JUS CIVILE

ANDREA FEDI

Avvocato

BIG DATA AND AI:
AN ISSUE FOR GENERAL LAW AND DATA PROTECTION

SUMMARY: *1. Introduction. – 2. A case study: Big Data and anti bribery. – 3. International principles and recent statements. – 4. Personal vs. non-personal data. – 5. Transparency and purpose limitation – 6. Profiling. – 7. Automated decisions. – 8. In a nutshell and an interpretative offer.*

1. – The term "Big Data" has long since appeared in the jargon of practitioners interested in legal ramifications of new technologies. By this term it is common to address extremely large data sets that may be analysed computationally to extract inferences about data patterns, trends and correlations [1]. In other terms, the term refers to a phenomenon characterised by:

i. a magnitude requirement (large data sets),

ii. a methodology (computational analysis, i.e., through a machine),

iii. a final goal (the extraction of knowledge from the data sets, as a miner extract a mineral from a mine).

The current technology has indeed opened doors to the possibility to collect and process huge amounts of data, and extract from their analysis new and predictive knowledge with great 'velocity', from large 'volume' databases containing a 'variety' of different

[1] Big Data makes it possible to collect and analyse large amount of data to identify attitude patterns and predict behaviours of individuals, groups and communities – it is necessary to secure the protection based on a person's right to control his/her personal data and the processing of such data. Big Data analytics identify computational technologies that analyse large amounts of data to uncover hidden patterns, trends and correlations: the whole data management lifecycle of collecting, organizing and analysing data to discover patterns, infer situations or states, predict and understand behaviours.

See also ICO, *Big data, artificial intelligence, machine, learning and data protection*, 2017.

data (so large and various that a natural person could not reasonably make such analysis), controlling their 'veracity' and, eventually, creating 'value'.

Indeed, the computational insight of data fields (also through artificial intelligence, AI) consents to find (often unexpected) correlations among data and draw consequences (forecasts) almost in real time, which of course delivers a high value capacity to predict (and influence) reactions of individuals and communities (elections, consumer preferences, investment appetites, etc.). Big Data have been therefore indicated as a new paradigm for the collection, storage, management, analysis and visualisation of large data collections with heterogeneous characters. That paradigm is based on the notorious 5 Vs (volume, velocity, variety, veracity and value) [2].

Obviously possibility to extract value from extensive data sets fundamentally requires the ability to gather together significant volumes of data and material technological investments to analyse them through new fast technologies [3]. It is no surprise, then, that big players in this game are on one hand governments (public authorities and agencies) [4] and large corporations [5].

Now, it is just honest to say that we should not indulge in (only) dystopic and alarming descriptions of data analytics. That analysis can be used for good reasons and give great results in fighting crimes, frauds and tax evasion, or organising reactions to emergencies (earthquakes, floods, etc.) or managing smart cities which more efficiently provide public services to citizens (traffic, garbage, mail, etc.) [6]. Big Data can also be used to forecast effects of public policy decisions and ponder pros and cons, for proactive and predictive healthcare, to peruse extensive sets of documents in trials or investigations.

---

[2] GIRINO, ESTRANGENORS, *Alla ricerca dei dati perduti*, in *Privacy &*, 2019, p. 74

[3] Anyone of us should be aware of how many digital traces we leave each time we surf on the internet, use a fidelity card or exploit geo location systems. In addition, all our technological devices share our data (Internet of Thighs) and communicate them to the service providers that supply TLC, domotics, maps and other services that we cannot but use in the world of today. Those digital traces are processed through algorithms and AI and provide a clear vision of who we are as individuals and of the communities to which we belong. From structured or unstructured data (input), through data mining and cleansing, aggregation and integration, analysing and modelling, Big Data players obtain an output that gives a precise insight on natural persons and groups.

[4] It is sufficient to consider the incredible amount of information that the governmental authorities process with respect to citizens, tax payers, immigrants, business undertakings, etc.

[5] Let's think to consumers'profiles, data on visitors of websites and social networks, etc. MAURO, *I big data tra protezione dei dati personali e diritto della concorrenza*, in *Circolazione e protezione dei dati personali, tra libertà e regole di mercato*, Milano, 2019, p. 643.

[6] IASELLI, *Lezioni di informatica giuridica*, ed. Key, Milano, 2019, p. 53 ff.

On the other hand, we should all be adverted that the algorithms used to inquire Big Data sets are written by human beings and (albeit the algorithms may thereafter develop and grow and transform themselves through machine learning and AI) they will inevitably be affected by prejudices or, simply, mistakes made by those persons [7]. If the tool (algorithm) is wrong it will originate wrong knowledge. Imagine an algorithm based on a scientific or statistical theory that proves wrong after many years, or an algorithm that simply mirrors a 'state of play' that changes over time. Just a century ago, the writer of an algorithm might assume that women were less productive than men or monopolies are more efficient than competition.

A wrong setting of the algorithm may do even worse. It may in fact orient forecasts and decisions in a way that perpetuates the mistake, say – denying loans to certain classes of people and therefore blocking the possibility itself that the individuals of that cluster may demonstrate the wrong assumption on which the algorithm has been conceived.

Also, Big Data are made to categorise knowledge but that categorisation may lead to 'imprisonment' of individuals in clusters and to treat each person in the cluster independently from his/her single, original and unique nature. In this way, everyone becomes prisoner of his/her cluster and his/her past. The facts of one's life are the fuel of the Big Data engine, which will not consider the possibility of change.

Again, there is wide literature and awareness of risks entailed by Big Data and only sectoral knowledge of how much they could improve our lives. However, we must combat one-sided perceptions. Let me just try to give an example.

2. – It is now two decades that many jurisdictions have adopted anti-bribery legislations and have called companies and legal entities to cooperate in the fight by creating a special incentive to prevent business crimes. In fact, the UK Anti Bribery Act, the Spanish Ley Organica, the French Loi Sapin II and the Italian law 231 all share a common trait, pursuant to which the company is responsible for bribes committed in the interest of the company by one of its representatives/employees unless the company has adopted a compliance system to prevent, detect and fight commission of bribes.

---

[7] *Algorithms, far from being neutral, reflect embedded choices about data, connections, inferences, interpretations and thresholds for inclusion which are "political" in nature and should therefore be subject to legal scrutiny* (ODDENINO, in *Dir. Comm. Int.*, 2017, p. 777).

Thus, company willing to go exempt from sanctions have a need for conducting a risk mapping, a risk assessment and a gap analysis through a specific due diligence exercise, in order to then marshal effective and regular audits and procedures for preventive controls and continuous monitoring.

Traditionally, at least in Italy, the above need has been transposed into long and heavy compilations of internal and detailed internal rules and procedures (on top of existing general laws), which hardly match with the rapidity and flexibility of business and have rarely consented to the company the exculpation of companies.

More, the need to select employees, agents and suppliers and check their "white" status (i.e., no past criminal records), in order to demonstrate a diligent investigation aimed at preventing future crimes, inevitably clashes with those employees', agents' and suppliers'privacy rights and with the General Data Protection Regulation 679/2016 (GDPR), whose Art. 10 consents only a very limited scope for those background checks [8].

However, a point is worth underlining.

Italian law does not require companies to write long manuals of internal rules, nor to investigate one's private life prior to hiring him or her.

The law indeed only requires a 'model of organisation and business' and a 'model' (a way of doing safe business) is not a 'manual' of prescriptions. What is pivotal is that the company must get organised and ready to prevent business crimes; it does not mean that the company should write or rephrase the law or replace a public prosecutor.

If one concurs with the above, then Big Data and AI could be a model of controlling a business without writing extensive rules and impinging in personal criminal data of employees. For example, Big Data may consent preventive knowledge of where and in which occasions business crimes are more likely to happen, where to direct audits, which facts or documents needs to be checked in depth, how to make sample controls, etc.

The above is even more evident considering the reporting duties laid down by the EU Directive 2014/95/EU and the metric exercise that it mandates for large corporations [9].

---

[8] Processing of personal data relating to criminal convictions and offences or related security measures shall be carried out only under the control of official authority or when the processing is authorised by EU or EU Member State law providing for appropriate safeguards for the rights and freedoms of data subjects.

[9] "*In order to enhance the consistency and comparability of non-financial information disclosed throughout the Union, certain large undertakings should prepare a non-financial statement containing information relating to at least environmental matters, social and employee-related matters, respect for human rights, anti-corruption and bribery matters. Such statement should include a description of the policies, outcomes and risks related to those matters and should be included in the management report of the undertaking con-*

Again, Big Data may furnish the metric to assess results of preventive law initiatives, consequences on environment and impacts on stakeholders, special needs to protect communities and human rights and methodologies to assess consequences of entrepreneurial decisions.

Use of data analytics for compliance matters (including KYC) will consent big achievements.

3. – That being said, it is evidently important to avoid abuses of Big Data and here we must remind the following pillar statements.

Art. 12 of the Universal Declaration of Human Rights: no one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, or to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

Art. 8.1 of the EU Convention on Human Rights: everyone has the right to respect for his private and family life, his home and his correspondence. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Art. 8.1 of the EU Charter of Fundamental Rights: everyone has the right to the protection of personal data concerning him or her. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. Compliance with these rules shall be subject to control by an independent authority.

All in all, Big Data impinge in our private life and extract knowledge on individuals often with the purpose to influence their choices (as consumers, citizens, investors). That may clash with some of the very foundation principles of EU [10] and, quite obviously, has

---

cerned. The non-financial statement should also include information on the due diligence processes implemented by the undertaking, also regarding, where relevant and proportionate, its supply and subcontracting chains, in order to identify, prevent and mitigate existing and potential adverse impacts".

[10] I cannot but remind the words of the former Chairman of the Italian Data Protection Authority, Antonello Soro (*Big data e libertà nella dimensione digitale*, August 23, 2018): "*The new emergency for global*

generated concern, which has been repeatedly expressed by EU bodies.

The COUNCIL OF EUROPE on 23 January 2017 issued its *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data* and insisted on the following cornerstone yardsticks:

a) ethical and socially aware use of data,

b) need of preventive policies and risk assessment,

c) respect of principles of purpose limitation and transparency,

d) privacy by-design approach,

e) pivotal role of individual's consent,

f) anonymization of data whenever possible,

g) centrality of human intervention (as opposed to AI) in Big Data supported decisions.

Consistently, the ADVISORY COMMITTEE OF THE CONVENTION FOR THE PROTECTION OF INDIVIDUALS WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA (Convention

---

privacy is called app economy" – The term "app economy" is one of the most appropriate definitions of the digital economy, where a significant part of trade is channelled through applications downloaded by consumers, for various purposes, on smartphones, tablets, etc.. In most cases, apps provide services free of charge or, more precisely, in exchange for non-monetary consideration: data. They are used to build consumer profiles useful for effectively directing marketing activities and therefore represent both an important asset from an economic point of view and the object of a fundamental right. Example: applications that allow obtaining information on the header of the numbers typed, drawing on a database compiled by the users themselves.

Consent: even assuming that it is actually acquired, one should ask if it is really informed and therefore if the user has actual awareness of the use that will be made of the data provided. In the absence of adequate information, the consent itself could not be considered validly given.

Transfer of contacts from the address book: the user cannot validly use these contacts – transferring them to a supplier who will use them in a commercial context – without the consent of the third party. This highlights the limits and risks of a commercial system in which the general conditions of contract, unilaterally established by digital companies (or by big techs), end up defining the perimeter of rights and freedoms.

Geopolitical impact and impact on the structure and transparency of the market: such commercial systems, with companies located mainly outside the European Union and in real data havens, are often much more elusive and "underground" than tax systems. And, perhaps, even more dangerous, if we consider that in the digital dimension.

Objective difficulty of enforcing sanctioning measures against parties towards which, in the event of non-compliance, it is rather problematic not only to resort to the criminal sanctions provided for in the event of non-compliance, but also, more simply, to effectively carry out inspection activities.

"In order to overcome these aporias, it is essential that the right to the protection of personal data and the necessary guarantees to ensure its effectiveness be recognised at international level. In a global context such as that of the network, based on interdependence and on the overcoming of borders and frontiers, the protection of rights can only be global and based on homogeneous levels of guarantee".

108) issued on January 25, 2019 its *Guidelines on data protection and artificial intelligence* and

i) offered a series of recommendations to developers of AI solutions and public decision-makers inviting them to consider in advance ("by design" [11]) any potential adverse effects of AI on rights of individuals, through impact assessment procedures that involve also independent committees of experts and academic institutions;

ii) asserted the need to design algorithms that do not take decisions conditioned by prejudices or on an unnecessary amount of data (Big Data) and, with a view to privacy and data protection, suggested the use of so-called 'synthetic data' in the training phase of the algorithm, *i.e.* a particular category of data anonymized by computer processing;

iii) recommended to maintain a 'human rights by design' approach, *i.e.* continuous vigilance over the algorithm and the guarantee for the right of data subjects not to be subject to decisions based on fully automated processes;

iv) called on governments and public decision-makers to ask bidders in public tenders special forms of transparency on algorithms and guarantees that preventive impact assessments on personal data have been carried out;

v) placed particular emphasis on the need to set up training programs and "digital literature", in order to spread awareness among the population on the potential and limits of AI.

Few months ago, the COMMITTEE OF EU MINISTERS of 13 February 2019, with its *Declaration on the manipulative capabilities of algorithmic processes* recognized that "*digital services are used today as an essential tool of modern communication […]. This results in unprecedented amounts of new data that are constantly created with mounting speed and scale. […] Technology is an ever growing presence in our daily lives and prompts users to disclose their relevant, including personal, data voluntarily and for comparatively small awards of personal convenience. These data are used to train machine-learning technologies to prioritise search results, to predict and shape personal preferences, to alter information flows, and, sometimes, to subject individuals to behavioural experimentation. […] Increasingly, computational means make it possible to infer intimate and detailed information about individuals from readily available data. This supports the sorting of individuals into categories, thereby reinforcing different forms of*

[11] EUROPEAN DATA PROTECTION BOARD, *Guidelines 4/2019 of November 13, 2019 on Article 25, data protection by design and by default.*

social, cultural, religious, legal and economic segregation and discrimination. It also facilitates the micro-targeting of individuals based on profiles in ways that may profoundly affect their lives. Moreover, data-driven technologies and systems are designed to continuously achieve optimum solutions within the given parameters specified by their developers. When operating at scale, such optimisation processes inevitably prioritise certain values over others, thereby shaping the contexts and environments in which individuals, users and non-users alike, process information and make their decisions. This reconfiguration of environments may be beneficial for some individuals and groups while detrimental to others […]. Contemporary machine learning tools have the growing capacity not only to predict choices but also to influence emotions and thoughts and alter an anticipated course of action, sometimes subliminally." Therefore "*the Committee of Ministers:*

i) draws attention to the growing threat to the right of human beings to form opinions and take decisions independently of automated systems, which emanates from advanced digital technologies. Attention should be paid particularly to their capacity to use personal and non-personal data to sort and micro-target people, to identify individual vulnerabilities and exploit accurate predictive knowledge, and to reconfigure social environments in order to meet specific goals and vested interests;

ii) *encourages member States to assume their responsibility to address this threat by […] considering the need for additional protective frameworks related to data that go beyond current notions of personal data protection and privacy […]; taking appropriate and proportionate measures to ensure that effective legal guarantees are in place against such forms of illegitimate interference; and empowering users by promoting critical digital literacy skills and robustly enhancing public awareness […];*

iii) *draws attention to the necessity of critically assessing the need for stronger regulatory or other measures to ensure adequate and democratically legitimated oversight over the design, development, deployment and use of algorithmic tools, with a view to ensuring that there is effective protection against unfair practices or abuse of position of market power."*

There are two remarks that seem worth emphasizing.

First, will all the above statements generate new EU rules on Big Data in the short term? We are under the impression that, albeit after a long and certainly articulated and non-easy process, we should expect that.

Second, even before the enactment of new rules, it would be a mistake to consider the

above statements as mere political orientations. Now that the data protection regulation is framed within a common EU regulation directly applicable in the whole EU, the above statements can also be taken as interpretative guidance on the interpretation and construction of individuals'rights in the Union.

The above can play an important role, all the more because the GDPR seems to fall short with respect to the need to regulate Big Data and AI.

4. – The GDPR contains itself a very strong statements: "*The processing of personal data should be designed to serve mankind. […] GDPR respects all fundamental rights and observes the freedoms and principles recognised in the Charter as enshrined in the Treaties, in particular the respect for private and family life, home and communications, the protection of personal data, freedom of thought, conscience and religion, freedom of expression and information, freedom to conduct a business, the right to an effective remedy and to a fair trial, and cultural, religious and linguistic diversity*".

That being said, the architecture of GDPR is however based on protection of 'personal data', i.e., data of identifiable individuals [12] and that limited scope leaves much out of the scope of GDPR's protections.

In fact, whilst personal data are subject to several controls and restrictions pursuant to the GDPR, EU flows of non-personal data are basically free [13], which means that the GDPR does not provide protections against abuses of Big Data each time they are (or appear) to process non personal data, e.g., data referred to communities and groups (as opposed to individuals) or anonymized data (that cannot be associated to identifiable natural persons) [14].

---

[12] Art. 4(1) GDPR: "*personal data means any information relating to an identified or identifiable natural person (the data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

[13] See EU Non Personal Data Regulation 1807/2018. GIANNELLA, VENDITTI, *Il Regolamento (UE) 2018/1807 e il confine con il GDPR*, in *Privacy &*, 2019, 17.

[14] Obviously, anonymization is different from mere pseudonymisation, which means only "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*". Pseudonymised data remain personal data and are subject to GDPR rules. "Whereas 26 GDPR: *Personal data which have undergone pseudonymisation,*

With respect to information on communities (neighbourhoods, work teams, churches, etc.) the risk of clustering them and taking automated decisions through AI thus remains on the table, because the GDPR is not designed to provide safeguards to communities but only to furnish a shield for individuals.

With respect to anonymized data, instead, all the legal framework of the GDPR bases on the possibility to draw a clear line distinguishing between personal data and anonym data [15], but that line appears today tiny, pale and controversial [16].

Here lays the first problem: data may appear anonymized but the great power entrenched in Big Data analysis may decrypt them and re-associate those data to identifiable individuals [17]. The EUROPEAN DATA PROTECTION SUPERVISOR indeed recognized that in its Opinion 7/2015 of November 19, 2015: "*One result is the emergence of a revenue model for Internet companies relying on tracking online activity. Such Big Data should be considered personal even where anonymization techniques have been applied: it is becoming and will be ever easier to infer a person's identity by combining allegedly 'anonymous' data with publicly available information such as on social media. Furthermore, with the advent of the 'Internet of Things', much of the data collected and communicated by the increasing number of personal and other devices and sensors will be personal data: the data collected by them can be easily related to the users of these devices whose behaviour they*

---

which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. GDPR does not therefore concern the processing of such anonymous information, including for statistical or research purposes."

[15] Whereas 26 of the GDPR explicitly states that "*principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or data rendered anonymous in such a manner that the data subject is no longer identifiable.*"

[16] Very much the same happens with respect to common data, sensitive data and judicial data; in a Big Data scenario the analytics may infer sensitive data from common data and the distinction is puts at stake.

[17] According to Whereas 26 "*To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments*".

will monitor. These may include highly sensitive data including health information and information relating to our thinking patterns and psychological make-up".

This is what is otherwise called in literature as 'singling-out': data which are apparently anonymized can be used to retrieve 'personal' data using analytics.

If we start considering that very few data are thus actually anonym, and the majority of those only pretend to be anonym (but in reality they are not), then the GDPR applies and a big clash emerges, principally because rules on transparency, profiling and automated decisions of the GDPR must apply.

5. – Among the foundations of the GDPR we easily trace the principles of transparency (data subjects must be aware of the processing of their data, by whom, for which purposes, how) and purpose limitation (a controller may not process data for purposes other than those disclosed).

The above triggers that data subjects must receive a full notice on the processing of their data (as dictated by Articles 13/14 GDPR) and, as stated in Art. 6.1(b) GDPR, personal data may be collected only for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes [18].

That poses at least three material issues in respect of Big Data and data analytics through AI [19].

The first issue is that Big Data are not always collected directly from the controller that pulls them in the data analytics engine. That may happen, of course; for instance when a social network collects digital traces of its members and then analyse them [20].

---

[18] And, where the processing for a purpose other than that for which the personal data have been collected is not based on the data subject's consent or on a EU or EU Member State law which constitutes a necessary and proportionate measure in a democratic society, the controller shall, in order to ascertain whether processing for another purpose is compatible with the purpose for which the personal data are initially collected, take into account, inter alia: (a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing; (b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller; (c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10; (d) the possible consequences of the intended further processing for data subjects; (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.

[19] MANTELERO, *La privacy all'epoca dei big data*, in *I dati personali nel diritto europeo*, Torno, 2019, p. 1181.

[20] Whereas 30 GDPR: *Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other*

Here, the Big Data collector and analyser is the direct point of contact of the social network members. But in certain instances the Big Data controller first collects data from various sources (social networks, public data, sellers of data bases, large retail businesses), then puts them together and finally makes the analysis without having had any direct relationship with the data subjects. In this second scenario, the data subject should be informed from the beginning (say – from the time he/she subscribes a fidelity card in a shop or a membership in a social network) that his/her personal data can go through the whole chain from the shop to the distributor to the manufacturer to the Big Data player and, here is the point, the data subject should be informed since that very moment of the purpose for which his/her data will be processed by each player along the chain.

The second issue is the legal relationship between controllers down the chain. Are they autonomous controllers, or co-controllers or one is a processor and the other the real controller? Clearly, the liability regime (vis-à-vis data subjects and authorities) may change.

The third related issue is that the data controller should be able to inform the customer of all uses of his/her data but nobody can precisely define that at that stage, also because the inferred knowledge that data analytics will provide (and the uses of those knowledge once it has emerged) is unknown before the analytics is done and the outcome of the analysis is delivered. The data analytics will in other terms often shed light on unexpected correlations or patterns and, only once those correlations have emerged (ex post), their value and the potential use of those will become clear.

Hence, either we give a formalistic application to the GDPR and consider sufficient to insert in the initial privacy notices a generic warn that data could be subject to data analytics and used for marketing; or, conversely, we impose new and more detailed notices to the data subject from time to time as soon as more information on purposes and methodologies of processing is available, which risks to compel data players to continuously update initial privacy notices with a storm of confusing and constant communications.

It seems evident that the mechanism itself (notice and purpose limitation) works well in a one-to-one legal relationship (one controller vis-à-vis one data subject) and does not work efficiently when there are intermediate controllers who handle billion of data on million people.

---

identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them.

6. – GDPR is strict on the ability of data controllers to profile natural persons. As a start, the notion of profiling [21] is broad and includes "*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*".

On top of that it should also be considered that the GDPR applies extensively in terms of territory [22], its Art. 3.2.b in fact clarifies amongst others that "*GDPR applies to the processing of personal data of data subjects who are in the Union by a controller or processor* (even) *not established in the Union, where the processing activities are related to the monitoring of their behaviour as far as their behaviour takes place within the Union*".

The above results in a large notion of profiling and a large scope of territorial application [23]. And many are the rules applicable to profiling under the GDPR that create issues in a Big Data economy.

a) Pursuant to Art. 13.2.f GDPR [24] (information to be provided where personal data

---

[21] DESTRI, LOTTO, *La profilazione*, in *Il processo di adeguamento al GDPR*, Milano, 2018, 459; PIERUCCI, *Elaborazione dei dati e profilazione delle persone*, in *I dati personali nel diritto europeo*, Torno, 2019, p. 131.

[22] ODDENINO, *Reflections on big data and international law*, in *Dir. Comm. Int.*, 2017, p. 777. EUROPEAN DATA PROTECTION BOARD, *Guidelines 3/2018 on the territorial scope of the GDPR (Article 3)*, November 12, 2019.

[23] That extra-territorial application has however some limits. In its recent decision, the European Court of Justice (ECJ, 24 September 2019, Google LLC vs. CNIL, Case C-507/17) has settled a dispute between Google and the CNIL (the French Data Protection Authority) as to how a search engine operator is to give effect to the right to de-referencing. Where a search engine operator grants a request for de-referencing, is that operator required to carry out that de-referencing on all versions of its search engine, or, on the contrary, only on the versions of that search engine corresponding to all the EU Member States? The fact that a search engine is operated by an undertaking that has its seat in a third State cannot result in the processing of personal data escaping the obligations and guarantees laid down by EU laws. The internet is a global network without borders and search engines render the information and links contained in a list of results displayed following a search conducted on the basis of an individual's name ubiquitous. In a globalised world, internet users'access — including those outside the EU — to the referencing of a link referring to information regarding a person whose centre of interests is situated in the EU is thus likely to have immediate and substantial effects on that person within the EU itself. Still, EU legislature have not chosen to confer a scope on the rights which goes beyond the territory of the Member States, not to impose on an operator, like Google, a de-referencing obligation which also concerns the national versions of its search engine that do not correspond to the Member States. It follows that, currently, there is no obligation under EU law, for a search engine operator, to carry out such a de-referencing on all the versions of its search engine.

[24] See VECCHI, MARCHESE, ORTILLO, *Le disposizioni generali e le definizioni in materia di protezione*

are collected from the data subject), "the data controller shall, at the time when personal data are obtained, provide the data subject with the following […] information necessary to ensure fair and transparent processing: […] the existence of automated decision-making, including profiling, […] and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [25]. But how may that happen when the "logic" is developed through AI (i.e., a non-human intelligence that develops and transform itself over time and departs from the initial human-written algorithm by machine learning)? In other terms, how can a controller inform a data subject on something that the controller itself does not know and cannot know because the logic is developed by an AI beyond human abilities?

b) One should also consider that the hidden patterns and correlations that are discovered through data analytics (and which allow the profiling of an individual) are the result of the analytics and become evident only at the end of it when their value and potential use is uncovered. We are used to learn on books about the scientific method: a scientist observes something that happens on nature, conceives a theory that could explain it and at this point tests the validity of such theory with experiments. In a Big Data world the process is reversed. There is no (human) theory at inception and then specific (human) test. What happens is often the reverse. Large amounts of different data are put in the turbine of data analytics and unexpected and non-searched correlations emerge, which of course have a value but were not intended to be found. But then, if it is so, how can a controller inform the data subject on consequences of the processing when the outcomes of the analytics are defined only at the end of the process and not before?

7. – As we have briefly mentioned earlier, the GDPR contains a very strict regulation of decisions taken with automated means (including decisions based on automated profiling) by processing personal data.

---

dei dati personali, in *Commentario al regolamento UE 2016/679 e al codice della privacy aggiornato*, Ed. Top Legal, 2019, p. 121.

[25] Similarly, Art. 15.1.h of GDPR provides that: "The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: […]the existence of automated decision-making, including profiling, [….] and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

Pursuant to Articles 22 and 70.1.f GDPR [26],

1. the data subject has the right not to be subject to a decision based 'solely' on automated processing, which produces legal effects concerning him or her or similarly significantly affects him or her;

2. the prohibition above does not apply only if the decision: (a) is (strictly) necessary for entering into, or performing a contract between the data subject and the data controller; or (b) is authorized by EU or EU Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's 'explicit' consent;

3. in the cases referred to in points 2.(a) and 2.(c), the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision; and, in any event,

4. automated decisions may not be based on special categories of personal data (health, political orientation, etc.) [27] unless there is the data subject's 'explicit' consent or there is substantial public interest [28] and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place [29].

Now, it is hard to reconcile that with a Big Data environment.

In fact:

---

[26] MESSINETTI, *La tutela della persona umana versus l'intelligenza artificiale. Potere decisionale dell'apparato tecnologico e diritto alla spiegazione della decisione automatizzata*, in *Contr. Impr.*, 2019, p. 861.

[27] Categories listed in Article 9(1) GDPR.

[28] Point (a) or (g) of Article 9(2) GDPR.

[29] Whereas 71 of the GDPR explains that "*the data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. […]. However, decision-making based on such processing, including profiling, should be allowed where expressly authorized by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision*".

– data analytics are often based on AI and AI produces automated decisions, including decisions based solely on AI; forbidding only automated decisions is a great roadblock on AI's full exploitation and moves from an assumption, that a human decision is qualitatively better than an AI decision;

– explicit, free, informed and unequivocal consent [30] (which is one of the ways to unblock the possibility of automated decisions) requires full data subject's information on purposes and methodologies of the data analytics but, as we have seen above, that is hardly doable when the purpose and the outcome of data analytics cannot be defined (and communicated) before the processing and inferences and correlations and their potential uses and value emerge only after such analytics; notices to data subjects at the start of the process will therefore be inevitably general, but that conflicts with the need of an explicit and full consent;

– stating that safeguards are necessary in just right, but which safeguards are necessary in an AI and Big Data scenario? I suspect that regulation should clarify it and clear the path for those who want to engage in this business without taking imponderable risks;

– which meaning can we give to the 'human intervention factor' that the controller must ensure upon data subject's request? When is that intervention enough to satisfy the rule and when is it a fake or irrelevant intervention?

Whereas 71 of the GDPR indicates a path: "*In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should*

a) *use appropriate mathematical or statistical procedures for the profiling,*

b) *implement technical and organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized,*

c) *secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect*".

---

[30] CAGGIA, *Il consenso al trattamento dei dati personali nel diritto europeo*, in *Riv. Dir. Comm.*, 2019, p. 1.

However, without guidance on what is fair and transparent, appropriate and secure I am afraid that we will be lost in translation between the general principles of the law and the need of specific rules which can give firm ground to investments in this field.

The very same problems arise when we confront ourselves with the need of a data protection impact assessment (DPIA) based on Art. 35 of GDPR. In accordance with that article a DPIA "*shall in particular be required in the case of […] a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person*" and the DPIA shall contain at least: "*(a) a systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller; (b) an assessment of the necessity and proportionality of the processing operations in relation to the purposes; (c) an assessment of the risks to the rights and freedoms of data subjects; and (d) the measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with GDPR taking into account the rights and legitimate interests of data subjects and other persons concerned. Where appropriate, the controller shall seek the views of data subjects or their representatives on the intended processing, without prejudice to the protection of commercial or public interests or the security of processing operations*.

Once again, the question is how that may happen with respect to data analytics and AI, when the outcome of the processing (and the use of it that the controller will do) come to the surface after the analytics being made and the unforeseen correlations are found.

More, can we really expect that views of data subjects ate sought? Regardless of the disposition of the data controller to request those views, it is the same width in the scope of data and multitude of data subjects that render the exercise extremely heavy.

8. – The legal philosophy at the base of the GDPR seems to diverge considerably as compared to the business philosophy that boosts data analytics [31].

[31] FAINI, *Big data e internet of thinghs: data protection e data governance alla luce del regolamento europeo*, in *Il processo di adeguamento al GDPR*, Milano, 2018, 459. CESANA, *Dati personali e big data: la nuova privacy europea*, in *Big Data: privacy, gestione, tutele*, Torino, 2018, p. 111. PIRAINO, *Il regolamento generale sulla protezione dei dati personali e i diritti dell'interessato*, in Nuove leggi civ. comm., 2017,

– The GDPR establishes the need of a risk based approach when processing data (the other tip of the coin is the accountability of the controller), whereas risks of big data analytics are inherently high.

– The GDPR is based on purpose limitation, whereas data analytics is carried out to produce outcome whose uses are not visible before analytics is made. Further, the GDPR imposes transparency whereas it is almost impossible to provide precise information on purpose and use of the output before data are collected and the analysis is done.

– The GDPR stipulates especially strict rules on profiling and automated decisions, but this hardly can match with AI.

– The GDPR protects individuals, whereas Big Data pose issues also for groups and communities.

This is why In July 2019, the Italian data protection, antitrust and TLC authorities issued guidelines that stressed the following points [32]:

a) Information duties in GDPR are not enough,

b) Approach limited to individuals does not suffice,

c) Big Data are inherently transnational and must be addressed through international cooperation,

d) Anonymization must be actual,

e) Control of unfair practices by antitrust authority has to be enhanced and new merger controls approach has to be taken into account,

---

p. 369. MAGLIO, *Ripensare il futuro della protezione dei dati personali. Le tecnologie in evoluzione nell'era dell'intelligenza artificiale e dei big data*, in *Manuale di diritto alla protezione dei dati personali*, Maggioli, 2019, p. 871.

[32] In July 2019, the Italian GARANTE PER LA PROTEZIONE DEI DATI PERSONALI, the AUTORITÀ GARANTE DELLA CONCORRENZA E DEL MERCATO and AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI issued guidelines that stressed the following points: 1. *Government and Parliament to consider new rules for full and effective transparency on use of personal data (towards individuals and collectivities), 2. International cooperation to be reinforced, 3. Single policy on extraction, access and use of publicly available data for public use and coordination with single digital market*, 4. *Reduce information asymmetries between users and digital players and between platforms and other business players, 5. Determine nature and title to data and effective anonymization prior to processing data*, 6. Provide n*ew instruments for on-line pluralism, transparency and awareness, 7. Consumerist approach through antitrust instruments to be extended to service level, innovation and equity, 8. Reform merger controls, 9. Enhance portability and mobility of data, 10. New audit rights for authorities and increase sanctions, 11. Permanent coordination among the three authorities.* See also Delibera 217/17/CONS of May 30, 2017 of the AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI, *Avvio di un'indagine conoscitiva sui big data*; and AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI, *Interim report nell'ambito dell'indagine conoscitiva di cui alla delibera 217/17/CONS* of June 2018.

f) Data portability is an effective tool to rebalance the position of data subjects vis-à-vis controller and it should be extended also beyond GDPR rules,

g) Powers of public authorities in TLC, antitrust and data protection must be exercised consistently in an integrated fashion.

We cannot be content with simply underlining the above legal issues. A legal solution should be sought. And we cannot but only looking forward new regulations.

Potentially, one way through could be to consider data protection rules not as an isolated archipelago but as part of the thorough civil and commercial law system and, hence, apply to Big Data a combination of rules taken from the GDPR and the legal standards of EU continental law (fairness and loyalty, consumer protection, etc.).

If we look to Big Data from that standpoint, some solutions seem to emerge.

i. On the inherent risk of Big Data [33]: Italian law does not prohibit an entrepreneurial activity because it is risky. On the contrary, specific standards are rules out in case of activities embedding significant risks. Pursuant to art. 2050 of the Italian civil code, the one who causes third party damage in the pursuance of the a risky activity, whether because of its very nature or the means that are used for that activity, shall redress the damage unless it proves to have used all the adequate measures to avoid the damage. That rule should apply to Big Data, it being however understood that the rules must be interpreted on the basis of the 'state of the art' without forcing players to unreasonable spending [34].

---

[33] See the extensive and argumented reasoning in *Intelligenza Artificiale e diritto*, in *Giur. It.* 2019, 9. 1657 ff., especially RUFFOLO, *Intelligenza artificiale, machine learning e responsabilità da algoritmo*, p. 1689.

[34] Whereas 78 GDPR states: "*When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations*". And Whereas 81 adds: "*In order to maintain security and to prevent processing in infringement of this Regulation, the controller or processor should evaluate the risks inherent in the processing and implement measures to mitigate those risks, such as encryption. Those measures should ensure an appropriate level of security, including confidentiality, taking into account the state of the art and the costs of implementation in relation to the risks and the nature of the personal data to be protected. In assessing data security risk, consideration should be given to the risks that are presented by personal data processing, such as accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed which may in particular lead to physical, material or non-material damage"*. The same principles can be traced in Art. 25.1 GDPR (privacy by design) and Art. 32.1 (security of processing).

ii. On data transparency and purpose limitation: the data subject must be informed that his/her data that those data could be used for data analytics and should also be informed at least of the general purposes of that analytics (i.e., whether for marketing, of public interest, or political surveys or other). This may be way too generic for full transparency and explicit consent and, conversely, mandating Big Data players to continuously integrate and detail information each time that analytics is done could be far too burdensome (considering that data subjects to be informed could be in the range of million). However, nothing forces the data controllers to provide the privacy notices with individual letters (one per each data subject). A privacy notice on the website could be enough [35] also because other way of notifying individuals can be disproportionate [36]. Hence, a general privacy notice at the time of collecting the data could cover data analytics by a Big Data player if that one publicly provides more detailed information on its website.

iii. On profiling and automated decisions: the same line of reasoning could be adopted and the requirements of the GDPR should be viewed taking into account fairness, state of the art and reasonable costs. If so, information on the logic and consequences of the processing and DPIA could be make available through the Big Data player's website and human intervention and safeguards adopted to govern automated decisions will need to be interpreted not as absolute guarantees (*obbligazioni di risultato*) but as best effort obligations to be bench marketed to the state of the art.

In sum, what really counts is to put data subjects in the position to exercise their right to object [37]. Once this result is achieved, the market will regulate itself.

---

[35] Art. 12.1 GDPR states that *"The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means"*.

[36] Art. 14.5(b) GDPR.

[37] The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions. The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims. Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.

Abstract

This short article recapitulates the notion of Big Data and the difficult match between Big Data and current data protection rules (purpose limitation, transparency, consent, profiling and automated decisions). After doing that, the Author tries to indicate an interpretative way to reconcile massive use of large data sets to profile individuals with the principles of the GDPR.