, as well as solely automated decision-making, has additional conditions that must be met to trigger respective provisions, some generated personal information will fall outside the definition, making the provisions not applicable. Generated personal data that lacks the third condition, i.e., the objective condition, then the protections do not apply.

### III. THE CHALLENGES RAISED BY GENERATED PERSONAL DATA

The concept of generated personal data not only raises complex and difficult issues,<sup>177</sup> but also exacerbates long-standing problems. For many years, numerous commentators have been discussing issues concerning profiling and inferred data.<sup>178</sup>

The advent and recent advancement of generative “artificial intelligence” (GenAI) will only accelerate and accentuate the problems and challenges posed by generated personal data. As Maria Lillà Montagnani and Mark Verstraete aptly observe that the development of data analytics and AI “marks a shift from companies primarily *collecting data* to companies *generating inferred data*,<sup>179</sup> generated personal data, as opposed to collected personal data, will become one of the main issues under data protection and privacy law. In fact, ICO has launched a consultation series on how aspects of data protection law should apply to the development and use of generative AI models.<sup>180</sup>

Of the many challenges and issues raised by generated personal data, this Part will explore three challenges very briefly.

---

<sup>177</sup> E.g., CEDPO AI Working Group, *Generative AI: The Data Protection Implications*, (2023), <https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf> [hereinafter *Generative AI: The Data Protection Implications*].

<sup>178</sup> E.g., Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 Nw. U. L. Rev. 357 (2022).

<sup>179</sup> Maria Lillà Montagnani and Mark Verstraete, *What Makes Data Personal?* (June 4, 2022). UC Davis Law Review, Vol. 56, No. 3, 2023 Bocconi Legal Studies Research Paper No. 4128080, <http://dx.doi.org/10.2139/ssrn.4128080> [hereinafter *What Makes Data Personal?*] (“*Inferred data* is increasingly important in the information economy. The development of cutting-edge data analytics and artificial intelligence (“AI”) marks a shift from companies primarily *collecting data* to companies *generating inferred data*. And further, the status of inferred data as personal data is deeply contested within information governance frameworks.”) (emphasis added).

<sup>180</sup> ICO, *Consultation series on generative AI and data protection*, <https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/> [ ].

## A. CHALLENGES TO THE CONCEPT OF PERSONAL DATA

In *The Law of Everything. Broad Concept of Personal Data and the Future of EU Data Protection Law*,<sup>181</sup> Nadezhda Purtova aptly argued that everything will be or will contain personal data, leading to the application of data protection to everything in the near future. The rise of generated personal data will push this problem further.

This Article has proceeded with the tone that if generated personal data is not included as personal data, there would be bad consequences because individuals' rights won't be protected. These are certainly legitimate concerns, and it is one of the approaches that this Project aims to take.

There is, however, the other side of the coin. A different, deeper concern is the implications of various generated personal data being simply considered as personal data under data protection/privacy law. One of the possible scenarios is that the concept of personal data will be further expanded or inflated, potentially leading to an overly broad and unwieldy definition of personal data. In turn, inconsistencies in principles as well as rights and duties can emerge. As discussed later, for example, the principle of accuracy and the right to rectification apply differently, or fail to apply, to generated personal data compared to collected personal data.

In such circumstances, data protection/privacy law would place various stakeholders involved in data protection and privacy issues -- such as individuals protected by it, companies seeking to comply with it, and courts and DPAs enforcing it -- in a difficult position when attempting to apply or interpret the law.

One of the challenges to the concept of personal data raised by generated personal data is the very foundation of the GDPR framework, which is predicated on an individual rights-based approach.

As with profiling was,<sup>182</sup> one of the challenges that the generated personal data raises is this notion of personal data relates to a particular person. At many levels and phases, generated personal data oscillates between data relating to a particular individual and data concerning multiple individuals or groups.

At a webinar on “synthetic data as a privacy enhancing technology” organized by the EDPS, Dara Hallinan made similar remarks, arguing that, *inter alia*, “the generation of synthetic group data profiles would also potentially fall outside the

---

<sup>181</sup> Nadezhda Purtova, *The Law of Everything. Broad Concept of Personal Data and the Future of EU Data Protection Law*, 10 *Law, Innovation, and Technology* 40, 43-45 (2018) [hereinafter *The Law of Everything*].

<sup>182</sup> *E.g.*, *Profiling the European citizen*, supra note \_\_, at \_\_.

individualistic framework of data protection law.”<sup>183</sup>

One way to reconcile the GDPR's individual rights-based approach with the proliferation of generated personal data that is non-individualistic in nature is to recognize the concept of generated personal data, and to articulate rules with its characteristics in mind.

## **B. CHALLENGES TO THE PRINCIPLE OF TRANSPARENCY**

The generation of personal data poses a challenge to the principle of transparency. The GDPR imposes on data controllers a duty to provide individuals with information about what is happening to their personal data (the principle of transparency), but this safeguard is challenged by generated personal data.

Article 13 and 14 of the GDPR regulate what information must be provided to data subjects by controllers.<sup>184</sup> Article 13 regulates when personal data is collected directly from data subjects (e.g., when they subscribe to a new service).<sup>185</sup>

Article 14 regulates the same obligation to provide information, but in cases where personal data has not been obtained directly from the data subject.<sup>186</sup> The text of the law does not limit the situation to where the data is collected from third-party sources.

Thus, in a case where personal data is generated by the data controller, Article 14 governs. Article 14 has more exceptions than these under Article 13,<sup>187</sup> and whether providing information to the data subject “proves impossible or would involve a disproportionate effort” depends on specific facts.

In theory, the duty generally applies (or should apply) whenever new personal data is generated by companies or data brokers. Nevertheless, in practice, individuals whose personal data have been generated rarely receive notification under Article 14.

Rainer Mùhlhoff aptly observes that “[a] person’s predictive privacy is violated when personal information about them is predicted without their knowledge and against their will based on the data of many other people.”<sup>188</sup>

---

<sup>183</sup> For more remarks by by Dr. Dara Hallinan, see *Are Synthetic Health Data “Personal Data”?*, *supra* note \_\_, at 41.

<sup>184</sup> Article 29 Working Party, *Guidelines on Transparency under Regulation 2016/269*, (2018), <https://ec.europa.eu/newsroom/article29/items/622227> [hereinafter *Transparency Guidelines*].

<sup>185</sup> See generally *GDPR Commentary*, *supra* note \_\_, at 413.

<sup>186</sup> Article 14(1), GDPR.

<sup>187</sup> *GDPR Commentary*, *supra* note \_\_, at 437.

<sup>188</sup> Rainer Mùhlhoff, *Predictive Privacy: Collective Data Protection in the Context of Artificial*

Considering that the *Transparency Guidelines* do not envision a case where personal data is generated within the same organization,<sup>189</sup> and also the case involved the Binding decision 1/2021 regarding WhatsApp Ireland<sup>190</sup>, the transparency principle and obligation to provide information to data subject under Article 14 is further challenged by the proliferation of generated personal data.

Not only does this impact the transparency principle, but it also undermines the individual rights of the GDPR, as individuals are not aware that their personal data are generated, and therefore not able to exercise their rights. Individuals are further required to exercise their right to access actively to learn what types of data are generated about them<sup>191</sup>.

### C. CHALLENGE TO THE RIGHT TO RECTIFICATION

Generated personal data pose a difficult challenge to the principle of accuracy and the right to rectification, both of which are closely related and linked to each other. Individuals will face challenges if they wish to contest and rectify generated personal data.<sup>192</sup> This is especially true when generated personal data is about a probable, or possible, but uncertain future -- predictions.<sup>193</sup>

The accuracy principle is one of the cornerstones of modern data protection and privacy law. The GDPR imposes on data controllers the duty to keep personal data as accurate as possible.<sup>194</sup> However, what or how accurate means in generated personal data is not an easy question to answer.<sup>195</sup>

---

*Intelligence and Big Data*, 10 Big Data & Society (2023).

<sup>189</sup> *Transparency Guidelines*, *supra* note \_\_, at para 26 (“Article 14 applies in the scenario where the data have not been obtained from the data subject. This includes personal data which a data controller has obtained from sources such as: third party data controllers; publicly available sources; data brokers; or other data subjects.”).

<sup>190</sup> EDPB, *Binding decision 1/2021 on the dispute arisen on the draft decision of the Irish Supervisory Authority regarding WhatsApp Ireland under Article 65(1)(a) GDPR*, [https://www.edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-12021-dispute-arisen\\_en](https://www.edpb.europa.eu/our-work-tools/our-documents/binding-decision-board-art-65/binding-decision-12021-dispute-arisen_en).

<sup>191</sup> Jef Ausloos, *The right to erasure in EU data protection law: From individual rights to effective protection* (Oxford University Press, 2020); René Mahieu, *The Right of Access to Personal Data in the EU: A Legal and Empirical Analysis* (Dissertation Vrije Universiteit Brussel, 2023).

<sup>192</sup> Hideyuki MATSUMI, *Data Protection, Privacy, and Unfalsifiable Predictions* in *Handbook on Law and Digital Technologies* (Ugo Pagallo, Roger Bronsword, Pompeu Casanovas eds., De Gruyter, forthcoming 2025) [hereinafter *Data Protection, Privacy, and Unfalsifiable Predictions*]; Hideyuki MATSUMI, *The Failure of Rectification Rights*, presented at PLSC 2022 in Boston, Massachusetts (Unpublished manuscript; On file with author) (discussing how the right to rectification under the current configuration of the GDPR does not empower individuals to rectify potentially inaccurate predictions, but actually burdens them).

<sup>193</sup> Matsumi & Solove, *The Prediction Society*, *supra* note \_\_.

<sup>194</sup> Article 5(1), GDPR (“(1) Personal data shall be: (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (‘accuracy’)”).

<sup>195</sup> See *Data Protection, Privacy, and Unfalsifiable Predictions*, *supra* note \_\_ (Taxonomy by

A review of various literature suggests at least two different views or approaches to accuracy assessment: (1) Accuracy assessment focusing on the substance of an assertion made in generated personal data (“substantive accuracy”); and (2) Accuracy assessment focusing on the procedure of making or generating personal data (“procedural accuracy”).

Substantive accuracy assesses accuracy by focusing on whether the substance of the assertion is true and accurate.<sup>196</sup> For example, when a system infers or generates a person's age, height, weight, or address, these generated personal data are compared against the ground truth (i.e., facts). This generated personal data is considered substantively accurate if it is the same as the facts.

Procedural accuracy assesses accuracy by examining the process or procedure of making an inference.<sup>197</sup> Consider an inference about the likelihood of developing a heart disease. In the *Profiling Guidelines*, the Art. 29 WP takes such a view. In explaining how the right to rectification applies to profiling, the *Guidelines* provide a hypothetical example in which a patient is categorized into a high-risk group for heart disease:

A local surgery's computer system places an individual into a group that is most likely to get heart disease. **This ‘profile’ is not necessarily inaccurate even if he or she never suffers from heart disease.** The profile merely states that he or she is *more likely* to get it. That may be **factually correct as a matter of statistics.**<sup>198</sup>

Each of these views to assess accuracy initially appears to be reasonable. However, the discrepancy becomes evident when the view of procedural accuracy is applied to different types of personal data and these are compared with each other: (1) a person's age, height, weight, or address, irrespective of whether they are collected or profiled; and (2) generated personal data about the future, or predictions.

If data about a person's age, height, weight, or address, irrespective of collected or generated, is substantively inaccurate, then they are inaccurate data, deserves to be rectified. We do not resort to procedural accuracy to argue that “it is not necessarily inaccurate because it was procedurally accurate.”

This view, however, changes when generated personal data about the future, or predictions, are at issue. If generated personal data concern a probable, or possible, but uncertain future, such as the risk of developing a heart disease, the *Profiling Guidelines* suggest “it is not necessarily inaccurate even if he or she never suffers from heart disease” because it can be “factually correct as a matter

---

Perspectives on Accuracy Assessment).

<sup>196</sup> Hence, referred to as “substantive accuracy.”

<sup>197</sup> Hence, referred to as “procedural accuracy.”

<sup>198</sup> *Profiling Guidelines*, *supra* note, at 18 (*Italic emphasis* in the original; **bold emphasis** added).

of statistics.”<sup>199</sup> By the same virtue, various generated scores, including COMPAS risk score and DEWS scores, would be deemed *not necessarily inaccurate*, if we follow the same view.<sup>200</sup>

As the accuracy principle is challenged, the right to rectification is also challenged. How to view or assess accuracy is of consequence because the right to rectification applies to *inaccurate* personal data. Individuals will face significant challenges when they want to falsify or rectify potentially inaccurate generated personal data.

## **IV. RECOGNIZING GENERATED PERSONAL DATA UNDER THE GDPR**

The challenges articulated in the previous Part are difficult problems to solve. Recognizing the concept of generated personal data, however, can be a starting point. Recognizing the concept of generated personal data under the GDPR has some virtues.

### **A. RECOGNIZING GENERATED PERSONAL DATA: ADVANCING THE LEGAL FRAMEWORK**

On 30 November 2021, a MEP asked at the European Parliament whether the Commission intends “to extend the list of special categories of personal data to include facial and voice data” in the next revision of the Data Protection Regulation,<sup>201</sup> after observing that “[d]eepfake technologies make use of personal data, harnessing the facial and voice characteristics used for personal identification.”<sup>202</sup> The Commission replied negatively to this question.<sup>203</sup> It is worthy to note that the MEP began its question by referring to deepfake

<sup>199</sup> *Profiling Guidelines*, *supra* note, at 18.

<sup>200</sup> See *Examples of Generated Personal Data*, *supra*.

<sup>201</sup> Ivo Hristov, *Risks in respect of the use of biometrics*, Question for written answer to the Commission, Parliamentary Question, available at [https://www.europarl.europa.eu/doceo/document/E-9-2021-005185\\_EN.html](https://www.europarl.europa.eu/doceo/document/E-9-2021-005185_EN.html).

<sup>202</sup> *Id.*, (“The growth in highly-realistic video and audio manipulations carries with it several risks, such as that of identity theft using biometric data that could be obtained during the verification process for online banking transactions. Deepfake technologies make use of personal data, harnessing the facial and voice characteristics used for personal identification. In view of the potentially adverse economic and social consequences of this, I would like to ask the following questions: 1. Does the Commission intend, in the next revision of the Data Protection Regulation, to extend the list of special categories of personal data to include facial and voice data?; 2. Does it plan to clarify still further the exemptions under which the use of facial and voice data is permitted?; 3. . .”).

<sup>203</sup> The European Commission, *Answer to the question of Risks in respect of the use of biometrics*, Parliamentary Question, available at [https://www.europarl.europa.eu/doceo/document/E-9-2021-005185-ASW\\_EN.html](https://www.europarl.europa.eu/doceo/document/E-9-2021-005185-ASW_EN.html) (“1. Where facial images and voice data are used for purposes of identifying a data subject, they are already covered under the definition of biometric data and are considered as a special category of data under Article 9 of the GDPR. When they are not, they are still personal data and fall under data protection rules.”).

technologies, implying generated facial and voice data, but the question does not specifically mention generated facial and voice data; rather the question refers to only “facial and voice data.” The answer by the Commission refers to “facial images and voice data” of existing individuals.

This section will very briefly explore one of the virtues of recognizing generated personal data under the GDPR, in favor of the argument for adopting such data in the next revision of the GDPR.

### 1. Articulating Rules on How Personal Data Can Be Generated

Under the current configuration of the GDPR, there is no regulation on how personal data can be generated. To highlight the point, consider the following, rather extreme, example. Suppose there is a company or data broker that makes and sells predictions<sup>204</sup> about individuals -- generated personal data -- by reading their hand palms. This act is unlikely covered by the profiling regulation, as it lacks the automated element. But what if the palmistry was carried out using automated means, such as image recognition and machine learning?

An example of palmistry might seem amusing. However, consider the history of phrenology and how similar ideas continue to resurface.<sup>205</sup> In an attempt to predict criminality based on physical characteristics, Cesare Lombroso, once referred to as “the father of criminology,” argued that it was possible to identify criminals by looking at facial features.<sup>206</sup> Police used phrenology to “typify criminals and arrest them, even in the absence of any evidence a crime had been committed.”<sup>207</sup> By the 1950s, the phrenology craze had burned out.<sup>208</sup>

The history of phrenology, however, doesn’t end there. Even if phrenology has long been recognized as a pseudo-science, similar claims are resurfacing. For example, a press release by Harrisburg University claimed that “[a] group of Harrisburg University professors and a Ph.D. student have developed automated computer facial recognition software capable of predicting whether someone is likely going to be a criminal.”<sup>209</sup> This example demonstrates that questionable attempts to make predictions persist and recur.

---

<sup>204</sup> E.g., performance at work, economic situation, health, personal preferences, interests, reliability, behavior, etc.

<sup>205</sup> *The Prediction Society*, *supra* note \_\_, at 54.

<sup>206</sup> Cesare Lombroso, *Criminal Man, According to the Classification of Cesare Lombroso* (1911) [hereinafter *Criminal Man*]. The original Italian book, entitled *L’Uomo delinquente*, was published in 1876, and the first English translation was published in 1911. See Hideyuki MATSUMI, *Predictions and Data Protection* (unpublished manuscript; on file with author), at 6.

<sup>207</sup> Ifeoma Ajunwa, *The Quantified Worker: Law and Technology in the Modern Workplace* (2023), at 145.

<sup>208</sup> *Id.*

<sup>209</sup> Harrisburg University, *HU Facial Recognition Software Predicts Criminality*. This press release is removed from Harrisburg University’s website, but can be accessed at <http://archive.is/N1HVe> (last visited 2023-04-28).

The point is, there will always be some dubious AI hypes claiming that “our AI can analyze, recognize, detect, or predict XYZ.”<sup>210</sup> The law needs to hold these companies and data brokers accountable for their data practices, including generating personal data and making predictions. And because “AI” inevitably uses personal data, data protection and privacy laws are one of the most useful and powerful regulatory tools to tackle dubious generation of personal data.

Admittedly, there are provisions apart from the profiling under the GDPR that would hinder and can challenge dubious data practices, such as palmistry and phrenology.<sup>211</sup> Yet, the current configuration of the GDPR does not stipulate on the conditions how personal data can be generated. The GDPR does have rules on what rights individuals have when they are profiled, what data controllers must do when profiling individuals, and when a particular profiling is prohibited unless exceptions are met. However, many data protection and privacy laws, including the GDPR, does not stipulate what are acceptable means to carry out profiling and what are not.

One very rare exception to this is the Federal Data Protection Act of Germany.<sup>212</sup> It stipulates the minimum standard for generating probability values, or credit scores. Under the Act, “the use of a probability value regarding specific future behaviour” of an individual “for the purpose of [scoring]” is prohibited unless it meets certain conditions, including that the calculation must be based on a “*scientifically recognised mathematical statistical method.*”<sup>213</sup>

Whether this meets the sufficient for protecting individuals from companies and data brokers' business of generating personal data and selling them is another question. But perhaps the very fact that *Schufa* case was brought before the referring court and the Court may suggest that such a rule brings the law a step closer to regulating the scoring business.

With the concept of generated personal data, the GDPR to begin articulating rules on the acceptable means and minimum standards for generating personal data in the EU.

## **B. RECOGNIZING GENERATED PERSONAL DATA: NECESSARY BUT NOT SUFFICIENT**

---

<sup>210</sup> See generally Arvind Narayanan & Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (2024) [hereinafter *AI Snake Oil*].

<sup>211</sup> E.g., DPAI in Article 35, various principles in Chapter II, rights of the data subjects in Chapter III.

<sup>212</sup> Bundesdatenschutzgesetz (“BDSG”) in German. The English translation is available at [https://www.gesetze-im-internet.de/englisch\\_bdsdg/](https://www.gesetze-im-internet.de/englisch_bdsdg/). See also *Schufa*, *supra* note \_\_.

<sup>213</sup> BDSG art 31(1)2 (“data used to calculate the probability value are demonstrably essential for calculating the probability of the action on the basis of a scientifically recognized mathematicstatistical procedure”) (emphasis added).

Recognizing the concept of generated personal data is only a half measure.<sup>214</sup> It does not magically solve the challenges discussed in Part III.<sup>215</sup>

However, by distinguishing the concept of generated personal data from other forms of personal data, such as collected ones, this recognition can serve as a starting point and catalyst for discussing issues and possible solutions specific to such data.

Several important debates can emerge from this distinction. For instance:

1. How the accuracy principle applies differently to generated personal data compared to collected personal data.
2. How the rights of data subjects and duties of data controllers apply differently to generated personal data.
3. Under what circumstances and for what purposes generating personal data, such as photos, videos, or audio about a deceased person, is acceptable.

These debates can focus on specific aspects of generated personal data.

The key point is that the concept of generated personal data enables debates that are distinct from, or additional to, those surrounding the general concept of personal data. This differentiation allows for more nuanced and targeted discussions about the unique challenges and ethical considerations posed by generated personal data in the evolving landscape of data protection and privacy.

## CONCLUSION

This Article explored and analyzed how the General Data Protection Regulation (GDPR) applies to, or is challenged by, data generated by data controllers, or “generated personal data.”

The concept of generated personal data is nothing new. Similar concepts have been referred to and discussed by many commentators, and the long discourse has indicated that the current data protection and privacy regime struggles with such data.

Today, the advent and recent advancement of generative “artificial intelligence” -- GenAI -- only accelerates and accentuates the problems and challenges posed by generated personal data. Data companies and brokers are generating a vast amount of personal data, probably more than they collect.

Because the way and what type of personal data are generated by those data

---

<sup>214</sup> Idea inspired by the *Prepared Testimony and Statement for The Record of Woodrow Hartzog, Neil Richards, and Ryan Durrie*, [https://www.judiciary.senate.gov/imo/media/doc/2023-09-12\\_pm\\_-\\_testimony\\_-\\_hartzog.pdf](https://www.judiciary.senate.gov/imo/media/doc/2023-09-12_pm_-_testimony_-_hartzog.pdf).

<sup>215</sup> See *The Challenges Raised by Generated Personal Data*.

controllers can impact the lives of individuals, it is critical that those data practices are held accountable by relevant law. Data protection and privacy law are some of the most useful and powerful tools for this purpose.

By recognizing the concept of generated personal data, data protection/privacy law can start discussing the challenges and issues surrounding such data, and articulate how the law can address these.

The author hopes that this Article will serve as a foundation for these discussions and contribute meaningfully to the ongoing debate.

[END OF DOCUMENT]