The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 641.530 – June 2020

EN

The impact of the General Data Protection Regulation (GDPR) on artificial intelligence

This study addresses the relationship between the General Data Protection Regulation (GDPR) and artificial intelligence (AI). After introducing some basic concepts of AI, it reviews the state of the art in AI technologies and focuses on the application of AI to personal data. It considers challenges and opportunities for individuals and society, and the ways in which risks can be countered and opportunities enabled through law and technology.

The study then provides an analysis of how AI is regulated in the GDPR and examines the extent to which AI fits into the GDPR conceptual framework. It discusses the tensions and proximities between AI and data protection principles, such as, in particular, purpose limitation and data minimisation. It examines the legal bases for AI applications to personal data and considers duties of information concerning AI systems, especially those involving profiling and automated decision-making. It reviews data subjects' rights, such as the rights to access, erasure, portability and object.

The study carries out a thorough analysis of automated decision-making, considering the extent to which automated decisions are admissible, the safeguard measures to be adopted, and whether data subjects have a right to individual explanations. It then addresses the extent to which the GDPR provides for a preventive risk-based approach, focusing on data protection by design and by default. The possibility to use AI for statistical purposes, in a way that is consistent with the GDPR, is also considered.

The study concludes by observing that AI can be deployed in a way that is consistent with the GDPR, but also that the GDPR does not provide sufficient guidance for controllers, and that its prescriptions need to be expanded and concretised. Some suggestions in this regard are developed.

AUTHOR

The study was led by Professor Giovanni Sartor, European University Institute of Florence, at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament. It was co-authored by Professor Sartor and Dr Francesca Lagioia, European University Institute of Florence, working under his supervision.

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in June 2020.

PE 641.530
ISBN: 978-92-846-6771-0
doi: 10.2861/293
QA-QA-02-20-399-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary

AI and big data

In the last decade, AI has gone through rapid development. It has acquired a solid scientific basis and has produced many successful applications. It provides opportunities for economic, social, and cultural development; energy sustainability; better health care; and the spread of knowledge. These opportunities are accompanied by serious risks, including unemployment, inequality, discrimination, social exclusion, surveillance, and manipulation.

There has been an impressive leap forward on AI since it began to focus on the application of machine learning to mass volumes of data. Machine learning systems discover correlations between data and build corresponding models, which link possible inputs to presumably correct responses (predictions). In machine learning applications, AI systems learn to make predictions after being trained on vast sets of examples. Thus, AI has become hungry for data, and this hunger has spurred data collection, in a self-reinforcing spiral: the development of AI systems based on machine learning presupposes and fosters the creation of vast data sets, i.e., big data. The integration of AI and big data can deliver many benefits for the economic, scientific and social progress. However, it also contributes to risks for individuals and for the whole of society, such as pervasive surveillance and influence on citizens' behaviour, polarisation and fragmentation in the public sphere.

AI and personal data

Many AI applications process personal data. On the one hand, personal data may contribute to the data sets used to train machine learning systems, namely, to build their algorithmic models. On the other hand, such models can be applied to personal data, to make inferences concerning particular individuals.

Thanks to AI, all kinds of personal data can be used to analyse, forecast and influence human behaviour, an opportunity that transforms such data, and the outcomes of their processing, into valuable commodities. In particular, AI enables automated decision-making even in domains that require complex choices, based on multiple factors and non-predefined criteria. In many cases, automated predictions and decisions are not only cheaper, but also more precise and impartial than human ones, as AI systems can avoid the typical fallacies of human psychology and can be subject to rigorous controls. However, algorithmic decisions may also be mistaken or discriminatory, reproducing human biases and introducing new ones. Even when automated assessments of individuals are fair and accurate, they are not unproblematic: they may negatively affect the individuals concerned, who are subject to pervasive surveillance, persistent evaluation, insistent influence, and possible manipulation.

The AI-based processing of vast masses of data on individuals and their interactions has social significance: it provides opportunities for social knowledge and better governance, but it risks leading to the extremes of 'surveillance capitalism' and 'surveillance state'.

A normative framework

It must be ensured that the development and deployment of AI tools takes place in a socio-technical framework – inclusive of technologies, human skills, organisational structures, and norms – where individual interests and the social good are preserved and enhanced.

To provide regulatory support for the creation of such a framework, ethical and legal principles are needed, together with sectorial regulations. The ethical principles include autonomy, prevention of harm, fairness and explicability; the legal ones include the rights and social values enshrined in the EU charter, in the EU treaties, as well as in national constitutions. The sectoral regulations involved include first of all data protection law, consumer protection law, and competition law, but also other

domains of the law, such as labour law, administrative law, civil liability etc. The pervasive impact of AI on European society is reflected in the multiplicity of the legal issues it raises.

To ensure adequate protection of citizens against the risks resulting from the misuses of AI, beside regulation and public enforcement, the countervailing power of civil society is also needed to detect abuses, inform the public, and activate enforcement. AI-based citizen-empowering technologies can play an important role in this regard, by enabling citizens not only to protect themselves from unwanted surveillance and 'nudging', but also to detect unlawful practices, identify instances of unfair treatment, and distinguish fake and untrustworthy information.

AI is compatible with the GDPR

AI is not explicitly mentioned in the GPDR, but many provisions in the GDPR are relevant to AI, and some are indeed challenged by the new ways of processing personal data that are enabled by AI. There is indeed a tension between the traditional data protection principles – purpose limitation, data minimisation, the special treatment of 'sensitive data', the limitation on automated decisions – and the full deployment of the power of AI and big data. The latter entails the collection of vast quantities of data concerning individuals and their social relations and processing such data for purposes that were not fully determined at the time of collection. However, there are ways to interpret, apply, and develop the data protection principles that are consistent with the beneficial uses of AI and big data.

The requirement of purpose limitation can be understood in a way that is compatible with AI and big data, through a flexible application of the idea of compatibility, which allows for the reuse of personal data when this is not incompatible with the purposes for which the data were originally collected. Moreover, reuse for statistical purposes is assumed to be compatible, and thus would in general be admissible (unless it involves unacceptable risks for the data subject).

The principle of data minimisation can also be understood in such a way as to allow for beneficial applications of AI. Minimisation may require, in some contexts, reducing the 'personality' of the available data, rather than the amount of such data, i.e., it may require reducing, through measures such as pseudonymisation, the ease with which the data can be connected to individuals. The possibility of re-identification should not entail that all re-identifiable data are considered personal data to be minimised. Rather the re-identification of data subjects should be considered as creation of new personal data, which should be subject to all applicable rules. Re-identification should indeed be strictly prohibited unless all conditions for the lawful collection of personal data are met, and it should be compatible with the purposes for which the data were originally collected and subsequently anonymised.

The information requirements established by the GDPR can be met with regard to AI-based processing, even though the complexity of AI application has to be taken into account. The information made available to data subjects should enable them to understand the purpose of each AI-based processing and its limits, even without going into unnecessary technical details.

The GDPR allows for inferences based on personal data, provided that appropriate safeguards are adopted. Profiling is in principle prohibited, but there are ample exceptions (contract, law or consent). Uncertainties exist concerning the extent to which an individual explanation should be provided to the data subject. It is also uncertain to what extent reasonableness criteria may apply to automated decisions.

The GDPR provisions on preventive measures, and in particular those concerning privacy by design and by default, do not hinder the development of AI systems, if correctly designed and implemented, even though they may entail some additional costs. It needs to be clarified which AI applications present high risks and therefore require a preventive data protection assessment, and possibly the preventive involvement of data protection authorities.

Finally, the possibility of using personal data for statistical purposes opens opportunities for the processing of personal data in ways that do not involve the inference of new personal data. Statistical processing requires security measures that are proportionate to the risks for the data subject, and which should include at least pseudonymisation.

The GDPR prescriptions are often vague and open-ended

The GDPR allows for the development of AI and big data applications that successfully balance data protection and other social and economic interests, but it provides limited guidance on how to achieve this goal. It indeed abounds in vague clauses and open standards, the application of which often requires balancing competing interests. In the case of AI/big data applications, the uncertainties are aggravated by the novelty of the technologies, their complexity and the broad scope of their individual and social effects.

It is true that the principles of risk-prevention and accountability potentially direct the processing of personal data toward a 'positive sum' game, in which the advantages of the processing, when constrained by appropriate risk-mitigation measures, outweigh its possible disadvantages. Moreover these principles enable experimentation and learning, avoiding the over- and under-inclusiveness issues involved in the applications of strict rules. However, by requiring controllers to rely on such principles, the GDPR offloads the task of establishing how to manage risk and find optimal solutions onto controllers, a task that may be challenging as well as costly. The stiff penalties for non-compliance, when combined with the uncertainty on the requirements for compliance, may constitute a novel risk, which, rather than incentivising the adoption of adequate compliance measure, may prevent small companies from engaging in new ventures.

Thus, the successful application of GDPR to AI-application depends heavily on what guidance data protection bodies and other competent authorities will provide to controllers and data subjects. Appropriate guidance would diminish the cost of legal uncertainty and would direct companies – in particular small ones that mostly need such advice – to efficient and data protection-compliant solutions.

Some policy indications

The study concludes with the following indications on AI and the processing of personal data.

- The GDPR generally provides meaningful indications for data protection in the context of AI applications.
- The GDPR can be interpreted and applied in such a way that it does not substantially hinder the application of AI to personal data, and that it does not place EU companies at a disadvantage by comparison with non-European competitors.
- Thus, the GDPR does not require major changes in order to address AI applications.
- However, a number of AI-related data-protection issues do not have an explicit answer in the GDPR. This may lead to uncertainties and costs, and may needlessly hamper the development of AI applications.
- Controllers and data subjects should be provided with guidance on how AI can be applied to personal data consistently with the GDPR, and on the available technologies for doing so. Such guidance can prevent costs linked to legal uncertainty, while enhancing compliance.
- Providing guidance requires a multilevel approach, which involves data protection authorities, civil society, representative bodies, specialised agencies, and all stakeholders.
- A broad debate is needed involving not only political and administrative authorities, but also civil society and academia. This debate needs to address the issues of determining what

standards should apply to AI processing of personal data, particularly to ensure the acceptability, fairness and reasonability of decisions on individuals. It should also address what applications are to be barred unconditionally, and which ones may instead be admitted only under specific circumstances and controls.

- Discussion of a large set of realistic examples is needed to clarify which AI applications are socially acceptable, under what circumstances and with what constraints. The debate on AI can also be an opportunity to reconsider in depth, with more precision and concreteness, some basic ideas of law and ethics, such as acceptable and practicable conceptions of fairness and non-discrimination.
- Political authorities, such as the European Parliament, the European Commission and the Council should provide general open-ended indications about the values at stake and ways to achieve them.
- Data protection authorities, and in particular the Data Protection Board, should provide controllers with specific guidance on the many issues for which no precise answer can be found in the GDPR. Such guidance can often take the form of soft law instruments designed with dual legal and technical competence, as in the case of Article 29 Working Party opinions.
- National Data Protection Authorities should also provide guidance, in particular when contacted for advice by controllers, or in response to data subjects' queries.
- The fundamental data protection principles – especially purpose limitation and minimisation – should be interpreted in such a way that they do not exclude the use of personal data for machine learning purposes. They should not preclude the creation of training sets and the construction of algorithmic models, whenever the resulting AI systems are socially beneficial and compliant with data protection rights.
- The use of personal data in a training set, for the purpose of learning general correlations and connection, should be distinguished from their use for individual profiling, which is about making assessments about individuals.
- The inference of new personal data, as it is done in profiling, should be considered as creation of new personal data, when providing an input for making assessments and decisions. The same should apply to the re-identification of anonymous or pseudonymous data.
- Guidance is needed on profiling and automated decision-making. It seems that an obligation of reasonableness – including normative and reliability aspects – should be imposed on controllers engaging in profiling, mostly, but not only when profiling is aimed at automated decision-making. Controllers should also be under an obligation to provide individual explanations, to the extent that this is possible according to the available AI technologies, and reasonable according to costs and benefits. The explanations may be high-level, but they should still enable users to contest detrimental outcomes.
- It may be useful to establish obligations to notify data protection authorities of applications involving individualised profiling and decision-making, possibly accompanied with the right to ask for indications on compliance.
- The content of the controller's obligation to provide information (and the corresponding rights of the data subject) about the 'logic' of an AI system need to be specified, with appropriate examples, an in relation to different technologies.

- It needs to be ensured that the right to opt out of profiling and data transfers can easily be exercised, through appropriate user interfaces. The same applies to the right to be forgotten.
- Normative and technological requirements concerning AI by design and by default need to be specified.
- The possibility of repurposing data for AI applications that do not involve profiling – scientific and statistical ones – need to be broad, as long as appropriate precautions are in place preventing abuse.
- Strong measures need to be adopted against companies and public authorities that intentionally abuse the trust of data subjects by using their data against their interests.
- Collective enforcement in the data protection domain should be enabled and facilitated.

In conclusion, controllers engaging in AI-based processing should endorse the values of the GDPR and adopt a responsible and risk-oriented approach. This can be done in ways that are compatible with the available technology and economic profitability (or the sustainable achievement of public interests, in the case of processing by public authorities). However, given the complexity of the matter and the gaps, vagueness and ambiguities present in the GDPR, controllers should not be left alone in this exercise. Institutions need to promote a broad societal debate on AI applications, and should provide high-level indications. Data protection authorities need to actively engage in a dialogue with all stakeholders, including controllers, processors, and civil society, in order to develop appropriate responses, based on shared values and effective technologies. Consistent application of data protection principles, when combined with the ability to efficiently use AI technology, can contribute to the success of AI applications, by generating trust and preventing risks.

Table of Contents

Table of figures

1. Introduction

This study aims to provide a comprehensive assessment of the interactions between artificial intelligence (AI) and data protection, focusing on the 2016 EU General Data Protection Regulation (GDPR).

Artificial intelligence systems are populating the human and social world in multiple varieties: industrial robots in factories, service robots in houses and healthcare facilities, autonomous vehicles and unmanned aircraft in transportation, autonomous electronic agents in e-commerce and finance, autonomous weapons in the military, intelligent communicating devices embedded in every environment. AI has come to be one of the most powerful drivers of social transformation: it is changing the economy, affecting politics, and reshaping citizens' lives and interactions. Developing appropriate policies and regulations for AI is a priority for Europe, since AI increases opportunities and risks in ways that are of the greatest social and legal importance. AI may enhance human abilities, improve security and efficiency, and enable the universal provision of knowledge and skills. On the other hand, it may increase opportunities for control, manipulation, and discrimination; disrupt social interactions; and expose humans to harm resulting from technological failures or disregard for individual rights and social values.

A number of concrete ethical and legal issues have already emerged in connection with AI in several domains, such as civil liability, insurance, data protection, safety, contracts and crimes. Such issues acquire greater significance as more and more intelligent systems leave the controlled and limited environments of laboratories and factories and share the same physical and virtual spaces with humans (internet services, roads, skies, trading on the stock exchange, other markets, etc.).

Data protection is at the forefront of the relationship between AI and the law, as many AI applications involve the massive processing of personal data, including the targeting and personalised treatment of individuals on the basis of such data. This explains why data protection has been the area of the law that has most engaged with AI, although other domains of the law are involved as well, such as consumer protection law, competition law, antidiscrimination law, and labour law.

This study will adopt an interdisciplinary perspective. Artificial intelligence technologies will be examined and assessed on the basis of most recent scientific and technological research, and their social impacts will be considered by taking account of an array of approaches, from sociology to economics and psychology. A normative perspective will be provided by works in sociology and ethics, and in particular information, computer, and machine ethics. Legal aspects will be analysed by reference to the principles and rules of European law, as well as to their application in national contexts. The report will focus on data protection and the GDPR, though it will also consider how data protection shares with other domains of the law the task of addressing the opportunities and risks that come with AI.

2. AI and personal data

This section introduces the technological and social background of the study, namely, the development of AI and its connections with the processing of personal and other data. First the concept of AI will be introduced (Section 2.1), then the parallel progress of AI and large-scale data processing will be discussed (Section 2.2), and finally, the analysis will turn to the relation between AI and the processing of personal data (Section 2.3).

2.1. The concept and scope of AI

The concept of AI will be introduced, as well as its connections with the robotics and algorithms.

2.1.1. A definition of AI

The broadest definition of artificial intelligence (AI) characterises it as the attempt to build machines that 'perform functions that require intelligence when performed by people.'[1] A more elaborate notion has been provided by the High Level Expert Group on AI (AI HLEG), set up by the EU Commission:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.[2]

This definition can be accepted with the proviso that most AI systems only perform a fraction of the activities listed in the definition: pattern recognition (e.g., recognising images of plants or animals, human faces or attitudes), language processing (e.g., understanding spoken languages, translating from one language into another, fighting spam, or answering queries), practical suggestions (e.g., recommending purchases, purveying information, performing logistic planning, or optimising industrial processes), etc. On the other hand, some systems may combine many such capacities, as in the example of self-driving vehicles or military and care robots.

The High-Level Expert Group characterises the scope of research in AI as follows:

> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

To this definition, we could also possibly add also communication, and particularly the understanding and generation of language, as well as the domains of perception and vision.

[1] Kurzweil (1990, 14), Russel and Norvig (2016, Section 1.1).

[2] AI-HLEG (2019).

2.1.2. AI and robotics

AI constitutes the core of robotics, the discipline that aims to build 'physical agents that performs tasks by manipulating the physical world.'[3] The High-Level Expert Group describes robotics as follows

> Robotics can be defined as 'AI in action in the physical world' (also called *embodied AI*). A robot is a physical machine that has to cope with the dynamics, the uncertainties and the complexity of the physical world. Perception, reasoning, action, learning, as well as interaction capabilities with other systems are usually integrated in the control architecture of the robotic system. In addition to AI, other disciplines play a role in robot design and operation, such as mechanical engineering and control theory. Examples of robots include robotic manipulators, autonomous vehicles (e.g. cars, drones, flying taxis), humanoid robots, robotic vacuum cleaners, etc.

In this report, robotics will not be separately addressed since embodied and disembodied AI systems raise similar concerns when addressed from the perspective of GDPR: in both cases personal data are collected, processed, and acted upon by intelligent system. Moreover, also software systems may have access to sensor on the physical world (e.g., cameras) or govern physical devices (e.g., doors, lights, etc.). This fact does not exclude that the specific types of interaction that may exists, or will exists, between humans and physical robots – e.g., in the medical or care domain– may require specific considerations and regulatory approaches also in the data protection domain.

2.1.3. AI and algorithms

The term 'algorithm' is often used to refer to AI applications, e.g., through locutions such 'algorithmic decision-making.' However, the concept of an algorithm is more general that the concept of AI, since it includes any sequence of unambiguously defined instructions to execute a task, particularly but not exclusively through mathematical calculations.[4] To be executed by a computer system, algorithms have to be expressed through programming languages, thus becoming machine-executable software programs. Algorithms can be very simple, specifying, for instance, how to arrange lists of words in alphabetical order or how to find the greatest common divisor between two numbers (such as the so-called Euclidean algorithm). They can also be very complex, such as algorithms for file encryption, the compression of digital files, speech recognition, or financial forecasting. Obviously, not all algorithms involve AI, but every AI system, like any computer system, includes algorithms, some dealing with tasks that directly concern AI functions.

AI algorithms may involve different kinds of epistemic or practical reasoning (detecting patterns and shapes, applying rules, making forecasts or plans), as well different ways of learning.[5] In the latter case the system can enhance itself by developing new heuristics (tentative problem-solving strategies), modifying its internal data, or even generating new algorithms. For instance, an AI system for e-commerce may apply discounts to consumers meeting certain conditions (apply rules), provide recommendations (e.g., learn and use correlations between users' features and their buying habits), optimise stock management (e.g., develop and deploy the best trading strategies). Though an AI system includes many algorithms, it can also be viewed as a single complex algorithm, combining the algorithms performing its various functions, as well as the top algorithms that orchestrate the system's functions by activating the relevant lower-level algorithms. For instance, a bot that answers queries in natural language will include an orchestrated combination of algorithms

[3] Russell and Norvig (2016).

[4] Harel (2004).

[5] According to Russel and Norvig (2016, 693), 'an agent is learning if it improves its performance on future tasks after making observations about the world'.

to detect sounds, capture syntactic structures, retrieve relevant knowledge, make inferences, generate answers, etc.

In a system that is capable of learning, the most important component will not be the learned algorithmic model, i.e., the algorithms that directly execute the tasks assigned to the system (e.g., making classifications, forecasts, or decisions) but rather the learning algorithms that modify the algorithmic model so that it better performs its function. For instance, in a classifier system that recognises images through a neural network, the crucial element is the learning algorithm (the trainer) that modifies the internal structure of the algorithmic model (the trained neural network) by changing it (by modifying its internal connections and weights) so that it correctly classifies the objects in its domain (e.g., animals, sounds, faces, attitudes, etc.).

2.1.4. Artificial intelligence and big data

The term *big data* identifies vast data sets that it is difficult to manage using standard techniques, because of their special features, the so-called thee V's: huge Volume, high Velocity and great Variety. Other features associated to big data are low Veracity (high possibility that at least some data are inaccurate), and high Value. Such data can be created by people, but most often they are collected by machines, which capture information from the physical word (e.g., street cameras, sensors collecting climate information, devices for medical testing, etc.), or from computer-mediated activities (e.g., systems recording transactions or tracking online behaviour etc.).

From a social and legal perspective what is most relevant in very large data sets, and which makes them 'big data' from a functional perspective, is the possibility of using such data sets for analytics, namely, for discovering correlations and making predictions, often using AI techniques, as we shall see when discussing machine learning.[6] In particular, the connection with analytics and AI makes big data specifically relevant to data protection.[7]

Big data can concern the non-human physical world (e.g. environmental, biological, industrial, and astronomical data), as well as humans and their social interactions (e.g., data on social networks, health, finance, economics or transportation). Obviously, only the second kind of data is relevant to this report.

2.2. AI in the new millennium

Over the last decades, AI has gone through a number of ups and downs, excessive expectations being followed by disillusion (the so-called AI winters).[8] In recent years, however, there is no doubt that AI has been hugely successful. On the one hand, a solid interdisciplinary background has been constructed for AI research: the original core of computing, mathematics, and logic has been extended with models and insights from a number of other disciplines, such as statistics, economics, linguistics, neurosciences, psychology, philosophy, and law. On the other hand, an array of successful applications has been built, which have already entered our daily lives: voice, image, and face recognition; automated translation; document analysis; question-answering; games; high-speed trading; industrial robotics; autonomous vehicles; etc.

Based on the current successes, it is most likely that current successful applications will not only be consolidate, but will be accompanied by further growth, following probably the middle path indicated in Figure 1.

[6] See Mayer-Schoenberger and Cukier (2013, 15).

[7] Hildebrandt (2014)

[8] Nilsson (2010).

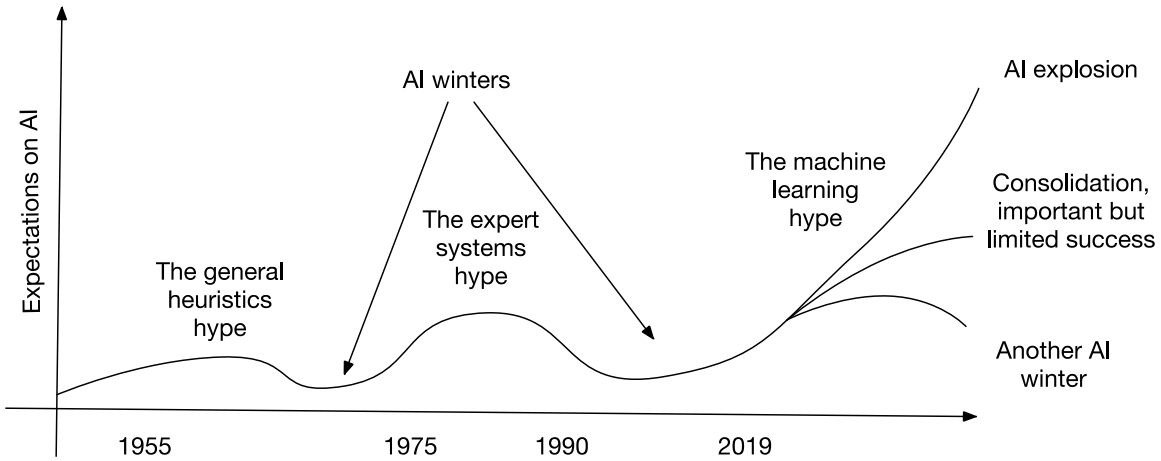


Figure 1 – Hypes and winters of AI

2.2.1. Artificial general and specific intelligence

AI research usually distinguishes two goals: 'artificial general intelligence,' also known as 'strong AI,' and 'artificial specialised intelligence,' also known as 'weak AI.' Artificial general intelligence pursues the ambitious objective of developing computer systems that exhibit most human cognitive skills, at a human or even a superhuman level.[9] Artificial specialised intelligence pursues a more modest objective, namely, the construction of systems that, at a satisfactory level, are able to engage in specific tasks requiring intelligence.

The future emergence of a general artificial intelligence is already raising serious concerns. A general artificial intelligence system may improve itself at an exponential speed and quickly become superhuman; through its superior intelligence it may then acquire capacities beyond human control.[10] In relation to self-improving artificial intelligence, humanity may find itself in a condition of inferiority similar to that of animals in relation to humans. Some leading scientists and technologists (such as Steven Hawking, Elon Musk, and Bill Gates) have argued for the need to anticipate this existential risk,[11] adopting measures meant to prevent the creation of general artificial intelligence or to direct it towards human-friendly outcomes (e.g., by ensuring that it endorses human values and, more generally, that it adopts a benevolent attitude). Conversely, other scientists have looked favourably on the birth of an intelligence meant to overcome human capacities. In an AI system's ability to improve itself could lie the 'singularity' that will accelerate the development of science and technology, so as not only to solve current human problems (poverty, underdevelopment, etc.), but also to overcome the biological limits of human existence (illness, aging, etc.) and spread intelligence in the cosmos.[12]

[9] Bostrom (2014)

[10] Bostrom (2014). This possibility was anticipated by Turing ([1951] 1966).

[11] Parkin (2015).

[12] See Kurzweil (2005) and Tegmark (2017).

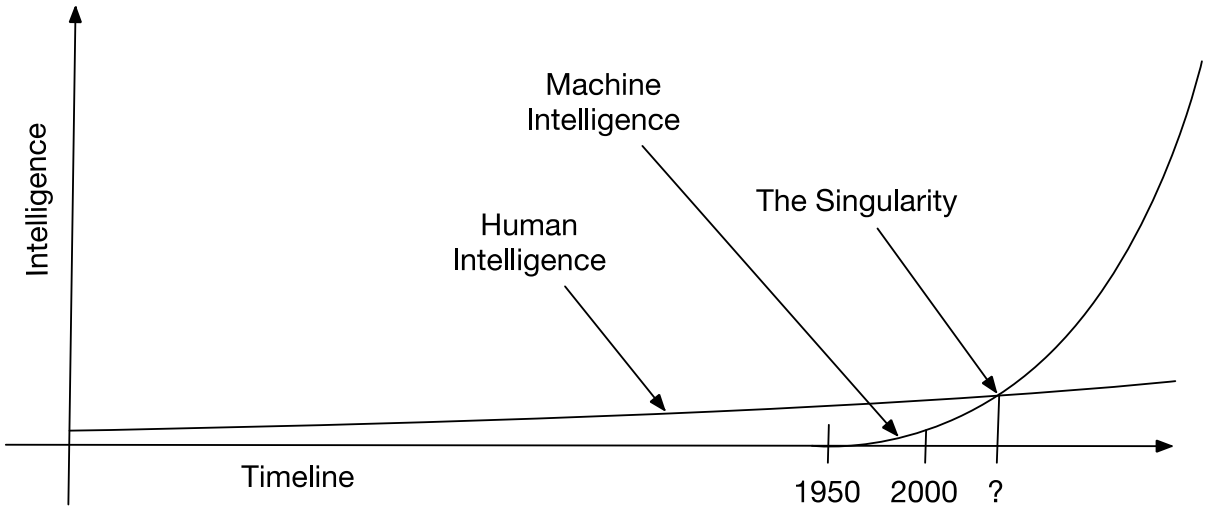


Figure 2 – General AI: The singularity

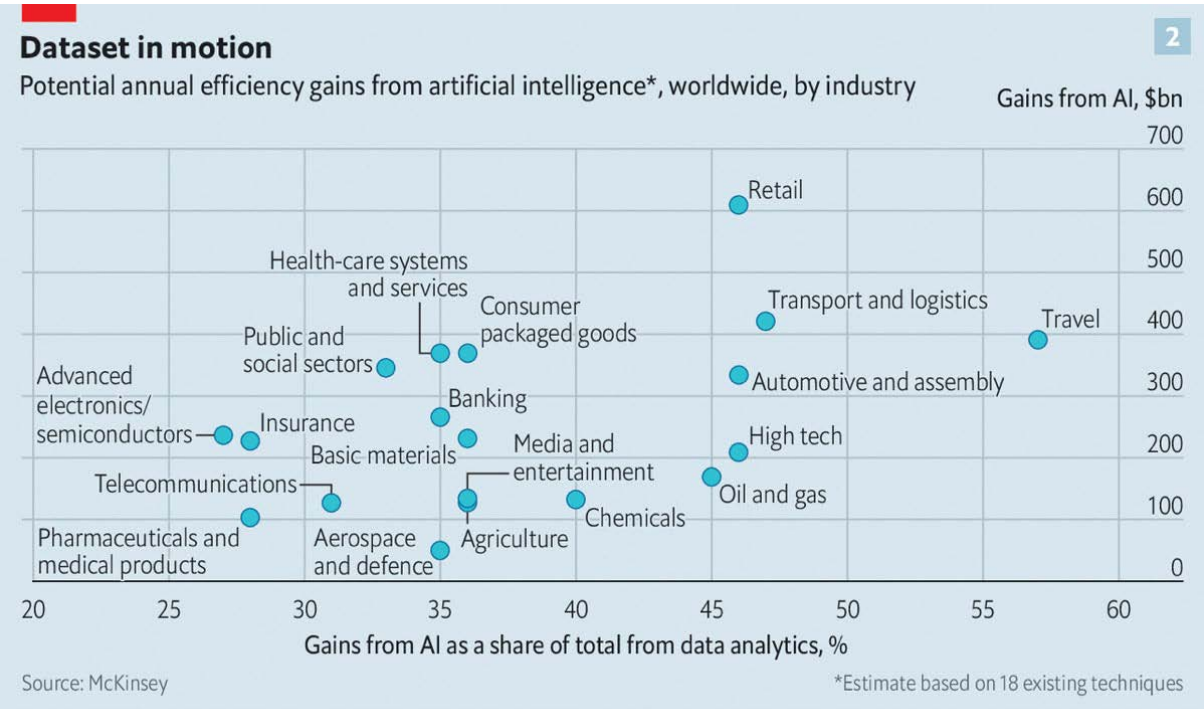
The risks related to the emergence of an 'artificial general intelligence' should not be underestimated: this is, on the contrary, a very serious problem that will pose challenges in the future. In fact, as much as scientists may disagree on whether and when 'artificial general intelligence,' will come into existence, most of them believe that this objective will be achieved within the end of this century.[13] In any case, it is too early to approach 'artificial general intelligence' at a policy level, since it lies decades ahead, and a broader experience with advanced AI is needed before we can understand both the extent and proximity of this risk, and the best ways to address it.

Conversely, 'artificial specialised intelligence' is already with us, and is quickly transforming economic, political, and social arrangements, as well as interactions between individuals and even their private lives. The increase in economic efficiency already is reality (see Figure 2), but AI provides further opportunities: economic, social, and cultural development; energy sustainability; better health care; and the spread of knowledge. In the very recent White Paper by the European Commission[14] it is indeed affirmed that AI.

> will change our lives by improving healthcare (e.g. making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans, and in many other ways that we can only begin to imagine.

[13] A poll among leading AI scientists can be found in Bostrom (2014).

[14] White Paper 'On artificial intelligence - A European approach to excellence and trust', Brussels, 19.2.2020 COM(2020)65 final.



The Economist

Figure 3 – Efficiency gains from AI

The opportunities offered by AI are accompanied by serious risks, including unemployment, inequality, discrimination, social exclusion, surveillance, and manipulation. It has indeed been claimed that AI should contribute to the realisation of individual and social interests, and that it should not be 'underused, thus creating opportunity costs, nor overused and misused, thus creating risks.'[15] In the just mentioned Commission's White paper, it is indeed observed that the deployment of AI

> entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.

Because the need has been recognised to counter these risks, while preserving scientific research and the beneficial uses of AI, a number of initiatives have been undertaken in order to design an ethical and legal framework for 'human-centred AI.' Already in 2016, the White House Office of Science and Technology Policy (OSTP), the European Parliament's Committee on Legal Affairs, and, in the UK, the House of Commons' Science and Technology Committee released their initial reports on how to prepare for the future of AI[16]. Multiple expert committees have subsequently produced reports and policy documents. Among them, the High-Level Expert Group on artificial intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), and the select committee on artificial intelligence of the United Kingdom (UK) House of Lords.[17]

The Commission's White Paper affirms that two parallel policy objectives should be pursued and synergistically integrated. On the one hand research and deployment of AI should be promoted, so

[15] Floridi et al (2018, 690).

[16] See Cath et al (2017).

[17] For a recent review of documents on AI ethics and policy, see Jobin (2019).

that the EU is competitive with the US and China. The policy framework setting out measures to align efforts at European, national and regional level should aim to mobilise resources

> to achieve an 'ecosystem of excellence' along the entire value chain, starting in research and innovation, and to create the right incentives to accelerate the adoption of solutions based on AI, including by small and medium-sized enterprises (SMEs)

On the other hand, the deployment of AI technologies should be consistent with the EU fundamental rights and social values. This requires measures to create an 'ecosystem of trust,' which should provide citizens with 'the confidence to take up AI applications' and 'companies and public organisations with the legal certainty to innovate using AI'. This ecosystem

> must ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk.

It is important to stress that the two objectives of excellence in research, innovation and implementation, and of consistency with individual rights and social values are compatible, but distinct. On the one hand the most advanced AI applications could be deployed to the detriment of citizens' rights and social values; on the other hand the effective protection of citizens' from the risks resulting from abuses AI does not provide in itself the incentives that are needed to stimulate research and innovation and promote beneficial uses. This report will argue that GDPR can contribute to address abuses of AI, and that it can be implemented in ways that do not hinder its beneficial uses. It will not address the industrial and other policies that are needed to ensure the EU competitiveness in the AI domain.

2.2.2. AI between logical models and machine learning

The huge success that AI has had in recent years is linked to a change in the leading paradigm in AI research and development. Until a few decades ago, it was generally assumed that in order to develop an intelligent system, humans had to provide a formal representation of the relevant knowledge (usually expressed through a combination of rules and concepts), coupled with algorithms making inferences out of such knowledge. Different logical formalisms (rule languages, classical logic, modal and descriptive logics, formal argumentation, etc.) and computable models for inferential processes (deductive, defeasible, inductive, probabilistic, case-based, etc.) have been developed and applied.[18]

The structure for expert systems is represented in Figure 4. Note that humans appear both as users of the system and as creators of the system's knowledge base (experts, possibly helped by knowledge engineers).

[18] Van Harmelen et al (2008).

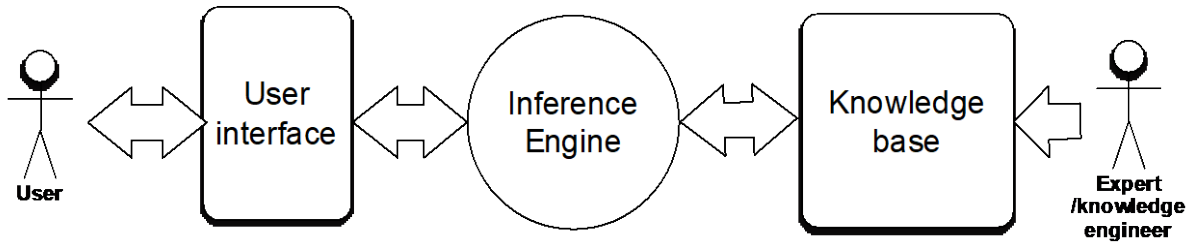


Figure 4 – Basic structure of expert systems

The theoretical results in knowledge representation and reasoning were not matched by disrupting, game-changing applications. Expert systems – i.e., computer systems including vast domain-specific knowledge bases, e.g., in medicine, law, or engineering, coupled with inferential engines – gave rise to high expectations about their ability to reason and answer users' queries. Unfortunately, such systems were often unsuccessful or only limitedly successful: they could only provide incomplete answers, were unable to address the peculiarities of individual cases, and required persistent and costly efforts to broaden and update their knowledge bases. In particular, expert-system developers had to face the so-called *knowledge representation bottleneck:* in order to build a successful application, the required information – including tacit and common-sense knowledge – had to be represented in advance using formalised languages. This proved to be very difficult and in many cases impractical or impossible.

In general, only in some restricted domains the logical models have led to successful application. In the legal domain, for example, logical models of great theoretical interest have been developed – dealing, for example, with arguments,[19] norms, and precedents[20] – and some expert systems have been successful in legal and administrative practice, in particular in dealing with tax and social security regulations. However, these studies and applications have not fundamentally transformed the legal system and the application of the law.

AI has made an impressive leap forward since it began to focus on the application of machine learning to mass amounts of data. This has led to a number of successful applications in many sectors – ranging from automated translation to industrial optimisation, marketing, robotic visions, movement control, etc. – and some of these applications already have substantial economic and social impacts. In machine learning approaches, machines are provided with learning methods, rather than, or in addition to, formalised knowledge. Using such methods, they can automatically learn how to effectively accomplish their tasks by extracting/inferring relevant information from their input data. As noted, and as Alan Turing already theorised in the 1950s, a machine that is able to learn will achieve its goals in ways that are not anticipated by its creators and trainers, and in some cases without them knowing the details of its inner workings.[21]

Even though the great success of machine learning has overshadowed the techniques for explicit and formalised knowledge representation, the latter remain highly significant. In fact, in many domains the explicit logical modelling of knowledge and reasoning can be *complementary* to machine learning. Logical models can explain the functioning of machine learning systems, check and govern their behaviour according to normative standards (including ethical principles and legal norms), validate their results, and develop the logical implications of such results according to conceptual knowledge and scientific theories. In the AI community the need to combine logical modelling and machine leaning is generally recognised, though different views exist on how to

[19] Prakken, and Sartor (2015).

[20] Ashley (2017).

[21] Turing ([1951] 1996)

achieve this goal, and on the aspects to be covered by the two approaches (for a discussion on the limits of machine learning, see recently Marcus and Davis 2019).

2.2.3. Approaches to learning

Three main approaches to machine learning are usually distinguished: supervised learning, reinforcement learning and unsupervised learning.

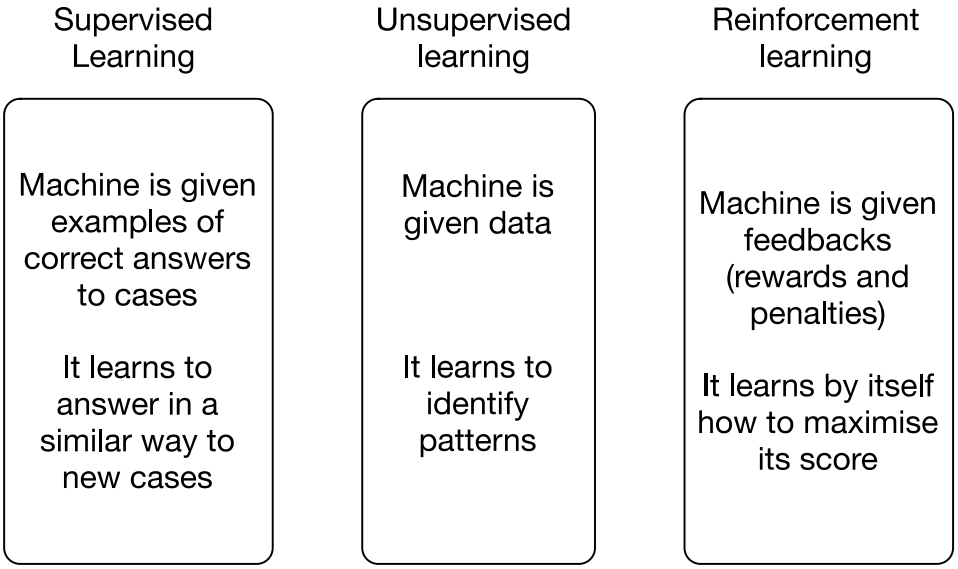


Figure 5 – Kinds of learning

Supervised learning is currently the most popular approach. In this case the machine learns through 'supervision' or 'teaching': it is given in advance a training set, i.e., a large set of (probably) correct answers to the system's task. More exactly the system is provided with a set of pairs, each linking the description of a case to the correct response for that case. Here are some examples: in systems designed to recognise objects (e.g. animals) in pictures, each picture in the training set is tagged with the name of the kind of object it contains (e.g., cat, dog, rabbit, etc.); in systems for automated translation, each (fragment of) a document in the source language is linked to its translation in the target language; in systems for personnel selection, the description of each past applicants (age, experience, studies, etc.) is linked to whether the application was successful (or to an indicator of the work performance for appointed candidates); in clinical decision support systems, each patient's symptoms and diagnostic tests is linked to the patient's pathologies; in recommendation systems, each consumer's features and behaviour is linked to the purchased objects; in systems for assessing loan applications, each record of a previous application is linked to whether the application was accepted (or, for successful applications, to the compliant or non-compliant behaviour of the borrower). As these examples show, the training of a system does not always require a human teacher tasked with providing correct answers to the system. In many case, the training set can be side-product of human activities (purchasing, hiring, lending, tagging, etc.), as is obtained by recording the human choices pertaining to such activities. In some cases the training set can even be gathered 'from the wild' consisting in data which is available on the open web. For instance, manually tagged images or faces, available on social networks, can be scraped and used for training automated classifiers.

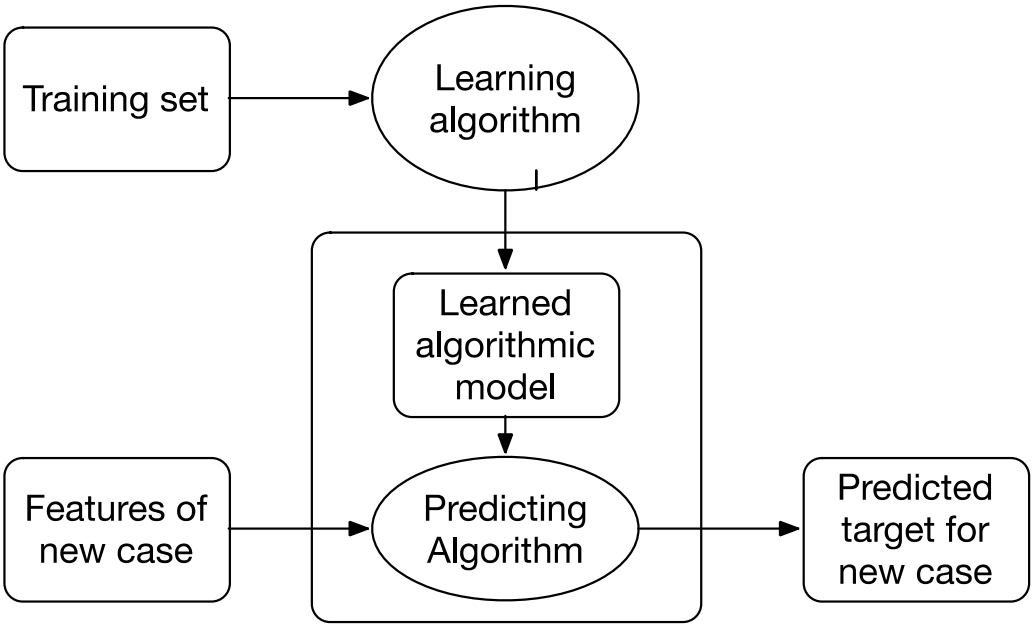


Figure 6 – Supervised learning

The learning algorithm of the system (its trainer), uses the training set to build an algorithmic model: a neural network, a decision tree, a set of rules, etc. The algorithmic model is meant to capture the relevant knowledge originally embedded in the training set, namely the correlations between cases and responses. This model is then used, by a predicting algorithm, to provide hopefully correct responses to new cases, by mimicking the correlations in the training set. If the examples in the training set that come closest to a new case (with regard to relevant features) are linked to a certain answer, the same answer will be proposed for the new case. For instance if the pictures that are most similar to a new input were tagged as cats, also the new input will also be tagged in the same way; if past applicants whose characteristic best match those of the new applicant were linked to rejection, the system will propose to reject also the new applicant; if the past workers who come closest to the new applicant performed well (or poorly), the systems will predict that also the applicant will perform likewise.

The answers by learning systems are usually called 'predictions'. However, often the context of the system's use often determines whether its proposals are be interpreted as forecasts, or rather as a suggestion to the system's user. For instance, a system's 'prediction' that a person's application for bail or parole will be accepted can be viewed by the defendant (and his or her lawyer) as a prediction of what the judge will do, and by the judge as a suggestion guiding her decision (assuming that she prefers not to depart from previous practice). The same applies to a system's prediction that a loan or a social entitlement will be granted.

There is also an important distinction to be drawn concerning whether the 'correct' answers in a training set are provided by the past choices by human 'experts' or rather by the factual consequences of such choices. Compare, for instance, a system whose training set consists of past loan applications linked to the corresponding lending decisions, and a system whose training set consists of successful applications linked to the outcome of the loan (repayment or non-payment). Similarly, compare a system whose training set consists of parole applications linked to judges' decisions on such application with a system whose training set consists of judicial decisions on parole applications linked to the subsequent behaviour of the applicant. In the first case, the system will learn to predict the decisions that human decision-makers (bank managers, or judges) would have made under the same circumstances. In the second case, the system will predict how a certain choice would affect the goals being pursued (preventing non-payments, preventing recidivism). In the first case the system would reproduce the virtues – accuracy, impartiality, farness – but also the

vices – carelessness, partiality, unfairness – of the humans it is imitating. In the second case it would more objectively approximate the intended outcomes.

As a simple example of supervised learning, Figure 7, shows a (very small) training set concerning bail decisions along with the decision tree that can be learned on the basis of that training set. The decision tree captures the information in the training set through a combination of tests, to be performed sequentially. The first test concerns whether the defendant was involved in a drug related offence. If the answer is positive, we have reached the bottom of the tree with the conclusion that bail is denied. If the answer is negative, we move to the second test, on whether the defendant used a weapon, and so on. Notice that the decision tree does not include information concerning the kind of injury, since all outcomes can be explained without reference to that information. This shows how the system's model does not merely replicate the training set; it involves generalisation: it assumes that certain combination of predictors are sufficient to determine the outcomes, other predictors being irrelevant.

Predictors

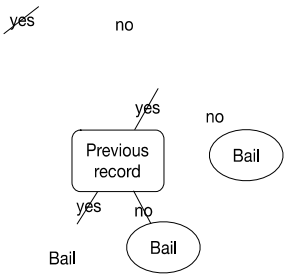


Figure 7 – Training set and decision tree for bail decisions

In this example we can distinguish the elements in Figure 6. The table in Figure 7 is the *training set*. The software that constructs the decision tree, is the *learning algorithm*. The decision tree itself, as shown in Figure 7 is the *algorithmic model,* which codes the logic of the human decisions in the training set. The software that processes new cases, using the decision tree, and makes predictions based on their features of such cases, is the *predicting algorithm*. In this example, as noted above, the decision tree reflects the attitudes of the decision-makers whose decisions are in the training set: it reproduces their virtues and biases.

For instance, according to the decision tree, the fact that the accuse concerns a drug-related offence is sufficient for bail to be denied. We may wonder whether this is a fair criterion for assessing bail requests. Note also that the decision tree (the algorithmic model) also provides answers for cases that do not fit exactly any example in the training set. For instance, no example in the training set concerns a drug-related offence with no weapon and no previous record. However, the decision tree provides an answer also for this case: there should be no bail, as this is what happens in all drug-related cases in the training set.

As another simplified example of supervised machine learning consider the training set and the rules in figure 7. In this case too, the learning algorithm, as applied to this very small set of past decisions, delivers questionable generalisation, such as the prediction that young age would always lead to a rejection of the loan applications and that middle age would always lead to acceptance. Usually, in order to give reliable prediction, a training set must include a vast number of examples, each described through a large set of predictors.

Reinforcement learning is similar to supervised learning, as both involve training by way of examples. However, in the case of reinforcement learning the systems learns from the outcomes of its own action, namely, through the rewards or penalties (e.g., points gained or lost) that are linked to the outcomes of such actions. For instance, in case of a system learning how to play a game, rewards

may be linked to victories and penalties to defeats; in a system learning to make investments, rewards may be linked to financial gains and penalties to losses; in a system learning to target ads effectively, rewards may be linked to users' clicks, etc. In all these cases, the system observes the outcomes of its actions, and it self-administers the corresponding rewards or penalties. Being geared towards maximising its score (its utility), the system will learn to achieve outcomes leading to rewards (victories, gains, clicks), and to prevent outcomes leading to penalties. With regard to reinforcement learning too, we can distinguish the *learner* (the algorithm that learns how to act successfully, based on the outcomes of previous actions by the system) and the learned *model* (the output of the learner, which determines the system's new actions).

In *unsupervised learning*, finally, AI systems learn without receiving external instructions, either in advance or as feedback, about what is right or wrong. The techniques for unsupervised learning are used in particular, for clustering, i.e., for grouping the set of items that present relevant similarities or connections (e.g., documents that pertain to the same topic, people sharing relevant characteristics, or terms playing the same conceptual roles in texts). For instance, in a set of cases concerning bail or parole, we may observe that injuries are usually connected with drugs (not with weapons as expected), or that people having prior record are those who are related to weapon. These clusters might turn out to be informative to ground bail or parole policies.

2.2.4. Neural networks and deep learning

Many techniques have been deployed in machine learning: decision trees, statistical regression, support vector machine, evolutionary algorithms, methods for reinforcement learning, etc. Recently, deep learning based on many-layered neural networks has been very successfully deployed especially, but not exclusively, where patterns have to be recognised and linked to classifications and decisions (e.g., in detecting objects in images, recognising sounds and their sources, making medical diagnosis, translating texts, choosing strategies in games, etc.). Neural networks are composed of a set of nodes, called neurons, arranged in multiple layers and connected by links. They are so-called, since they reproduce some aspects of the human nervous system, which indeed consists of interconnected specialised cells, the biological neurons, which receive and transmit information. Neural networks were indeed developed under the assumption that artificial intelligence could be achieved by reproducing the human brain, rather than by modelling human reasoning, i.e., that artificial reasoning would naturally emerge out of an artificial brain (though we may wonder to what extent artificial neural networks and human brains really share the similar structures and processes

Each neuron receives signals (numbers) from connected neurons or from the outside, and these signals are magnified or diminished as they cross incoming links, according to the weights of the latter. The neuron applies some calculations to the input it receives, and if the result reaches the neuron's threshold, the neuron becomes active sending signals to the connected neurons or outside of the network. The activation starts from nodes receiving external inputs and spreads through the network. The training of the network takes place by telling the network whether its answers (its outputs) are right or wrong. If an answer by the network is wrong, the learning algorithm updates the network – i.e., it adjusts the weights of the connections – so that next time the network is presented with that input, it will give the correct answer. Figure 8 shows a simplified representation of a multi-layered neural network (real networks may have many more layers of neurons) for face recognition, where the initial layers learn very generic aspects of the images (border, colours, shapes, etc.) while higher layers engage with the elements of human faces.

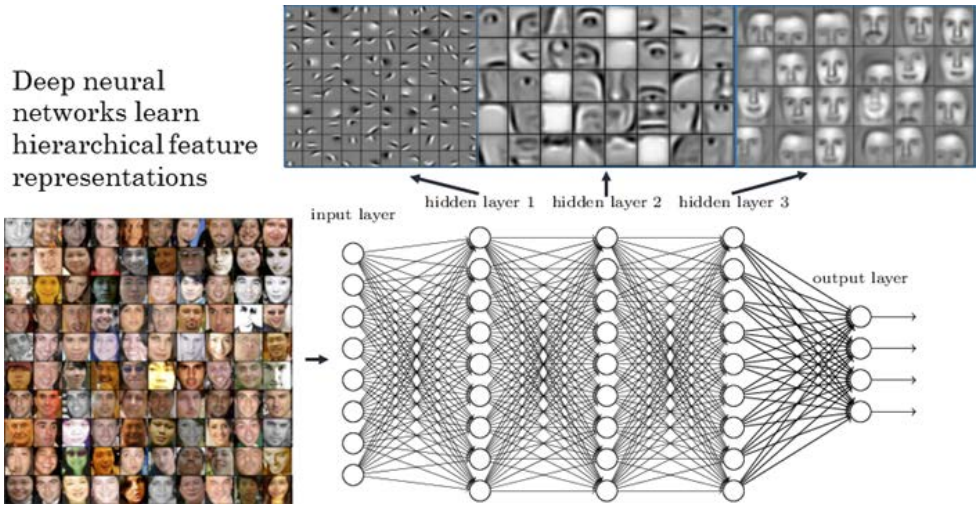


Figure 8 – Multilayered (deep) neural network for face recognition

In the case of the neural network, the learning algorithm modifies the network until it achieves the desired performance level, while the outcome of the learning – algorithmic model – is the network in its final configuration.

As previously noted, the learning algorithm is able to modify the neural network (the weights in connections and neurons) so that the network is able to provide the most appropriate answers. Under the supervised learning approach, the trained network will reproduce the behaviour in the training set; under the reinforcement learning approach, the network will adopt the behaviour that maximises its score (e.g. the reward points linked to gains in investments or to victories in games).

2.2.5. Explicability

Different machine learning approaches differ in their ability to provide explanations. For instance, the outcome of a decision tree can be explained through the sequence of tests leading to that outcome. In our example, if bail is refused after testing No for Drug, Yes for Weapons and Yes for Previous record, an explanation is provided by a corresponding rule: *if No Drug and Weapons and Previous Record, then No Bail*.

Unlike a decision tree, a neural network does not provide explanations of its outcomes. It is possible to determine how a certain output has resulted from the network's activation, and how that activation, in response to a given input, was determined by the connections between neurons (and by the weights assigned to such connections as a result of the network's training) and by the mathematical functions governing each neuron. However, this information does not show a rationale that is meaningful to humans: it does not tell us why a certain response was given.

Many approaches exist to providing explanations of the behaviour of neural networks and other opaque systems (also called 'black boxes'). Some of these approaches look into the system to be explained, and build explanations accordingly (e.g., looking at the outcomes of the network's different layers, as in the example in Figure 8). Other approaches build explanations on the basis of the network's external behaviour: they only consider the relation between the inputs provided by the network and the outcomes it delivers, and build arguments or other explanations accordingly. However, advancements of human-understandable explanation of neural networks have so far been quite limited still.[22] Unfortunately, in many domains, the systems whose functioning is less explicable provide higher performance. Thus, comparative advantages in performance and in explicability may have to be balanced, in order to determine what approach should be adopted in

[22] Guidotti et al (2018).

a machine learning system. The best balance also depends on the domain in which the system is used and on the importance of the interests that are affected. When public action is involved and key human interests are at stake (e.g., as in judicial decisions) explanation is paramount.

Even when a system can only be viewed as a black box, however, some critical analyses of its behaviour are still possible. Through sensitivity analysis – i.e., by systematically checking whether the output changes if the value of certain input features is modified, leaving all other features unchanged – we can understand what features determine the system's output. For instance, by checking whether the prediction of a system meant to assess creditworthiness changes if we modify the place of birth or residence of the applicant, we can determine whether this input feature is relevant to the system's output. Consequently, we may wonder whether the system unduly discriminated people depending on their ethnicity or social status, which may be linked to place of birth or residence.

2.3. AI and (personal) data

The following sections will consider the interaction between AI and big data. First, the use of big data for AI-bases predictions and assessments will be introduced. The ensuing risks and opportunities will be analysed. Then, decision-making concerning individuals will be addressed, with a focus on fairness and non-discrimination. Finally, the issues concerning profiling, influence and manipulation will be analysed, including those related to pervasive surveillance by private actors and governments.

2.3.1. Data for automated predictions and assessments

To predict a certain outcome in a new case means to jump from certain known features of that case, the so-called *predictors* (also called independent variables, or features), to an unknown feature of that case, the *target* to be predicted (also called dependent variable, or label). This forecast is based of models that capture general aspects of the contexts being considered, on the basis of which it is possible to connect the values of predictors and targets. For instance a model in the medical domain may connect symptoms to diseases, a psychometric model may connect online behaviour (e.g., friends, posts and likes on a social network) to psychological attitudes; etc.

Such models may be created by humans (who formulate the rules and concepts in the model), even when the application of the models is delegated to a machine (as in rule-based expert systems). However, as noted in Section 2.2.2, the construction (learning) of the models, and not only their application is increasingly entrusted to machines. In the machine learning approach, machines discover the probabilistic correlations between predictors and targets, and then apply these correlations to make predictions in new cases. Thanks to the combination of AI techniques, vast masses of data, and computational power, it has become possible to base automated predictions and assessments on a much larger sets of examples, taking into account a much larger set of features of each of them, so as to achieve useful level of accuracy in many domains.

For instance, targeted advertising may be based on records linking the characteristics and behaviour of consumers (gender, age, social background, purchase history, web browsing, etc.) to their responses to ads. Similarly, the assessment of job applications may be based on records linking characteristics of previous workers (education, employment history, jobs, aptitude tests, etc.), to their work performance; the prediction of the likelihoods of recidivism by a particular offender may be based on records combining characteristics of past offenders (education, employment history, family status, criminal record, psychological tests, etc.) with data or assessments on their recidivism; the prediction of a prospective borrower's creditworthiness may be based on records linking the characteristics of past borrowers to data or assessments about their creditworthiness; the diagnosis of diseases or the suggestion of personalised medical treatments may be based on the records of

past patients, linking their characteristics and medical tests to subsequent medical conditions and treatments.

As a result of the need to learn by analysing vast amount of data, AI has become hungry for data, and this hunger has spurred data collection, in a self-reinforcing spiral.[23] Thus, the development of AI systems based on machine learning presupposes and fosters the creation of vast data sets, i.e., *big data*[24].

The collection of data is facilitated by the availability of electronic data as a by-product of using any kind of ICT system. Indeed, a massive digitisation has preceded most AI applications, resulting from the fact that data flows are produced in all domains where computing is deployed.[9] For instance, huge amounts of data are collected every second by computers that execute economic transactions (as in e-commerce)[10], by sensors monitoring and providing input to physical objects (e.g., vehicles or smart home devices), by the workflows generated by economic and governmental activities (e.g., banking, transportation, or taxation, etc.); by surveillance devices (e.g. traffic cameras, or access control systems); and systems supporting non-market activities (e.g. internet access, searching, or social networking).

In recent years, these data flows have been integrated into a global interconnected data-processing infrastructure, centred on, but not limited to, the Internet. This infrastructure constitutes a universal medium for communicating, accessing data, and delivering any kind of private and public services. It enables citizens to shop, use banking and other services, pay taxes, get government benefits and entitlements, access information and knowledge, and build social connections. Algorithms – often powered by AI – mediate citizens' access to content and services, selecting information and opportunities for them, while at the same time recording any activity. Today, this global interconnected data-processing infrastructure seems to include about 30 billion devices – computers, smartphones, industrial machines, cameras, etc. – which generate masses of electronic data (see Figure 9).

[23] Cristianini (2016).

[24] Mayer-Schönberger and Cukier (2013).

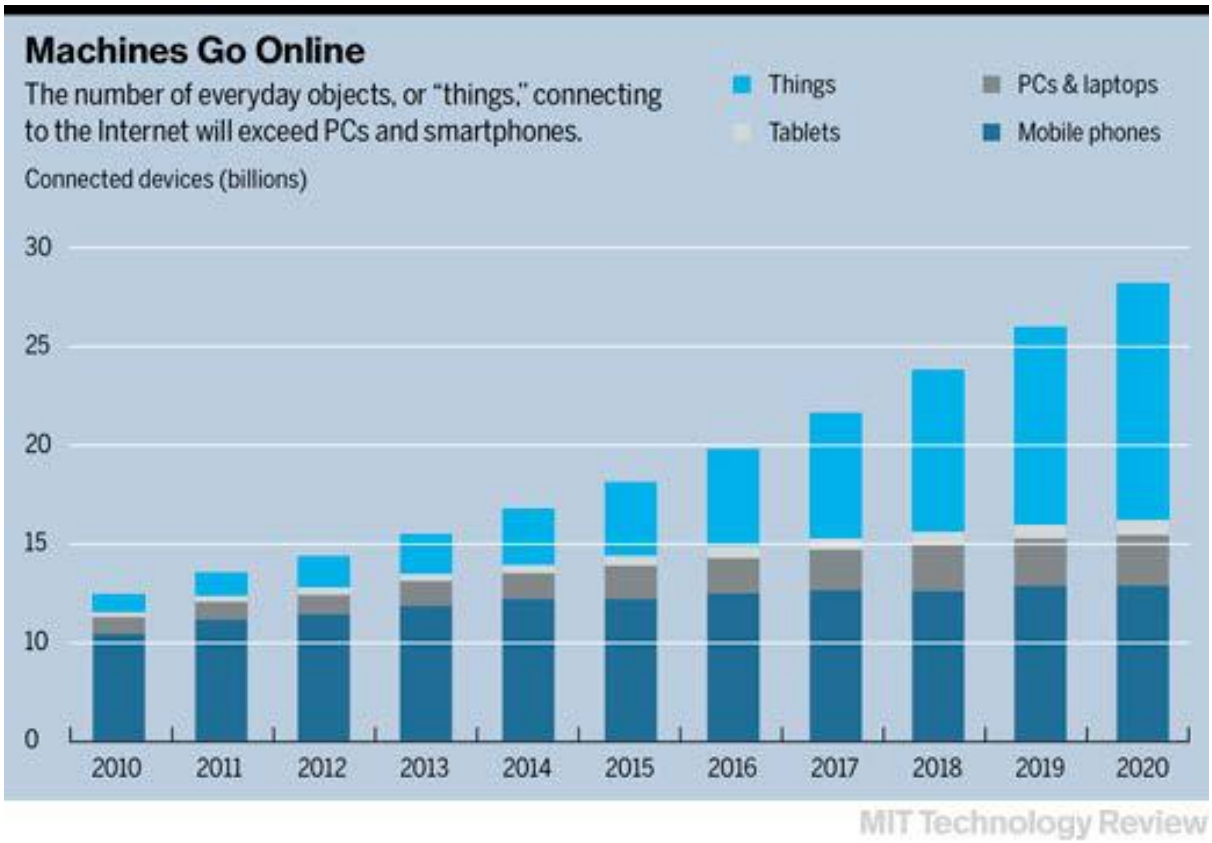


Figure 9 – Number of connected devices

Figure 10 provides a comparative overview of what takes place online every minute.

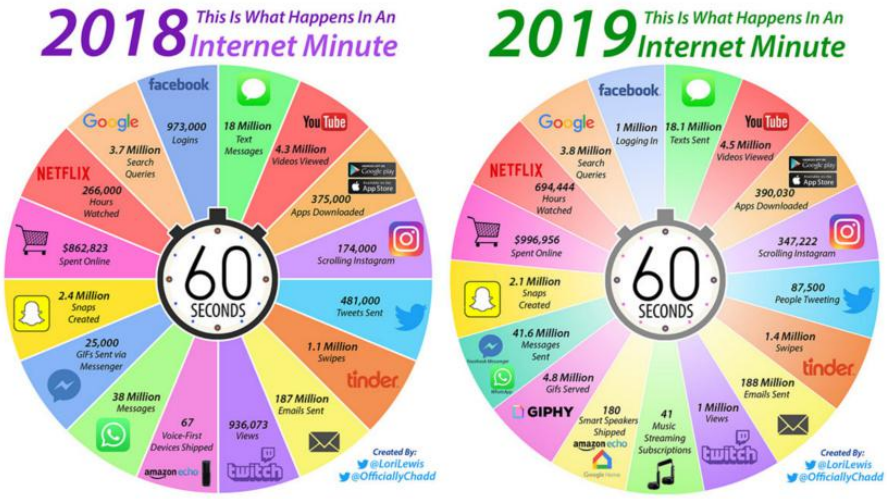


Figure 10 – Data collected in a minute of online activity worldwide

AI's hunger for data concerns any kind of information: from meteorological data, to environmental ones, to those concerning industrial processes. Figure 4 gives an idea of the growth of data creation.

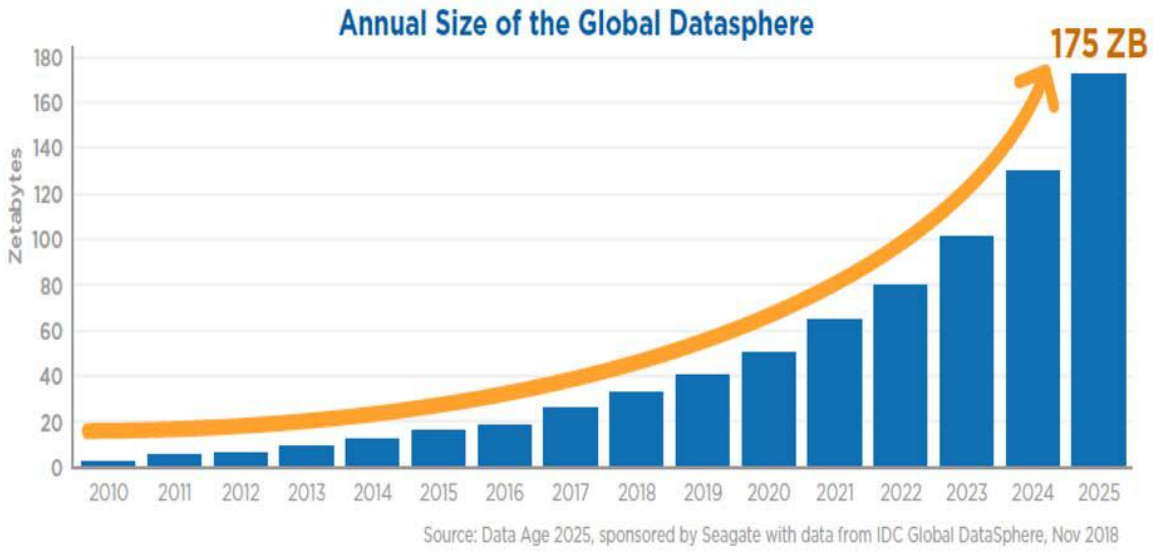


Figure 11 – Growth of global data

2.3.2. AI and big data: risks and opportunities

The integration of AI and big fata technologies into the global data-processing infrastructure can deliver a lot of benefits: better access to information; generation and distribution of knowledge across the globe; cost savings, greater productivity, and value creation; new creative and well-paying jobs; individualised private and public services; environmentally-friendly management of utilities and logistics; novel information and consulting services; support for transparency; remedies against biases and discriminations, etc. Great advances are enabled in many domains: scientists can discover correlations, formulate hypotheses and develop evidence-based models; doctors can provide better diagnosis and personalised and targeted therapies; firms can anticipate market trends and make more efficient decisions; consumers can make more informed choices and obtain personalised services; public authorities can anticipate risks, prevent damages, optimise the management of public goods (such as the environment) and coordinate citizens' actions (e.g., the management of traffic, energy consumption, and utilities). And more good can come in the future. As has been argued by Ray Kurzweil, an inventor, futurist, and director of engineering at Google:

> Through [information] technologies we can address the grand challenges of humanity, such as maintaining a healthy environment, providing the resources for a growing population (including energy, food, and water), overcoming disease, vastly extending human longevity, and eliminating poverty. It is only by extending ourselves with intelligent technology that we can deal with the scale of complexity needed. [25]

In some cases, AI can fully replace human activities (e.g., in driverless vehicles, cleaning robots, and certain planning and scheduling tasks in logistics). In many cases it rather complements human capacities: it enhances the human ability to know and act, it supports creativity and invention.[26] Thanks to AI, it may be possible to achieve a new cooperation between humans and machines, which overcomes the classical model in which machines only performed routine and repetitive tasks. This integration was already predicted in the early 1960s' by JK Licklider, a scientist who played a key role in the development of the Internet. He argued that in the future, cooperation between human and computer would include creative activities, i.e., 'making decisions and controlling

[25] Kurzweil (2012).

[26] McAfee, and Brynjolfsson (2019).

complex situations without inflexible dependence on predetermined programs.[27] Today, it is indeed possible to integrate humans and machines in new ways that not only exploit synergies, but may also preserve and enhance human initiative and work satisfaction.[28]

However, the development of AI and its convergence with big data also leads to serious risks for individuals, for groups, and for the whole of society. For one thing, AI can eliminate or devalue the jobs of those who can be replaced by machines: many risk losing the 'race against the machine',[29] and therefore being excluded from or marginalised in the job market. This may lead to poverty and social exclusion, unless appropriate remedies are introduced (consider, for instance, the future impact of autonomous vehicles on taxi and truck drivers, or the impact of smart chatbots on call-centres workers).

Moreover, by enabling big tech companies to make huge profits with a limited workforce, AI contributes to concentrating wealth in those who invest in such companies or provide them with high-level expertise. This trend favours economic models in which '*the winner takes all*'. Within companies, monopoly positions tend to prevail, thanks to the network effect (users' preference for larger networks), coupled with economies of scale (enabled by automation) and exclusive or preferential access to data and technologies. Within workers, financial and other benefits, as well as work satisfaction, tend to accrue only to those who can engage in high-level functions that have not yet been automated. To address the adverse impact of AI, appropriate political and social strategies must ensure that everyone will benefit from AI, thanks to workers' training, human-machine interactions focused on engagement and creativity, broader access to data and technologies, wealth redistribution policies.

There is also a need to counter the new opportunities for illegal activities offered by AI and big data. In particular, AI and big data systems can fall subject to cyberattacks (designed to disable critical infrastructure, or steal or rig vast data sets, etc.), and they can even be used to commit crimes (e.g., autonomous vehicles can be used for killing or terrorist attacks, and intelligent algorithms can be used for fraud or other financial crimes).[30] Even beyond the domain of outright illegal activities, the power of AI can be used to purse economic interests in ways that are harmful to individuals and society: users, consumers, and workers can be subject to pervasive surveillance, controlled in their access to information and opportunities, manipulated in their choices.

Certain abuses may be incentivised by the fact that many tech companies – such as major platforms hosting user-generated content – operate in two- or many-sided markets. Their main services (search, social network management, access to content, etc.) are offered to individual consumers, but the revenue stream comes from advertisers, influencers, and opinion-makers (e.g., in political campaigns). This means not only that any information that is useful for targeted advertising will be collected and used for this purpose, but also that platforms will employ any means to capture users, so that they can be exposed to ads and attempts at persuasion. This may lead not only to a massive collection of personal data about individuals, to the detriment of privacy, but also to a pervasive influence on their behaviour, to the detriment of both individual autonomy and collective interests. Additionally, profit-driven algorithms can combine in order to advance anticompetitive strategies, to the detriment not only competitors but also of consumers. AI also can contribute to polarisation and fragmentation in the public sphere,[31] and to the proliferation of sensational and fake news,

[27] Licklider (1960).

[28] McAfee and Brynjolfsson (2019), Mindell (2015).

[29] Brynjolfsson and McAfee (2011).

[30] Bhuta et al (2015).

[31] Sunstein (2007).

when used to capture users by exposing them to information they may like, or which accords with their preferences, thereby exploiting their confirmation biases.[32]

Just as AI can be misused by economic actors, it can also be misused by the public section. Governments have many opportunities to use AI for legitimate political and administrative purposes (e.g., efficiency, cost savings, improved services), but they may also employ it to anticipate and control citizens' behaviour in ways that restrict individual liberties and interfere with the democratic process.

2.3.3. AI in decision-making concerning individuals: fairness and discrimination

The combination of AI and big data enables automated decision-making even in domains that require complex choices, based on multiple factors, and on non-predefined criteria. In recent years, a wide debate has taken place on the prospects and risks of algorithmic assessments and decisions concerning individuals

Some scholars have observed that in many domains automated predictions and decisions are not only cheaper, but also more precise and impartial than human ones. AI systems can avoid the typical fallacies of human psychology (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc.), and the widespread human inability to process statistical data,[33] as well as typical human prejudice (concerning, e.g., ethnicity, gender, or social background). In many assessments and decisions – on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism – algorithmic systems have often performed better, according to usual standards, than human experts.[34]

Others have underscored the possibility that algorithmic decisions may be mistaken or discriminatory. Only in rare cases will algorithms engage in explicit unlawful discrimination, so-called disparate treatment, basing their outcome on prohibited features (predictors) such as race, ethnicity or gender. More often a system's outcome will be discriminatory due to its disparate impact, i.e., since it disproportionately affects certain groups, without an acceptable rationale.

As noted in Section 2.2.3, systems based on supervised learning may be trained on past human judgements and may therefore reproduce the strengths and weaknesses of the humans who made these judgements, including their propensities to error and prejudice. For example, a recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work. If past decisions were influenced by prejudice, the system will reproduce the same logic.[35] Prejudice baked into training sets may persist even if the inputs (the predictors) to the automated systems do not include forbidden discriminatory features, such as ethnicity or gender. This may happen whenever a correlation exists between discriminatory features and some predictors considered by the system. Assume, for instance, that a prejudiced human resources manager did not in the past hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods. A training set of decisions by that manager will teach the systems not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity.

[32] Pariser (2011).

[33] Kahneman (2011).

[34] Kahneman (2011, Ch. 21), Kleinberg et al (2019).

[35] Kleinberg et all (2019).

In other cases, a training set may be biased against a certain group, since the achievement of the outcome being predicted (e.g., job performance) is approximated through a proxy that has a disparate impact on that group. Assume, for instance, that the future performance of employees (the target of interest in job hiring) is only measured by the number of hours worked in the office. This outcome criterion will lead to past hiring of women – who usually work for fewer hours than men, having to cope with heavier family burdens – being considered less successful than the hiring of men; based on this correlation (as measured on the basis of the biased proxy), the systems will predict a poorer performance of female applicants.

In other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors. A system may perform unfairly, since it uses a favourable predictor (input feature) that only applies to members of a certain group (e.g., the fact of having attended a socially selective high-education institution). Unfairness may also result from taking biased human judgements as predictors (e.g., recommendation letters).

Finally, unfairness may derive from a data set that does reflect the statistical composition of the population. Assume for instance that in applications for bail or parole, previous criminal record plays a role, and that members of a certain groups are subject to stricter controls, so that their criminal activity is more often detected and acted upon. This would entail that members of that group will generally receive a less favourable assessment than members of other groups having behaved in the same ways.

Members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set, since this will reduce the accuracy of predictions for that group (e.g., consider the case of a firm that has appointed few women in the past and which uses its records of past hiring as its training set).

It has also been observed that it is difficult to challenge the unfairness of automated decision-making. Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operation, giving rise to additional costs and uncertainties. In fact, the predictions of machine-learning systems are based on statistical correlations, against which it may be difficult to argue on this basis of individual circumstances, even when exceptions would be justified. Here is the perspective of Cathy O'Neil, a machine-learning expert who has become a critic of the abuses of automation:

> An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, 'suggestive' countervailing evidence simply won't cut it. The case must be ironclad. The human victims of WMDs, we'll see time and again, are held to a far higher standard of evidence than the algorithms themselves.[36]

These criticisms have been countered by observing that algorithmic systems, even when based on machine learning, are more controllable than human decision-makers, their faults can be identified with precision, and they can be improved and engineered to prevent unfair outcomes.

> [W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central trade-

[36] O'Neil (2016)

> offs among competing values. Algorithms are not only a threat to be regulated; with the right safeguards in place, they have the potential to be a positive force for equity.[37]

In conclusion, it seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making. The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker. In many cases, the best solution consists in integrating human and automated judgements, by enabling the affected individuals to request a human review of an automated decision as well as by favouring transparency and developing methods and technologies that enable human experts to analyse and review automated decision-making. In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, such as medical diagnosis, financial investment, the granting of loans, etc. The future challenge will consist in finding the best combination between human and automated intelligence, taking into account the capacities and the limitations of both.

2.3.4. Profiling, influence and manipulation

The use of automated assessment systems may be problematic where their performance is not worse, or even is better, than what humans would do. This is due to the fact that automation diminishes the costs of collecting information on individuals, storing this information and process it in order to evaluate individuals and make choices accordingly. Thus, automation paves the way for much more persistent and pervasive mechanisms for assessment and control.

In general, thanks to AI, all kind of personal data can be used to analyse, forecast and influence human behaviour, an opportunity that transforms them into valuable commodities. Information that was not collected or was discarded as worthless 'data exhaust' – e.g., trails of online activities – has now become a prized resource.

Through AI and big data technologies – in combination with the panoply of sensor that increasingly trace any human activity – individuals can be subject to surveillance and influence in many more cases and contexts, on the basis of a broader set of personal characteristics (ranging from economic conditions to health situation, place of residence, personal life choices and events, online and offline behaviour, etc.). By correlating data about individuals to corresponding classifications and predictions, AI increases the potential for *profiling*, namely, for inferring information about individuals or groups, and adopting assessments and decisions on that basis. The term 'profile' derives from the Italian 'profilo,' from "profilare," originally meaning to draw a line, especially the contour of an object: that is precisely the idea behind profiling through data processing, which means to expand the available data of individuals of groups, so as to sketch – describe or anticipate – their traits and propensities.

A profiling system establishes (predicts) that individuals having certain features F_1, also have a certain likelihood of possessing certain additional features F_2. For instance, assume that the system establishes (predicts) that those having a genetic patterns have the tendency to develop a higher than average chance to develop cancer, or that those having a certain education and job history or ethnicity have a certain higher-than-average likelihood to default of their debts). Then we may say that this system has profiled the group of the individuals possessing features F_1: it has added to the description (the profile) of these group a new segment, namely, the likelihood of possessing the additional features F_2. If the system is then given the information that a specific individual has features F_1, then the system can infer that it likely that this individual also has feature F_2. This may lead to the individual being treated accordingly, in a beneficial or a detrimental way. For instance, in the case in which the inferred feature of an individual is his or her higher susceptibility to cancer,

[37] Kleinberg, Ludwig, Mullainathan, and Sunstein (2018, 113).

the system's indication may provide the basis for preventive therapies and tests, or rather for a raise in the insurance premium.

The information so inferred may also be conditional, that is, it may consist in the propensity to react in a certain way to given inputs. For instance, it may consist in in the propensity to respond to a therapy with improved medical condition, or in the propensity to respond to a certain kind of ad or to a certain price variation with a certain purchasing behaviour, or in the propensity to respond a certain kind of message with a change in mood or preference (e.g., relatively to political choices). When that is the case, profiling potentially leads to influence and manipulation.

Assume, too, that the system connects certain values for input features (e.g., having a certain age, gender, social status, personality type, etc.) to the propensity to react to a certain message (e.g., a targeted ad) with a certain response (e.g., buying a certain product). Assume also that the system is told that a particular individual has these values (he is a young male, working class, extrovert, etc.). Then the system would know that by administering to the individual that message, the individual can probably be induced to deliver the response.

The notion of profiling just presented corresponds to this more elaborate definition:

> Profiling is a technique of (partly) automated processing of personal and/or non-personal data, aimed at producing knowledge by inferring correlations from data in the form of profiles that can subsequently be applied as a basis for decision-making. A profile is a set of correlated data that represents a (individual or collective) subject. Constructing profiles is the process of discovering unknown patterns between data in large data sets that can be used to create profiles. Applying profiles is the process of identifying and representing a specific individual or group as fitting a profile and of taking some form of decision based on this identification and representation.[38]

The notion of profiling in the GDPR only covers assessments or decisions concerning individuals, based on personal data, excluding the mere construction of group profiles:

> 'profiling'[…] consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

Even when an automated assessment and decision-making system – a profile-based system – is unbiased, and meant to serve beneficial purposes, it may negatively affect the individuals concerned. Those who are subject to pervasive surveillance, persistent assessments and insistent influence come under heavy psychological pressure that affects their personal autonomy, and they are susceptible to deception, manipulation and exploitation in multiple ways.

2.3.5. The dangers of profiling: the case of Cambridge Analytica

The dangers involved in profiling have emerged with clarity in the Cambridge Analytica case, concerning attempts at influencing voting behaviour – in the United States' 2016 election and possibly also in the Brexit referendum – based of massive processing of personal data. Figure 12 shows the main steps concerning Cambridge Analytica involvement in the US elections.

[38] Bosco et al (2015); see also Hildebrandt, M. (2009).

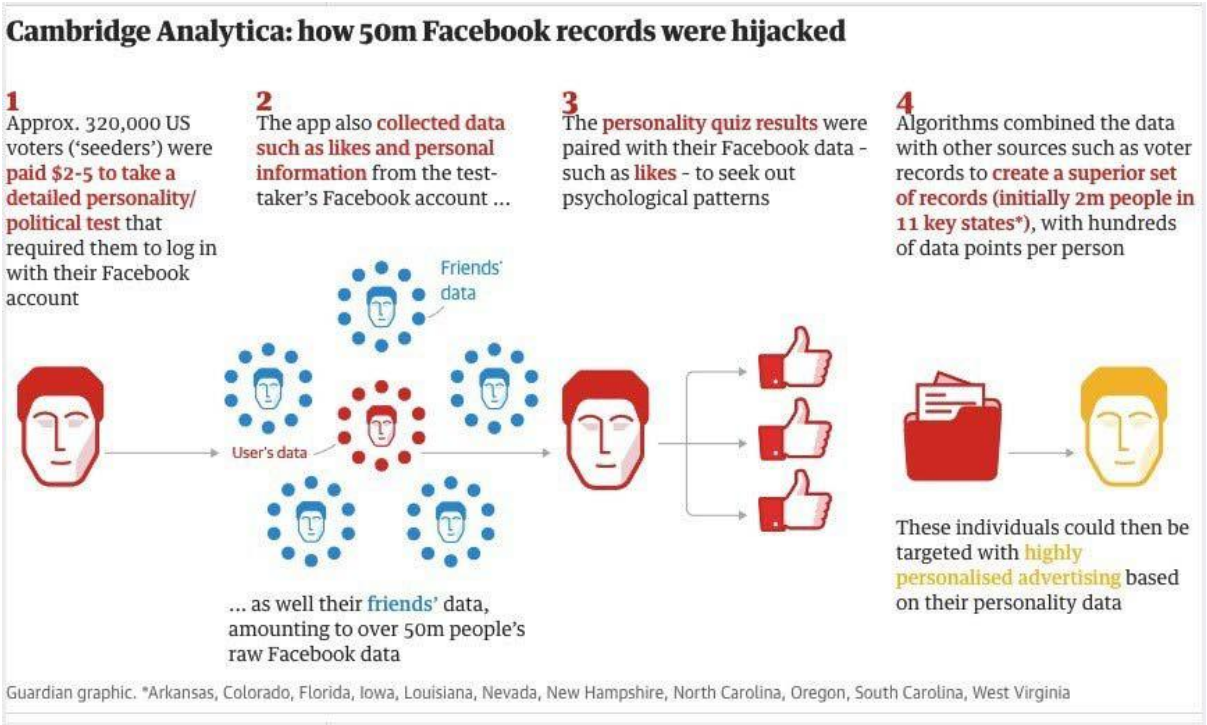


Figure 12 – The Cambridge Analytica case

First of all, people being registered as voters in the USA were invited to take a detailed personality/political test (about 120 questions), available online. The individuals taking the test would be rewarded with a small amount of money (from two to five dollars). They were told that their data would only be used for the academic research.

About 320 000 voters took the test. In order to be receive the reward each individual taking the test had to provide access to his or her Facebook page (step 1). This allowed the system to connect each individual's answers to the information included in his or her Facebook page.

When accessing a test taker's page, Cambridge Analytica collected not only the Facebook page of test takers, but also the Facebook pages of their friends, between 30 and 50 million people altogether (step 2). Facebook data was also collected from other sources.

After this data collection phase, Cambridge Analytica had at is disposition two sets of personal data to be processed (step 3): the data about the test takers, consisting in the information on their Facebook pages, paired with their answers to the questionnaire, and the data about their friends, consisting only in the information on their Facebook pages.

Cambridge Analytica used the data about test-takers as a training set for building a model to profile their friends and other people. More precisely, the data about the test-takers constituted a vast training set, where the information on an individual's Facebook pages (likes, posts, links, etc.) provided values for predictors (features) and the answers to the questionnaire (and psychological and political attitudes expressed by such answers) provided values the targets. Thanks to its machine leaning algorithms Cambridge Analytica could use this data to build a model correlating the information in people's Facebook pages to predictions about psychology and political preferences. At this point Cambridge Analytica engaged in massive profiling, namely, in expanding the data available on the people who did not take the test (their Facebook data, and any further data that was available on them), with the predictions provided by the model. For instance, if test-takers having a certain pattern of Facebook likes and posts were classified as having a neurotic personality, the same assessment could be extended also to non-test-takers having similar patterns in their Facebook data.

Finally (stage 4), based on this personality/political profiling, potential voters who were likely to change their voting behaviour were identified (in US States in which a small change could make a difference) if prodded with appropriate messages. These voters where targeted with personalised political ads and with other messages that could trigger the desired change in voting behaviour, possibly building upon their emotions and prejudice and without making them aware of the purpose of such messages.[39]

2.3.6. Towards surveillance capitalism or surveillance state?

Some authors have taken a positive view of the development of systems based on the massive collection of information. They have observed that the integration of AI and big data enables increased efficiency and provides new means for managing and controlling individual and social behaviour.

When economic transactions – and more generally social interaction and individual activities– are computer-mediated, they provide for a ubiquitous and granular recording of data: computer systems can observe, verify and analyse any aspects of the activities in question.[40] The recorded data can be used to construct user profiles, to personalise interactions with users (as in targeted commercial communication), to engage in experimentation (e.g., to evaluate user responses to changes in prices and messaging), to guide and control behaviour (e.g., for the purpose of economic or political persuasion). In this context, new models of economic and social interaction become possible, which are based on the possibility of observing every behaviour, and of automatically linking penalties and rewards to it. Consider for instance how online consumers trust vendors of goods and services with whom they have never had any personal contact, relying on the platform through which such goods and services are provided, and on the platform's methods for rating, scoring, selecting, and excluding. Consider too how blockchain systems – through a shared unmodifiable ledger recording all transactions – enable the creation of digital currencies, self-executing smart contracts, and digital organisations.

According to Alex Pentland the director of the Human Dynamics Lab at the MIT Media Lab, AI and big data may enable the development of a 'social physics', i.e., a rigorous social science.[41] The availability of vast masses of data and of methods and computational resources to process these data could support a social science having solid theoretical-mathematical foundations as well as operational capacities for social governance.

> By better understanding ourselves, we can potentially build a world without war or financial crashes, in which infectious disease is quickly detected and stopped, in which energy, water, and other resources are no longer wasted, and in which governments are part of the solution rather than part of the problem.

The prospect for economic and social improvement offered by AI and big data is accompanied by the risks referred to as 'surveillance capitalism' and the 'surveillance state'.

According to Shoshana Zuboff, *surveillance capitalism* is the leading economic model of the present age.[42] Zuboff points out to the classic analysis by historian Karl Polanyi[43] who observed that industrial capitalism also treats as commodities (products to be sold in the market) entities that are not produced for the market: human life becomes 'labour' to be bought and sold, nature becomes 'land' or 'real estate', exchange becomes 'money.' As a consequence, the dynamics of capitalism

[39] On the problems related to disinformation and propaganda, see Bayer et al (2019).

[40] Varian (2010, 2014),

[41] Pentland (2015,28),

[42] Zuboff (2019), see also Cohen (2019) who prefers to speak of 'informational capitalism."

[43] Polanyi [1944] 2001),

produces destructive tensions – exploitation, destruction of environment, financial crises – unless countervailing forces, such as law, politics and social organisations (e.g., workers' and consumers' movements), intervene to counteract, moderate and mitigate excesses. According to Zuboff, the surveillance capitalism further expands commodification, extending it to human experience, which it turns into recorded and analysed behaviour, i.e., it transforms into marketable opportunities to anticipate and influence.

> Surveillance capitalism annexes human experience to the market dynamic so that it is reborn as behavior: the fourth 'fictional commodity.' Polanyi's first three fictional commodities – land, labor, and money – were subjected to law. Although these laws have been imperfect, the institutions of labor law, environmental law, and banking law are regulatory frameworks intended to defend society (and nature, life, and exchange) from the worst excesses of raw capitalism's destructive power. Surveillance capitalism's expropriation of human experience has faced no such impediments.[44]

Zuboff observes that in the case of surveillance capitalism, raw market dynamics can lead to novel disruptive outcomes. Individuals are subject to manipulation, are deprived of control over their future and cannot develop their individuality. Social networks for collaboration are replaced by surveillance-based mechanism of incentives and disincentives.

Consider for instance, how service platforms – such as Uber or Lyft in the ridesharing section –record the performance of workers as well the mutual reviews of workers and clients, and link multiple aspects of job performance to rewards or penalties. This new way of governing human behaviour may lead to efficient outcomes, but it affects the mental wellbeing and autonomy of the individuals concerned.[45] According to Zuboff, we have not yet developed adequate legal, political or social measures by which to check the potentially disruptive outcomes of surveillance capitalism and keep them in balance. However, she observes, the GDPR could be an important step in this direction, as a 'springboard to challenging the legitimacy of surveillance capitalism and ultimately vanquishing its instrumentarian power', towards 'society's rejection of markets based on the dispossession of human experience as a means to the prediction and control of human behavior for others' profit.'

The need to limit the commercial use of personal data has led to new legal schemes not only in Europe, but also in California, the place where many world-leading 'surveillance capitalists' have their roots; the CCPA (California Consumer Privacy Act), which came into effect on January 2020, provides consumers with rights to access their data and to prohibit data sales (broadly understood).

At the governmental level, surveillance capitalism finds its parallel in the so-called 'surveillance state', which is characterised as follows:

> In the National Surveillance State, the government uses surveillance, data collection, collation, and analysis to identify problems, to head off potential threats, to govern populations, and to deliver valuable social services. The National Surveillance State is a special case of the Information State-a state that tries to identify and solve problems of governance through the collection, collation, analysis, and production of information.[46]

In government too, AI and big data can bring great advantages, supporting efficiency in managing public activities, coordinating citizens' behaviour, and preventing social harms. However, they may also enable new kinds of influence and control, underpinned by purposes and values that may conflict with the requirements of democratic citizenship. A paradigmatic example is that of the

[44] Zuboff (2019, 507).

[45] Cristianini, and Scantamburlo (2019).

[46] Balkin (2008, 3).

Chinese Social credit systems, which collects data about citizens and assigns to those citizens scores that quantify their social value and reputation. This system is based on the aggregation and analysis of personal information. The collected data cover financial aspects (e.g., timely compliance with contractual obligations), political engagement (e.g., participation in political movements and demonstrations), involvement in civil and criminal proceedings (past and present) and social action (e.g. participation in social networks, interpersonal relationships, etc.). On the basis these data items, citizens may be assigned positive or negative points, which contribute to their social score. A citizen's overall score determines his or her access to services and social opportunities', such as universities, housing, transportation, jobs, financing, etc. The system's purported objective is to promote mutual trust, and civic virtues. One may wonder whether opportunism and conformism may be rather promoted to the detriment of individual autonomy and genuine moral and social motivations.

Thus, the perspective of an integration or symbiosis between humans and intelligent machine, while opening bright prospects, does not entail that all applications of AI should be accepted as long as they meet technological and fairness standards. It has been argued that following this approach

> What is achieved is resignation – the normalization of massive data capture, a one-way transfer to technology companies, and the application of automated, predictive solutions to each and every societal problem.[47]

Indeed, in some cases and domain AI and big data applications – even when accurate and unbiased– may have individual and social costs that outweigh their advantages. To address these cases, we need to go beyond requiring unbiasedness and fairness, and ask further questions, which may challenge the very admissibility of the AI applications at stake.

> Which systems really deserve to be built? Which problems most need to be tackled? Who is best placed to build them? And who decides? We need genuine accountability mechanisms, external to companies and accessible to populations. Any A.I. system that is integrated into people's lives must be capable of contest, account, and redress to citizens and representatives of the public interest.[48]

Consider, for instance, systems that are able to recognise sexual orientation, or criminal tendencies from the faces of persons. Should we just ask that whether these systems provide reliable assessments, or should we rather ask whether they should be built at all. Should we 'ban them, or at least ensure they are only licensed for socially productive uses?'[49] The same may concern extremely intrusive ways to monitor, analyse, punish or reword the behaviour of workers by online platforms for transportation (e.g. Uber) or other services. Similarly, some AI-based financial application, even when inclusive, may have a negative impact on their addressees, e.g., pushing them into perpetual debt.[50]

2.3.7. The general problem of social sorting and differential treatment

The key aspect of AI system, of the machine learning type, is their ability to engage in differential inference: different combinations of predictor-values are correlated to different predictions. As discussed above, when the predictors concern data on individuals and their behaviour, the prediction also concerns features or attitudes of such individuals. Thus, for instance, as noted above,

[47] Powles and Nissenbaum (2018).

[48] Powles and Nissenbaum (2018).

[49] Pasquale (2019).

[50] Pasquale (2019).

a certain financial history, combined with data on residence or internet use, can lead to a prediction concerning financial reliability and possibly to a credit score.

A new dynamic of stereotyping and differentiation takes place. On the one hand, the individuals whose data support the same prediction, will be considered and treated in the same way. On the other hand, the individuals whose data support different predictions, will be considered and treated differently.

This equalisation and differentiation, depending on the domains in which it is used and on the purposes that it is meant to serve, may affect positively or negatively the individuals concerned but also broader social arrangements.

Consider for instance the use of machine learning technologies to detect or anticipate health issues. When used to direct patients to therapies or preventive measures that are most suited to their particular conditions, these AI applications are certainly beneficial, and the benefits outweigh – at least when accompanied by corresponding security measures – whatever risks that may be linked to the abuse of patients' data. The benefits, moreover, concern in principle all data subjects whose data are processed for this purpose, since each patient has an interest in a more effective and personalised treatment. Processing of health-related data may also be justified on grounds of public health (Article 9 (2)(h)), and in particular for the purpose of 'monitoring epidemics and their spread' (Recital 46). This provision has become hugely relevant in the context of the Coronavirus disease 2019 (COVID-19) epidemics. In particular a vast debate has been raised by development of applications for tracing contacts, in order to timely monitor the diffusion of the infection.[51] AI is being applied in the context of the epidemics in multiple ways, e.g., to assess symptoms of individuals and to anticipate the evolution of the epidemics. Such processing should be viewed as legitimate as long as it effectively contributes to limit the diffusion and the harmfulness of the epidemics, assuming that the privacy and data protection risks are proportionate to the expected benefit, and that appropriate mitigation measures are applied.

The use of the predictions based on health data in the context of insurance deserves a much less favourable assessment. In this case there would be some gainers, namely the insured individuals getting a better deal based on their favourable heath prospects, but also some losers, namely those getting a worse deal because of their unfavourable prospects. Thus, individuals who already are disadvantaged because of their medical conditions would suffer further disadvantage, being excluded from insurance or being subject to less favourable conditions. Insurance companies having the ability (based on the data) to distinguish the risks concerning different applicants would have a competitive advantage, being able to provide better conditions to less risky applicants, so that insurers would be pressured to collect as much personal data as possible.

Even less commendable would be the use of health predictions in the context of recruiting, which would involve burdening less healthy people with unemployment or with harsher work conditions. Competition between companies would also be affected, and pressure for collecting health data would grow.

Let us finally consider the domain of targeted advertising. In principle, there seems to be nothing wrong in providing consumers with ads match their interests, helping them to navigate the huge set of options that are available online. However, personalised advertising involves the massive collection of personal data, which is used in the interests of advertisers and intermediaries, possibly against the interests of data subjects. Such data provide indeed new opportunities for influence and

[51] See the European Data Protection Board Guidelines 04/2020 on the use of location data and contact-tracing tools in the context of the Covid-19 outbreak.

control, they can be used to delivers deceitful, or aggressive messages, or generally messages that bypass rationality by appealing to weaknesses and emotions.

> Rather than predominantly stimulating the development and exercise of conscious and deliberate reason, today's networked information flows [...] employ a radical behaviorist approach to human psychology to mobilize and reinforce patterns of motivation, cognition, and behavior that operate on automatic, near-instinctual levels and that may be manipulated instrumentally.[52]

Thus, people may be induced to purchase goods they do not need, to overspend, to engage in risky financial transactions, to indulge in their weaknesses (e.g. gambling or drug addiction). The opportunity for undue influence is emphasised by the use of psychographic techniques that enable psychological attitudes to be inferred from behaviour, and thus disclose opportunities for manipulation.[53]

Even outside of the domain of aggressive or misleading advertising, we may wonder what real benefits to consumers and to society may be delivered by practices such as price discrimination, namely, the policy of providing different prices and different conditions to different consumers, depending on predictions on their readiness to pay. Economist have observed that this practice may not only harm consumers but also affect the functioning of markets.

> Because AI and big data enable firms to assess how much each individual values different products and is therefore willing to pay, they give these firms the power to price discriminate, to charge more to those customers who value the product more or who have fewer options. Price discrimination not only is unfair, but it also undermines the efficiency of the economy: standard economic theory is based on the absence of discriminatory pricing.[54]

The practice of price discrimination shows how individuals may be deprived of access to some opportunities when they are provided with personalised informational environment engineered by third parties, i.e., with informational cocoons where they are presented with data and choices that are selected by others, according to their priorities.

Similar patterns characterise the political domain, where targeted ads and messages can enable political parties to selectively appeal to individuals having different political preferences and psychological attitudes, without them knowing what messages are addressed to other voters, in order to direct such individuals towards the desired voting behaviour, possibly against their best judgement. In this case too, it may be wondered whether personalisation really contributes to the formation of considered political opinions, or whether it is averse to it. After the Cambridge Analytica case, some internet companies have recognised how microtargeted political advertising may negatively affect the formation of political opinion, and have consequently adopted some remedial measures. Some have refused to transmit paid political ads (Twitter), others have restricted the factors used for targeting, only allowing general features such as age, gender, or residence code, to the exclusion of other aspects, such as political affiliation or public voter records (Google).

In conclusion we may say that AI enables new kinds of algorithmic mediated differentiations between individuals, which need to be strictly scrutinised. While in the pre-AI era differential treatments could be based on the information extracted through individual interactions (the typical job interview) and human assessments, or on few data points whose meaning was predetermined, in the AI era differential treatments can be based on vast amounts of data enabling probabilistic

[52] Cohen 2019

[53] Burr and Cristianini (2019).

[54] Stiglitz (2019, 115).

predictions, which may trigger algorithmically predetermined responses. The impacts of such practices can go beyond the individuals concerned, and affect important social institution, in the economical as well as in the political sphere.

The GDPR, as we shall see in the following section, provides some constraints: the need for a legal basis for any processing of personal data, obligations concerning information and transparency, limitations on profiling and automated decision-making, requirements on anonymisation and pseudonymisation, etc. These constraints, however, need to be coupled with strong public oversight, possibly leading to the ban of socially obnoxious forms of differential treatment, or to effective measures that prevent abuses. The decision on what forms of algorithmic differentiations to allow is a highly political one, which should be entrusted to technical authorities only under the direction of politically responsible bodies, such as in particular, parliamentary assemblies. It is a decision that concerns what society we want to live in, under what arrangement of powers and opportunities.

2.4. AI, legal values and norms

To promote valuable practices around the use of AI, we need to ensure that the development and deployment of AI takes place in a sociotechnical framework (inclusive of technologies, human skills, organisational structures, and norms) where individual interests and social goods are both preserved and enhanced.

To provide regulatory support to the creation of such a framework, we need to focus not only on existing regulations, but also on first principles, given that the current rules may fail to provide appropriate solutions and directions to citizens, companies and enforcement authorities. First principles include fundamental rights and social values at both the ethical and the legal level.

2.4.1. The ethical framework

A high-level synthesis of the ethical framework for AI is provided for instance by the AI4People document, which describes the opportunities provided by AI and the corresponding risks as follows:[15]

- enabling human self-realisation, without devaluing human abilities;
- enhancing human agency, without removing human responsibility; and
- cultivating social cohesion, without eroding human self-determination.

The High-Level Expert Group on Artificial Intelligence, set up by the European Commission, recently published a set of ethics guidelines for trustworthy AI. According to the expert group, the foundation of legal, ethical and robust AI should be grounded on fundamental rights and reflect the following four ethical principles:

- Respect for human autonomy: humans interacting with AI must be able to keep full and effective self-determination over themselves. AI should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans, but should be rather designed to augment, complement and empower human cognitive, social and cultural skills.
- Prevention of harm: the protection of human dignity as well as mental and physical integrity should be ensured. Under this principle, AI systems and the environments in which they operate must be safe and secure, they should neither cause nor exacerbate harm or otherwise adversely affect human beings.
- Fairness: it should be intended under its substantive and procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of

both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. The procedural dimension entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.

- Explicability: algorithmic processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions explainable to those affected both directly and indirectly.

According to the High-Level Expert Group, in order to implement and achieve trustworthy AI, seven requirements should be met, building on the principles mentioned above:

- Human agency and oversight, including fundamental rights;
- Technical robustness and safety, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility;
- Privacy and data governance, including respect for privacy, quality and integrity of data, and access to data;
- Transparency, including traceability, explainability and communication;
- Diversity, non-discrimination and fairness, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation;
- Societal and environmental wellbeing, including sustainability and environmental friendliness, social impact, society and democracy;
- Accountability, including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Implementation of these requirements should occur throughout an AI system's entire life cycle as required by specific applications.

A recent comparative analysis of documents on the ethics of AI has noted a global convergence around the values of transparency, non-maleficence, responsibility, and privacy, while dignity, solidarity and responsibility are less often mentioned.[55] However, substantial differences exists on how to how to balance competing requirements, i.e., on how to address cases in which some of the values just mentioned are affected, but at the same time economic, administrative, political or military advantages are also obtained.

2.4.2. Legal principles and norms

Moving from ethics to law, AI may both promote and demote different fundamental rights and social values included in the EU Charter and in national constitutions. AI indeed can magnify both the positive and the negative impacts of ICTs on human rights and social values.[56] The rights to privacy and data protection (Articles a 7 and 8 of the Charter) are at the forefront, but other rights are also at stake: dignity (article 1), right to liberty and security, freedom of thought, conscience and religion (Article 10), freedom of expression and information (Article 11), freedom of assembly and association (Article 12), freedom of arts and science (Article 13), right to education (article 14), freedom to choose an occupation and right to engage in work (Article 15), right to equality before the law (Article 20), right to non-discrimination (article 21), equality between men and women (Article 23), rights of the child (Article 24), right to fair and just working conditions (Article 31), right

[55] Jobin et al (2019).

[56] For a review of the impacts of ICTs on rights and values, see Sartor (2017), De Hert and Gutwirth (2009).

to health care (article 35), right to access to services of general economic interest (Article 36), consumer protection (Article 38), right to good administration (Article 41), right to an effective remedy and to a fair trial (Article 47). Besides individual right also social values are at stake, such as democracy, peace, welfare, competition, social dialogue efficiency, advancement in science, art and culture, cooperation, civility, and security.

Given the huge breath of its impacts on citizens' individual and social lives, AI falls under the scope of different sectorial legal regimes. These regimes include especially, though not exclusively, data protection law, consumer protection law, and competition law. As has been observed by the European Data Protection Supervisor (EDPS) in Opinion 8/18 on the legislative package 'A New Deal for Consumers,' there is synergy between the three regimes. Consumer and data protection law share the common goals of correcting imbalances of informational and market power, and, along with competition law, they contribute to ensuring that people are treated fairly. Other domains of the law are also involved in AI: labour law relative to the new forms of control over worker enabled by AI; administrative law relative to the opportunities and risk in using AI to support administrative decision-making; civil liability law relative to harm caused by AI driven systems and machines; contract law relative to the use of AI in preparing, executing and performing agreements; laws on political propaganda and elections relatively to the use of AI in political campaigns; military law on the use of AI in armed conflicts; etc.

2.4.3. Some interests at stake

The significance that AI bears to different areas of the law has to do with the nature of the interest that are affected by the deployment of AI technologies. Here are some of the interests more directly and specifically involved.

First, there is the interest in data protection and privacy, namely, the interest in a lawful and proportionate processing of personal data subject to oversight. This is hardly compatible with an online environment where every action is tracked, and the resulting data is used to extract further information about the individuals concerned, beyond their control, and to process this information in ways that may run counter to their interests.

The processing of personal data through AI systems may also affect citizens' interest in fair algorithmic treatment, namely, their interest in not being subject to unjustified prejudice resulting from automated processing.

The possibility of algorithmic unfairness, as well as the need to keep the processing of personal data under control and to understand (and possibly challenge) the reasons for determinations that affect individuals, raises concern from an algorithmic transparency/explicability standpoint. Citizens want to know how and why a certain algorithmic response has been given or a decision made, so as 'to understand and hold to account the decision-making processes of AI.' [57]

Individual autonomy is affected when citizens interact with black boxes,[17] whose functioning is not accessible to them, and whose decisions remain unexplained and thus unchallengeable.[58]

As observed above, since AI systems have access to a huge amount of information about individuals and about people similar to them, they can effortlessly use this information to elicit desired behaviour for purposes that citizens may not share, possibly in violation of fiduciary expectations they have toward the organisation that is deploying the AI system in question.[59] Thus, individuals

[57] Floridi et al (2018).

[58] Pasquale (2015).

[59] On fiduciary obligations related to the use of AI, see Balkin (2017).

have an interest in not being misled or manipulated by AI systems, but they also have an interest in being able to trust such systems, knowing that the controllers of those systems will not profit from the people's exposure (possibly resulting from personal data). Reasonable trust is needed so that individuals do to waste their limited and costly cognitive capacities in trying to fend off AI systems' attempts to mislead and manipulate them.

Finally, citizens have an indirect interest in fair algorithmic competition, i.e., in not being subject to market-power abuses resulting from exclusive control over masses of data and technologies. This is of direct concern to competitors, but the lack of competition may negatively affect consumers, too, by depriving them of valuable options and restricting their sphere of action. Moreover, the lack of competition enables the leading companies to obtain huge financial resources, which they can use to further increase their market power (e.g., by preventively buying potential competitors), or to promote their interests through influence.ng public opinion and politics.

2.4.4. AI technologies for social and legal empowerment

To ensure an effective protection of citizens' rights and to direct AI towards individual and social goods, regulatory initiatives are an essential element. However, regulatory instruments and their implementation by public bodies may be insufficient. Indeed, AI and big data are employed in domains already characterised by a vast power imbalance, which they may contribute to accentuate. In fact, these technologies create new knowledge (analytical and forecasting abilities) and powers (control and influence capacities) and make them available to those who govern these technologies.

To ensure an adequate protection of citizens, beside regulation and public enforcement, also the countervailing power of civil society[60] is needed to detect abuses, inform the public, activate enforcement, etc. In the AI era, an effective countervailing power needs also to be supported by AI: only if citizens and their organisations are able to use AI to their advantage, can they resist, and respond to, AI-powered companies and governments.[61] Moreover, active citizenship is an important value in itself, that needs to be preserved and advanced at a time in which we tend to delegate to technology (and in particular to AI) a vast amount of relevant decisions.

A few examples of citizen-empowering technologies are already with us, as in the case of ad-blocking systems as well as more traditional anti-spam software and anti-phishing techniques. Yet, there is a need to move a step forward. Services could be deployed with the goal of analysing and summarising massive amounts of product reviews or comparing prices across a multitude of platforms. One example in this direction is offered by CLAUDETTE:[62] an online system for the automatic detection of potentially unfair clauses in online contracts and in privacy policies.[63] Considerable effort has also been devoted to the development of data mining techniques for detecting discrimination with the aim to build supporting tools that could identify prejudice and unfair treatments in decisions that regard consumers.[64]

The growing interest in privacy and data protection has resulted in several proposals for automatically extracting, categorising and summarising information from privacy documents, and assisting users in processing and understanding their contents. Multiple AI methods to support data protection could be merged into integrated PDA-CDA (Privacy digital assistants/consumer digital assistants), meant to prevent excessive/unwanted/unlawful collection of personal data and well as

[60] Galbraith (1983).

[61] Lippi et al (2020).

[62] https://claudette.eui.eu/

[63] Contissa et al (2018), Lippi et al (2019).

[64] Ruggeri, Pedreschi, and Turini (2010).

to protect users from manipulation and fraud, provide them with awareness of fake and untrustworthy information, and facilitate their escape from 'filter bubbles' (the unwanted filtering/pushing of information).

It may be worth considering how the public could support and incentivise the creation and distribution of AI tools to the benefit of data subject and citizens. Such tools would provide new opportunities for research, development, and entrepreneurship. They would contribute to reduce unfair and unlawful market behaviour and favour the development of legal and ethical business models. Finally, citizen-empowering technologies would support the involvement of civil society in monitoring and assessing the behaviour of public and private actors and of the technologies deployed by the latter, encouraging active citizenship, as a complement to the regulatory and law-enforcement activity of public bodies.

3. AI in the GDPR

In this section the provisions of the GDPR are singularly analysed to determine the extent to which their application is challenged by of AI as well as the extent to which they may influence the development of AI applications.

3.1. AI in the conceptual framework of the GDPR

Unlike the 1995 Data Protection Directive, the GDPR contains some terms referring to the Internet (Internet, social networks, website, links, etc.), but it does not contain the term 'artificial intelligence', nor any terms expressing related concepts, such as intelligent systems, autonomous systems, automated reasoning and inference, machine learning or even big data. This reflects the fact that the GDPR is focussed on the challenges emerging for the Internet – which were not considered in the 1995 Data Protection Directive, but were well present at the time when GDPR was drafted – rather than on new issues pertaining to AI, which only acquired social significance in most recent years. However, as we shall see, many provisions in the GDPR are very relevant to AI.

3.1.1. Article 4(1) GDPR: Personal data (identification, identifiability, re-identification)

The concept of personal data plays a key role in the GDPR, characterising the material scope of the regulation. The provision in the GDPR only concern personal data, to the exclusion of information that does not concerns humans (e.g., data on natural phenomena), and also to the exclusion of information that, though concerning humans does not refer to particular individuals (e.g., general medical information on human physiology or pathologies) or has been effectively anonymised so that it has lost its connection to particular individuals. Here is how personal data are defined in Article 4 (1) GDPR:

> 'personal data' means any information relating to an identified or identifiable natural person ('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;

Recital (26) addresses identifiability, namely, the conditions under which a piece of data which is not explicitly linked to a person, still counts as personal data, since the possibility exists to identify the person concerned. Identifiability depends on the availability of 'means reasonably likely to be used' for successful re-identification, which in its turn, depends on the technological and sociotechnical state of the art:

> 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.

Through pseudonymisation, the data items that identify a person (i.e., the name) are substituted with a pseudonym, but the link between the pseudonym and the identifying data items can be retraced by using separate information (e.g., through a table linking pseudonyms and real names, or through cryptography key to decode the encrypted names). Recital (26) specifies that pseudonymised data still are personal data.

> Personal data which have undergone pseudonymisation, 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.

The connection between the personal nature of information and technological development is mentioned at Recital (9) of Regulation 2018/1807:

> If technological developments make it possible to turn anonymised data into personal data, such data are to be treated as personal data, and Regulation (EU) 2016/679 is to apply accordingly.

The concept of non-personal data is not positively defined in the EU legislation, as it includes whatever data that are not personal data as defined in the GDPR. Regulation 2018/1807,[65] at Recital 9 provides the following examples of non-personal data: aggregate and anonymised datasets used for big data analytics, data on precision farming that can help to monitor and optimise the use of pesticides and water, or data on maintenance needs for industrial machines.'

In connection with the GDPR definition of personal data, AI raises in particular two key issues: (1) the 're-personalisation' of anonymous data, namely the re-identification of the individuals to which such data are related; (2) and the inference of further personal information from personal data that are already available.

Re-identification

The first issue concerns of identifiability. AI, and more generally methods for computational statistics, increases the identifiability of apparently anonymous data, since they enable nonidentified data (including data having been anonymised or pseudonymised) to be connected to the individuals concerned

> [N]umerous supposedly anonymous datasets have recently been released and reidentified. In 2016, journalists reidentified politicians in an anonymized browsing history dataset of 3 million German citizens, uncovering their medical information and their sexual preferences. A few months before, the Australian Department of Health publicly released de-identified medical records for 10% of the population only for researchers to reidentify them 6 weeks later. Before that, studies had shown that de-identified hospital discharge data could be reidentified using basic demographic attributes and that diagnostic codes, year of birth, gender, and ethnicity could uniquely identify patients in genomic studies data. Finally, researchers were able to uniquely identify individuals in anonymized taxi trajectories in NYC27, bike sharing trips in London, subway data in Riga, and mobile phone and credit card datasets.[66]

The re-identification of data subjects is usually based on statistical correlations between non-identified data and personal data concerning the same individuals.

[65] 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.

[66] Rocher et al (2019).

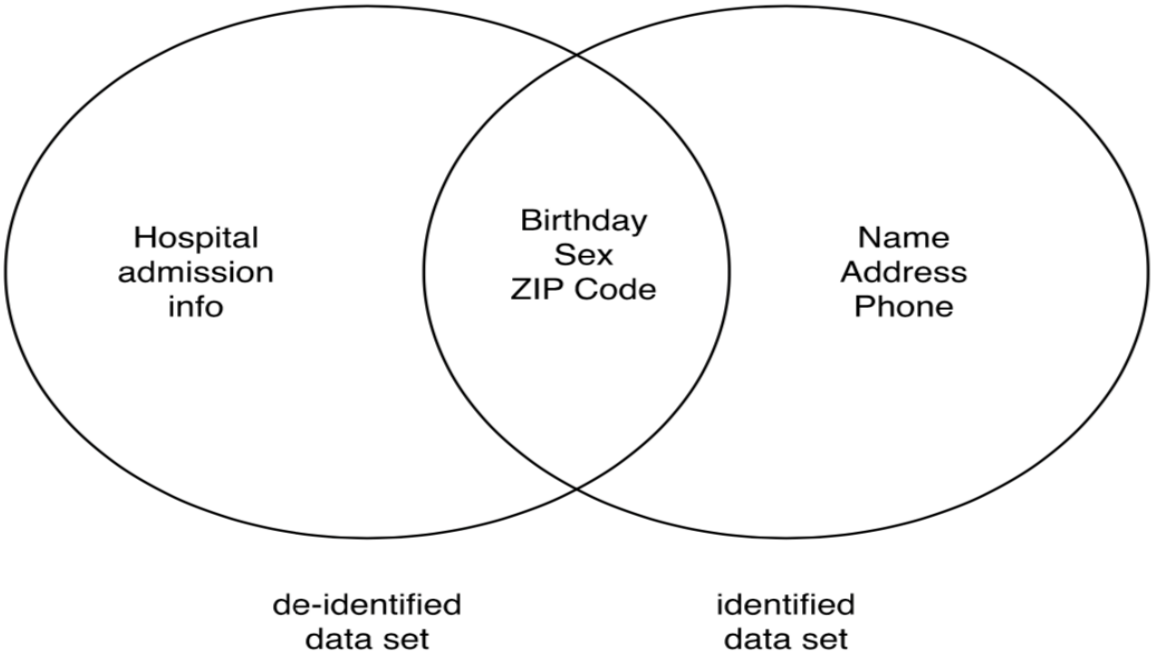


Figure 13 – The connection between identified and de-identified data

Figure 13 illustrates a connection between an identified and a de-identified data set that enabled the re-identification of the health record of the governor of Massachusetts. This result was obtained by searching for de-identified data that matched the Governor's date of birth, ZIP code and gender.[67] Another classic example is provided the Netflix price database case, in which anonymised movie ratings could be re-identified by linking them to non-anonymous ratings in IMDb (Internet Movie Database). In fact, knowing only two non-anonymous reviews by an IMDb user, it was possible to identify the reviews by the same user in the anonymous database. Similarly, it has been shown that an anonymous user of an online service can be re-identified by that service, if the service knows that the user has installed four apps on his or her device, and the service has access to the whole list of apps installed by each user.[68]

Re-identification can be viewed as a specific kind of inference of personal data: through re-identification. A personal identifier is associated to previously non-identified data items, which, as a consequence, become personal data. Note that for an item to be linked to a person, it is not necessary that the data subject be identified with absolute certainty; a degree of probability may be sufficient to enable a differential treatment of the same individual (e.g., the sending of targeted advertising).

Thanks to AI and big data the identifiability of the data subjects has vastly increased. The personal nature of a data idem no longer is a feature of that item separately considered. It has rather become a contextual feature. As shown above, an apparently anonymous data item becomes personal in the context of further personal data that enable re-identification. For instance, the identifiability of the Netflix movie reviewers supervened on the availability of their named reviews on IMDb. As it has been argued, 'in any "reasonable" setting there is a piece of information that is in itself innocent, yet in conjunction with even a modified (noisy) version of the data yields a privacy breach.'[69]

[67] Sweeney (2000).

[68] Achara et al (2015)

[69] Dwork and Naor (2010, 93).

This possibility can be addressed in two ways, neither of which is fail-proof. The first consists in ensuring that data is de-identified in ways that make it more difficult to re-identify the data subject; the second consists in implementing security processes and measures for the release of data that contribute to this outcome.[70]

Inferred personal data

As noted above, AI systems may infer new information about data subjects, by applying algorithmic models to their personal data. The key issue, from a data protection perspective, is whether the inferred information should be considered as new personal data, distinct from the data from which it has been inferred. Assume for instance, that an individual's sexual orientation is inferred from his or her facial features or that an individual's personality type is inferred from his or her online activity. Is the inferred sexual orientation or personality type a new item of personal data? Even when the inference only is probabilistic? If the inferred information counts as new personal data, then automated inferences would trigger all the consequences that the processing of personal data entails according to the GDPR: the need of a legal basis, the conditions for processing sensitive data, the data subject's rights, etc.

Some clues on the legal status of automatically inferred information can be obtained by considering the status of information inferred by humans. There is uncertainty about whether assertions concerning individuals, resulting from human inferences and reasoning may be regarded as personal data. This issue has been examined by the ECJ in Joint Cases C-141 and 372/12, where it was denied that the legal analysis, by the competent officer, on an application for a residence permit could be deemed personal data.[71] According to the ECJ and the Advocate General, only the data on which the analysis was based (the input data about the applicant) as well as the final conclusion of the analysis (the holding that the application was to be denied) were to be regarded as personal data. This qualification did not apply to the intermediate steps (the intermediate conclusions in the argument chain) leading to the final conclusion.

In the subsequent decision on Case C-434/16,[72] concerning a candidate's request to exercise data protection rights relative to an exam script and the examiners' comments, the ECJ apparently departed from the principle stated in Joint Cases C-141 and 372/12, arguing that the examiner's comments, too, were personal data. However, the Court held that data protection rights, and in particular the right to rectification, should be understood in connection with the purpose of the data at issue. Thus, according to the Court, the right to rectification does not include a right to correct a candidate's answers or the examiner's comments (unless they were incorrectly recorded). In fact, according to the ECJ, data protection law is not intended to ensure the accuracy of decision-making processes or good administrative practices. Thus, an examinee has the right to access both to the exam data (the exam responses) and the reasoning based on such data (the comments), but he or she does not have a right to correct the examiners' inferences (the reasoning) or the final result.

The view that inferred data are personal data was endorsed by the Article 29 WP, being implied in particular by the broad concept of personal data adopted in Opinion 4/2007.[73] This broad concept of personal data is presupposed by the Article 29 WP's statement, that in case of automated inference (profiling) data subjects have the right to access both the input data and the (final or intermediate) conclusions automatically inferred from such data.[74]

---

[70] Rubinstein and Harzog (2016).

[71] Joint cases c-141 and 372/12. See Joined Cases C–141 & 372/12, *YS, M and S* v. *Minister voor Immigratie, Integratie en Asiel,* 2014 E.C.R. I-2081, ¶ 48.

[72] Case C-434/16, *Peter Nowak* v. *Data Protection Commissioner*, 34.

[73] Opinion 4/2007

[74] Opinion 216/679, adopted on 3 October 2017, revised in 6 February 2018.

3.1.2. Article 4(2) GDPR: Profiling

The definition of profiling, while not using explicitly referring to AI, addresses processing that is today is typically accomplished using AI technologies. This processing consists in using the data concerning person to infer information on further aspects of that person:

> 'profiling' means 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;

According to the Article 29 WP,[75] profiling aims at classifying persons into categories of groups sharing the features being inferred:

> 'broadly speaking, profiling means gathering information about an individual (or group of individuals) and evaluating their characteristics or behaviour patterns in order to place them into a certain category or group, in particular to analyse and/or make predictions about, for example, their:
>
> - ability to perform a task;
> - interests; or
> - likely behaviour.'

AI and profiling

AI and big data, in combination with the availability of extensive computer resources, have vastly increased the opportunities for profiling. Indeed, machine learning-based approaches, as described in the previous sections, are often meant to provide inferences – classifications, predictions or decisions – when applied to data concerning individuals.

Assume that a classifier has trained on a vast set of past examples, which link certain features of individuals (the predictors), to another feature of the same individuals (the target). Through the training, the system has learned an algorithmic model can be applied to new cases: if the model is given predictors-values concerning a new individual, it infers a corresponding target value for that individual, i.e., a new data item concerning him or her.

For instance, the likelihood of heart disease of applicants for insurance may be predicted on the basis of their health records, but also on the basis of their habits (on eating, physical exercise, etc.) or social conditions; the creditworthiness of loan applicants may be predicted on the basis of their financial history but also on the basis of their online activity and social condition; the likelihood that convicted persons may reoffend may be predicted on the basis their criminal history, but also possibly their character (as identified by personality test) and personal background. These predictions may trigger automated determinations concerning, respectively, the price of a health insurance, the granting of a loan, or the release on parole.

A learned correlation may also concern a person's propensity to respond in certain ways to certain stimuli. This would enable the transition from prediction to behaviour modification (both legitimate influence and illegal or unethical manipulation). Assume, for instance that a system learns a correlation between certain features and activities (purchases, likes, etc.) of a person and his or her profile as a specific type of consumer, and that the system has also learned (or has been told) that this kind of consumer is interested in certain products and is likely to respond to certain kinds of ads. Consequently, a person who has these features and has engaged in such activities may be sent the

[75] Opinion 216/679, adopted on 3 October 2017, revised in 6 February 2018.

messages that are most likely to trigger the desired purchasing behaviour. The same model can be extended to politics, with regard to messages that may trigger desired voting behaviour.

Inferences as personal data

As noted above, the data inferred through profiling should be considered personal data. In this connection, we need to distinguish the general correlations that are captured by the learned algorithmic model, and the results of applying that model to the description of a particular individual. Consider for instance a machine learning system that has learned a model (e.g., a neural network or a decision tree) from a training set consisting of previous loan applications and outcomes.

In this example, the system's training set consists of personal data: e.g., for each borrower, his name, the data collected on him or her – age, economic condition, education, job, etc. – and the information on whether he or she defaulted on the loan. The learned algorithmic model no longer contains personal data, since it links any possible combinations of possible input values (predictors) to a corresponding likelihood of default (target). The correlations embedded in the algorithmic model are not personal data, since they apply to all individuals sharing similar characteristics. We can possibly view them as group data, concerning the set of such individuals (e.g., those who are assigned a higher likelihood of default, since they have a low revenue, live in a poor neighbourhood, etc.).

Assume that the algorithmic model is then applied to the input data consisting in the description of a new applicant, in order to determine that applicant's risk of default. In this case both the description of the applicant and the default risk attributed to him or her by the model represent personal data, the first being collected data, and the second inferred data.

Rights over inferences

Since inferred data concerning individuals also are personal data under the GDPR – at least when they are used to derive conclusions that are or may be acted upon – data protection rights should in principle also apply, though concurrent remedies and interests have to be taken into account. As noted above, according to the Article 29 Working Party, in the case of automated inferences (profiling) data subjects have a right to access both the personal data used as input for the inference, and the personal data obtained as (final or intermediate) inferred output. On the contrary, the right to rectification only applies to a limited extend. When the data are processed by a public authority, it should be considered whether review procedures already exist which provide for access and control. In the case of processing by private controllers, the right to rectify the data should be balanced with the respect for autonomy of private assessments and decisions.[76]

According to the Article 29 Working Party data subjects have a right to rectification of inferred information not only when the inferred information is 'verifiable' (its correctness can be objectively determined), but also when it is the outcome of unverifiable or probabilistic inferences (e.g., the likelihood of developing heart disease in the future). In the latter case, rectification may be needed not only when the statistical inference was mistaken, but also when the data subject provides specific additional data that support a different, more specific, statistical conclusion. This is linked to the fact that statistical inferences concerning a class may not apply to subclasses of it: it may be the case that students from university *A* usually have lower skills that students from university *B*, but this does not apply to the *A* students having top marks. Accordingly, a top student from university *A* should have the right to contest the inference that put him or her at a disadvantage relative to an average student from *B*.

[76] Wachter and Mittelstadt (2019).

Legal scholars have argued that data subjects should be granted a general right to 'reasonable inference' namely, the right that any assessment of decision affecting them is obtained through automated inferences that are reasonable, respecting both ethical and epistemic standards. Accordingly, data subject should be entitled to challenge the inferences (e.g. credit scores) made by an AI system, and not only the decisions based on such inferences (e.g., the granting of loans). It has been argued that for an inference to be reasonable it should satisfy the following criteria:[77]

(a) Acceptability: the input data (the predictors) for the inference should be normatively acceptable as a basis for inferences concerning individuals (e.g., to the exclusion of prohibited features, such as sexual orientation);

(b) Relevance: the inferred information (the target) should be relevant to the purpose of the decision and normatively acceptable in that connection (e.g., ethnicity should not be inferred for the purpose of giving a loan).

(c) Reliability: both input data, including the training set, and the methods to process them should be accurate and statistically reliable (see Section 2.3.3).

Controllers, conversely, should be prohibited to base their assessment or decisions on unreasonable inferences, and they should also have the obligation to demonstrate the reasonableness of their inferences.

The idea the unreasonable automated inference should be prohibited only applies to inferences meant to lead to assessments and decisions affecting the data subject. They should not apply to inquiries that are motivated by merely cognitive purposes, such as those pertaining to scientific research.

3.1.3. Article 4(11) GDPR: Consent

Consent according to Article 4(11) GDPR should be freely given, specific, informed and unambiguous, and be expressed through a clear affirmative action:

> 'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;

This definition is complemented by Recital (32) which specifies that consent should be granular, i.e., it should be given for all the purposes of the processing.

> Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them.

Consent plays a key role in the traditional understanding of data protection, based indeed on the 'notice and consent' model, according to which data protection is aimed at protecting a right to 'informational self-determination.' This right is indeed exercised by consenting or refusing to content to the processing of one's data, after having been given adequate notice. Against this approach two main criticism have been raised.[78]

The first criticism it that consent is most often meaningless: usually is not based on real knowledge of the processing at stake, nor on a real opportunity to choose. On the one hand, today's processing of personal data is so complex that most data subjects to do not have the skills to understand them

[77] Wachter and Mittelstadt (2019).

[78] See Cate et all (2014).

and anticipate the involved risks. Moreover, even if data subjects possessed such skills, still they would not have the time and energy to go through the details of each privacy policy. On the other hand, a refusal to consent may imply the impossibility to use (or limitation in the use of) services that are important or even necessary to the data subjects.

The second criticism is that consent, when targeted on specific purposes, does not include (and therefore precludes, when considered a necessary basis of the processing) future, often unknown, uses of the data, even when such uses are socially beneficial. Thus, the requirement of consent can 'interfere with future benefits and hinder valuable new discoveries', as exemplified in 'myriad examples', including 'examining health records and lab results for medical research, analysing billions of Internet search records to map flu outbreaks and identify dangerous drug interactions, searching financial records to detect and prevent money laundering, and tracking vehicles and pedestrians to aid in infrastructure planning.'[79]

These criticisms of consent have been countered by observing that it is possible to implement the principles of consent and purpose limitation in ways that are both meaningful to the data subject and consistent with allowing for future beneficial uses of the data.[80]

Firstly, it has been argued that notices should focus on most important issue, and that they should be user-friendly and direct. In particular, simple and clear information should be given on how to opt-in or opt-out relative to critical processing, such as those involving the tracking of users or the transmission of data to third parties. An interesting example is provided by the new California Data Privacy Act, which requires companies to include in their website a link with the words 'do not sell my data' (or a corresponding logo-button) to enable users to exclude transmission of their data to third parties. Further opt-out or opt-in buttons could be presented to all users, to provide ways to express their preferences relatively to tracking, profiling, etc.

Secondly, the GDPR allows that the data that were collected for certain purposes are processed for further purposes, as long as the latter purposes are compatible with the original ones (see Section 3.3.4).

In conclusion, it seems that, as we shall see in the following, the concepts of consent and purpose limitation can be interpreted in ways that are consistent with both the protection of the data subject and the need of enabling beneficial uses of AI. However, AI and big data raise three key issues concerning consent: specificity, granularity, and freedom.

Specificity

The first issue pertains to the specificity of consent: does consent to the processing for a certain purpose also cover further AI-based processing, typically for data analytics and profiling? – e.g., can data on sales be used to analyse consumer preferences and send targeted advertising? This seems to be ruled out, since consent needs to be specific, so that it cannot extend beyond what is explicitly indicated. However, the fact that the data subject has only consented to processing for a certain purpose (e.g., client management) does not necessarily rule out that the data can be processed for a further legitimate purpose (e.g., business analytics): the further processing is permissible when it is covered by a legal basis, and it is not incompatible with the purpose for which the data were collected.

The requirement of specificity is attenuated for scientific research as stated in Recital (33), which allows consent to be given not only for specific research projects, but also for areas of scientific research.

[79] Cate et al (2014, 9).

[80] Cavoukian (2015), Calo (2012).

> It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.

Granularity

The second issue pertains to the granularity of consent. For instance, is a general consent to any kind of analytics and profiling sufficient to authorise the AI-based sending of targeted commercial or political advertising? Recital (43) addresses granularity as follows:

> Consent is presumed not to be freely given if it does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case.

This has two implications for AI application. First it seems that the data subject should not be required to jointly consent to essentially different kinds of AI-based processing (e.g., to economic and political ads). Second, the use of a service should not in principle be dependent on an agreement to be subject to profiling practices. Consent to profiling must be separate from access to the service.[81]

Freedom

The third issue pertains to the freedom of consent: can consent to profiling be considered freely given? This issue is addressed in Recital (42), which excludes the freedom of consent when 'the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment.' According to Recital (43), consent is not free under situations of 'clear imbalance:'

> In order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller.

Situations of imbalance are prevalent in the typical contexts in which AI and data analytics are applied to personal data. Such situations exist in the private sector, especially when a party enjoys market dominance (as is the case for leading platforms), or a position of private power (as is the case for employers relative to their employees). They also exist between public authorities and the individuals who are subject to the powers by such authorities. In all these cases, consent cannot provide a sufficient legal basis, unless it can be shown that there are no risks of 'deception, intimidation, coercion or significant negative consequence if [the data subject] does not consent.'[82]

Finally, consent should be invalid when refusal or withdrawal of consent is linked to a detriment that is unrelated to the availability of the personal data for which consent was refused (e.g., a patients are told that in order to obtain a medical treatment they must consent that their medical data are used for purposes that are not needed for that treatment). This also applies to cases in which consent is required by the provider of a service, even though the processing is not necessary for performing the service.

> if the performance of a contract, including the provision of a service, is dependent on the consent despite such consent not being necessary for such performance.

[81] Article 29 Working Party Guidelines on consent under Regulation 2016/679. Wp259

[82] Article 29 Working Party Guidelines on consent under Regulation 2016/679. Wp259, 7

This typically is the case when the closing of a contract for a service is conditioned on the user's consent to being profiled, the profiling not being needed to provide the service to the individual user.

3.2. AI and the data protection principles

As many authors have observed, AI and big data challenge key data protection principles. In this section, we shall consider each principle separately, so as to determine the extent to which it may constrain intelligent processing.

3.2.1. Article 5(1)(a) GDPR: Fairness, transparency

Article 5(1)(a) requires that personal data should be processed 'lawfully, fairly and in a transparent manner in relation to the data subject.'

Transparency

The idea of transparency is specified in Recital 58, which focuses on conciseness, accessibility and understandability.

> The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used.

As we shall clarify in what follows, this idea is related, but distinct, from the idea of transparent and explainable AI. In fact, the latter idea involves building a 'scientific' model of the functioning of an AI system, rather than providing sufficient information to lay people, relatively to issues that are relevant to them.

Informational fairness

Two different concepts of fairness can be distinguished in the GDPR. The first, which we may call 'information fairness' is strictly connected to the idea of transparency It requires that data subjects are not deceived or misled concerning the processing of their data, as is explicated in Recital (60):

> The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed.

The same recital explicitly requires that information is provided on profiling:

> Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling.

Informational fairness is also linked to accountability, since it presumes that the information to be provided makes it possible to check for compliance. Informational fairness raises specific issues in connection with AI and big data, because of the complexity of the processing involved in AI-applications, the uncertainty of its outcome, and the multiplicity of its purposes. The new dimension of the principle pertains to the explicability of automated decisions, an idea that is explicitly affirmed in the GDPR, as we shall see in the following section. Arguably, the idea of transparency as explicability can be extended to automated inferences, even when a specific decision has not yet been adopted.

A specific aspect of transparency in the context of machine learning concerns access to data, in particular to the system's training set. Access to data may be needed to identify possible causes of unfairness resulting from inadequate or biased data or training algorithm. This is particularly

important when the learned algorithmic model is opaque, so that possible flaws cannot be detected though its inspection.

Substantive fairness

Recital (71) points to a different dimension of fairness, i.e. what we may call substantive fairness, which concerns the fairness of the content of an automated inference or decision, under a combination of criteria, which may be summarised by referring to the aforementioned standards of acceptability, relevance and reliability (see Section3.1.2):

> 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 use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, 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.

3.2.2. Article 5(1)(b) GDPR: Purpose limitation

Article 5(1)(b) sets forth the principle of purpose limitation, according to which personal data should be

> collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation')

The concept of a purpose also figures in Article 6, which establishes a link between the purpose of processing operations and their legal basis. The notion of a purpose is explicitly mentioned in Article 6 only in relation to the first legal basis, namely, consent, which should be given 'for one or more specific purposes', and for the last legal basis, namely 'the purposes of the legitimate interests pursued by the controller or by a third party'. However, the need for legitimate purpose is implicit in the other legal bases, which consist in the necessity of the processing for performing a contract, complying with a legal obligation, protecting vital interests, performing a task in the public interest or exercising a legitimate authority. Finally, the notion of a purpose also comes up in Articles 13(1)(c) and 14(1)(c), requiring controllers to provide information concerning 'the purposes of the processing for which the personal data are intended as well as the legal basis for the processing.'

AI and repurposing

A tension exists between the use of AI and big data technologies and the purpose limitation requirement. These technologies enable the useful reuse of personal data for new purposes that are different from those for which the data were originally collected. For instance, data collected for the purpose of contract management can be processed to learn consumers' preferences and send targeted advertising; 'likes' that are meant to express and communicate one's opinion may be used to detect psychological attitudes, political or commercial preferences, etc.

To establish whether the repurposing of data is legitimate, we need to determine whether a new purpose is 'compatible' or 'not incompatible' with the purpose for which the data were originally collected. According to the Article 29 WP, the relevant criteria are (a) the distance between the new

purpose and the original purpose, (b) the alignment of the new purpose with the data subjects' expectations, the nature of the data and their impact on the data subjects' interests, and (c) the safeguards adopted by the controller to ensure fair processing and prevent undue impacts.[83]

Though all these criteria are relevant to the issue of compatibility, they do not provide a definite answer to the typical issues pertaining to the reuse of personal data in AI applications. To what extent can the repurposing of personal data for analytics and AI be compatible with the purpose of the original collection? Should the data subjects be informed that their data is being repurposed? To address such issues, we need to distinguish what is at stake in the inclusion of a person's data in a training set from the application of a trained model to a particular individual.

Personal data in a training set

In general, the inclusion of a person's data in a training set is not going to affect to a large extent that particular person, since the record concerning a single individual is unlikely to a make a difference in a model that is based in a vast set of such records. However, the inclusion of a single record exposes the data subject to risks concerning the possible misuse of his or her data, unless the information concerning that person is anonymised or deleted once the model is constructed.

Moreover, when considered together with the data provided by similar individuals, the data concerning a person, once included in the training set, contribute to enabling the system's inference concerning a group of people, i.e., the group of all the individuals who share the similarities supporting the inference. Therefore, we may say that the set of all such records affects the common interest of the group in which that person is included. Consider for instance the use of a patient's genetic data to train a model that is then used to diagnose present diseases, or to determine their propensity to develop a disease in the future. The inclusion of a patient's data in a training set will contribute little to the model's predictive power, and it will not specifically affect the patient (unless his or her data are misused). However, the inclusion of the patient's data, alongside with the data of other similar patients, may create a risk for the group of all the patients who might be affected by predictions based on such data. For instance, assume that the trained model links certain predictors to a high probability of a future health issue. Patients who share such predictors, when their data is fed to the model, may either find themselves at an advantage (prevention based on predictive medicine) or at a disadvantage (e.g., discrimination in recruitment or insurance) depending on the how the prediction is used. The risks for the group increase if the predictive model is made available to third parties, which may use it in ways that the data subjects did not anticipate when providing their data.

Personal data for individualised inferences

While, as just noted, the inclusion of a person's data in a training set does not lead to significant impacts on that person, an individual is directly affected when his or her personal data are used as input in the algorithmic model that has been created on the basis of that training set, in order to make inferences concerning that individual. Consider, for instance, the case in which someone's medical data are entered into a model to make a medical diagnosis or to determine that person's prospective health condition. In such a case, we are clearly in the domain of profiling, since the input data (the predictors) concerning an individual are used to infer further personal data concerning him or her.

Let us now consider how the criteria for non-incompatibility established by the Article 29 WP apply on the one hand to the inclusion of personal data in a training set, and on the other hand to the use of personal data as input to profiling algorithms.

[83] Opinion 03/2013 on purpose limitation.

With regard to the use of a person's data in a training set, it seems that since the person is not directly affected by the use of her personal data, the distance between the new purpose and the original purpose should not be a primary concern, nor should be the data subject's expectations. However, we need to consider the risk that the data are misused, against the interest of the data subject (the risk is particularly serious for data on health or other sensitive conditions), as well as the possibility of mitigating this risk through anonymisation or pseudonymisation. Adequate security measures also are the key precondition for the legitimate use of personal data in a training set.

Different considerations pertain to the use of a personal data as input to algorithmic models that provide inferences concerning the data subject. This case clearly falls within the domain of profiling as the inference directly affects the individuals concerned. Therefore, the criteria indicated by the Article 29 WP have to be rigorously applied.

Obviously, the two uses of personal data may be connected in practice: personal data (for instance data outlining an individual's clinical history, or the history of his or her online purchases) can be processed to learn an algorithmic model, but they can also be used as inputs for the same or other algorithmic models (e.g., to predict additional health issues, or further purchases).

3.2.3. Article 5(1)(c) GDPR: Data minimisation

Article 5(1)(c) states the principle of data minimisation, according to which personal data should be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.' The principle of minimisation is also contained in Recital 78, requiring the 'minimisation of personal data' as an organisational measure for data protection by design and by default.

There is a tension between the principle of minimisation and the very idea of big data and data analytics, which involves using AI and statistical methods to discover new unexpected correlations in vast datasets. This tension may be reduced by the following considerations.

First, the idea of minimisation should be linked to an idea of proportionality. Minimisation does not exclude the inclusion of additional personal data in a processing, as long as the addition of such data provides a benefit, relatively to the purposes of the processing that outweigh the additional risks for the data subjects. Even the utility of future processing may justify retaining the data, as long as adequate security measures are in place. In particular, pseudonymisation, in combination with other security measures, may contribute to limit risks and increase therefore the compatibility of retention with minimisation.

Second, the processing of personal data for merely statistical purposes may be subject to looser minimisation requirements. In such a case the data subjects' information is considered only as an input to a training set (or a statistical database) and is not used for predictions or decisions concerning individuals. This is stated in Recital (162) which links statistical processing to the objective of producing statistical surveys or results:

> Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose.

Thus, the processing of personal data for statistical purposes should not deliver personal data as its final result. In particular, the personal data processed for statistical purpose should not be used for adopting decisions on individuals.

> The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.

Since the data subject is not individually affected by statistical processing, the proportionality assessment, as far as data protection is concerned, concerns the comparison between the (legitimate) interest in obtaining the statistical results, and the risks of the data being misused for non-statistical purposes.

It is true that the results of statistical processing can affect the collective interests of the data subjects who share the factors that are correlated to certain inferences (e.g., the individuals whose live style and activities are correlated to certain pathologies, certain psychological attitudes, or certain market preferences or political views). The availability of this correlation exposes all members of the group – as soon as their membership in the group is known – to such inferences. However, as long as the correlation is not meant to be applied to particular individuals, on the basis of data concerning such individual (data determining its belonging to the group) statistical processing remains outside of data protection. On the contrary, the information used to ascribe a person to a group and the person's ascription to that group are personal data, and so are the consequentially inferred data concerning that person. This idea is expressed in at footnote 5 in the 2017 Council of Europe Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data

> personal data are also any information used to single out people from data sets, to take decisions affecting them on the basis of group profiling information.

Thus, neither in the GDPR nor in the in Guidelines can we yet find an explicit endorsement of group privacy as an aspect of data protection. On the contrary, the need to take into account group privacy has been advocated by many scholars.[84] However, as we shall see in the following, a preventive risk-management approach can contribute to the protection of group privacy also in the context of GPDR.

3.2.4. Article 5(1)(d) GDPR: Accuracy

The principle of accuracy is stated in Article 5(1) GDPR that requires data to be 'accurate and, where necessary, kept up to date,' and that initiatives are taken to address inaccuracies:

> 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.

This principle also applies to personal data that are used as an input for AI system, particularly when personal data are used to make inferences or decisions about data subjects. Inaccurate data may expose data subjects to harm, whenever they are considered and treated in ways that do not fit their identity.

With regard to machine learning systems, we need to distinguish whether personal data are used only in a training set, to learn general statistical correlations, or rather as input to a profiling algorithm. Obviously, once that the data are available for the training set, the temptation to use the same data to make also individualised inferences will be very strong. Anonymisation, or pseudonymisation, with strong security measures can contribute to reducing the risk

3.2.5. Article 5(1)(e) GDPR: Storage limitation

The principle of storage limitation is stated in GDPR at Article 5(1)(e), which prohibits to keep personal data when they are no longer needed for the purposes of the processing.

[84] On the Guidelines, see Mantelero (2017).

> [Personal data should be] kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

Longer storage is however allowed for archiving, research, or statistical purposes.

> [P]ersonal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');

There is undoubtable tension between the AI-based processing of large sets of personal data and the principle of storage limitation. This tension can be limited to the extent that the data are used for statistical purposes, and appropriate measures are adopted at national level, as discussed above in 3.2.3.

3.3. AI and legal bases

Article 6 GDPR states that all processing of personal data requires a legal basis. This idea was first introduced in the 1995 Data Protection Directive, and was subsequently constitutionalised in Article 8 of the European Charter of Fundamental Rights, according to which personal data 'must be processed […] on the basis of the consent of the person concerned or some other legitimate basis laid down by law.'

The processing of personal data in the context of AI application raises some issues relating to the existence of a valid legal basis. To determine when a legal basis may support AI-based processing, we need to separately consider the legal bases set forth in Article 6 GDPR, which states that the processing of personal data only is lawful under the following conditions: (a) consent of the data subject, or necessity (b) for performing or entering into a contract, (c) for complying with a legal obligation, (d) for protecting vital interests (e) for performing a task in the public interest or in the exercise of public authority, or (f) for a legitimate interest.

3.3.1. Article 6(1)(a) GDPR: Consent

A data subject's consent to the processing of his or her personal data by an AI system can have two possibly concurring objects: including such data in a training set, or providing them to an algorithmic model meant to deliver individualised responses. Usually, the data subject's consent covers both. As noted in Section 3.1.3, consent has to be specific, granular and free. It is not easy for all these conditions to be satisfied with regard to the AI-based processing of personal data. Thus, this processing usually needs to rely alternatively or additionally on other legal bases.

The processing of personal data for scientific or statistical purposes may be based on the social significance of such purposes (Article 6(1)(f)), beside the endorsement of such purposes by the data subject. Consent to individual profiling may concur with the necessity or usefulness of such processing for the purposes indicated in the subsequent items of Article 6.

3.3.2. Article 6(1)(b-e) GDPR: Necessity

The legal bases from (b) to (e) can be treated together here since they all involve establishing the necessity of the processing for a certain aim: (b) performing or entering (at the request of the data subject) into a contract, (c) for complying with a legal obligation, (d) protecting vital interests (e) performing a task in the public interest or in the exercise of public authority. Thus, such legal bases

do not apply to the AI-based processing that is subsequent to or independent of such aims in the specific case at hand.

For instance, the necessity of using personal data for performing or entering a particular contract does not cover the subsequent use of such data for purposes of business analytics. Similarly, this legal basis does not cover the subsequent use of contract data as input to a predictive-decisional model concerning the data subject, even when the data are used for offering a different contract to the same person. Assume, for instance that the data subject's health data are necessary for performing an insurance contract with the data subject. This necessity would not cover to the use of the same data for offering a new contract to the same data subject, unless the data subject has requested to be considered for a new contract, i.e., unless the data are necessary 'in order to take steps at the request of the data subject prior to entering into a contract' (Article 6(b)).

3.3.3. Article 6(1)(f) GDPR: Legitimate interest

Article 6(1)(f) provides a general legal basis to the processing of personal data, namely, the necessity of the processing

> for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data,

We may wonder to what extent Article 6(1)(f) may apply to the AI-processing of personal data.[85] We have to distinguish the use of personal data in a training set to build/learn an algorithmic model, and their use as an input to a given algorithmic model. In the first case, as long as strong security measures are adopted it – which usually should involve pseudonymisation of the data, and their anonymisation as soon as the model has been completed – seems that the data subject's interests are not severely affected. If the controller is pursuing an interest that is permissible under the law (including an economic interests), it seems that the standard set forth in Article 6(1)(f) could be met.

The situation is much different when the data subjects' data are used in an algorithmic model, to derive conclusions concerning the data subject. Under such a case, the interest of the data subject should be given priority, according to his or her assessment. Thus, the data subject should be asked for his or her consent and have the opportunity to opt out.

The legitimate interest test may be important to address the admissibility of those applications that may seriously affect individuals and society, even when they are technologically sound and non-discriminatory. When an application provides benefits that are outweighed by the disadvantages imposed on the data subjects, we should conclude that the application fails to have a basis according to Article 6(1)(f). This may be the case, as noted above, for systems meant to detect individuals' attitudes from faces, or also to assess workers' performance based on pervasive surveillance, or to detect and influence political views, etc. In all such instances, given the difference in knowledge and power and lack of adequate information, consent by the data subject would not meet the requirement of freedom and information in the GDPR, and thus could not provide an alternative legal basis. Thus, the processing should be considered to be unlawful.

A limitation of the scope of Article 6(1)(f) may consist in the fact that it seems to adopt individualistic perspective, as it only requires a balance between the interests of controllers and on data subject, without taking into accounts broader interests, pertaining to groups or even to society as a whole. However, this limitation of the scope of the balancing test according to Article 6(1)(f) may have a reason, since the assessment of the social merit of a processing operation, and the decision to

[85] On legitimate interest, see Kamara and De Hert (2019).

outlaw it based on this assessment, should be adopted on the basis of on a wide debate, and according to the determination or at least to the directions, of politically responsible bodies.

3.3.4. Article 6(4) GDPR: Repurposing

A key issue concerning AI applications pertains to repurposing of personal data. This is an issue on which the provision of the GDPR are unclear. The general idea is stated Article 5(1)(b) as an articulation of the principle of purpose limitation. Personal data shall be 'not further processed in a manner that is incompatible' with the original purposes. The prohibition of repurposing is also affirmed Recital 50, according to which the further processing of personal data for new purposes is only allowed when it is compatible with the original purposes:

> The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected.

Compatibility is however presumed, according to 5(1)(b) when the further processing is meant to serve purposes pertaining to archiving, scientific or historical research or statistics:

> further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes

Compatibility is also presumed when the new processing is based on a law, for reasons of public interest:

> If the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, Union or Member State law may determine and specify the tasks and purposes for which the further processing should be regarded as compatible and lawful.

Article 6(4) specifies that the law allowing for repurposing 'constitutes a necessary and proportionate measure in a democratic society' and that compatibility is established (or substituted) by the data subject's consent. It also spells out possible factors to be taken into account to determine compatibility:

> 4. 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 Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard the objectives referred to in Article 23(1), 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.

The issues of the admissibility of processing personal data for new and different purposes has become crucial in the era of AI and big data, when vast and diverse masses of data are available and artificial intelligence or statistical methods are then deployed to discover correlations and identify possible causal links. As noted above this may lead to the discovery of unexpected connections based on the combination of disparate sets of data (e.g., connections between lifestyle preferences in social networks and health conditions, between consumer behaviour and market trends, between internet queries and the spread of diseases, between internet likes and political preferences, etc.). The results of these analyses (e.g., correlations discovered between consumers' data and their preferences, spending capacities and purchasing propensities, etc.) can then be used to assess or influence individual behaviour (e.g., by sending targeted advertisements).

Repurposing is key in the domain of big data and AI, since the construction of big data sets often involves merging data that had been separately collected for different purposes, and processing such data to address issues that were not contemplated at the time of collection. A key issue for the future of the GDPR pertains to the extent to which the compatibility test will enable us to draw a sensible distinction between admissible and inadmissible reuses of the data for the purposes of analytics.

Recital (50) does not help us much in addressing this issue, since it seems to indicate that no legal basis is required for compatible repurposing: 'where the processing is compatible with the purposes for which the personal data were initially collected […] no legal basis separate from that which allowed the collection of the personal data is required.' Moreover, Recital (50) seems to presume that all processing for statistical purposes is admissible, by affirming that 'further processing for … statistical purposes should be considered to be compatible lawful processing operations.' This presumption has been limited by the Article 29 WP, who has argued that compatibility must be checked also in the case of statistical processing.

In conclusion, it seems that two requirements are needed for repurposing to be permissible: (a) the new processing must be compatible with the purpose for which the data were collected, and (b) the new processing must have a legal basis (that may be, but is not necessarily, the same of the original processing). Following Recital (50) it seems that statistical processing should be presumed to be compatible, unless reasons for incompatibility appear to exist.

By applying these criteria to the AI-based reuse of data, we must distinguish whether the data are reused for statistical purposes or rather for profiling. Reuse for a merely statistical purpose should in general be acceptable since it does affect individually the data subject, and thus it should be compatible with the original processing. If the statistical processing is directed towards a permissible goal, such as security or market research, it can also rely on the legal basis of Article 6(1)(f), i.e., on its necessity for achieving purposes pertaining to legitimate interests.

Different would be the case for profiling. In such a case, the compatibility assessment is much more uncertain. It should lead to a negative outcome whenever AI-based predictions or decisions may affect the data subject in a way that negatively reverberates on the original purpose of the processing. Consider, for instance, the case in which a person's data collected for medical purpose are inputted to an algorithmic model that determines an insurance price for that person.

It has been argued that the possibility to repurpose personal data for statistical processing is very important for European economy, since European companies need to extract information on markets and social trends – as US and Asian companies do – in order to be competitive.[86] The use of personal data for merely statistical purposes should enable companies to obtain the information

[86] On statistical uses and big data, see Mayer-Schonberger and Padova 2016)

they need without interfering with the data subjects rights. In fact, as we noted above, according to Recital (162) the processing remains statistical only as long as the result the processing

> is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.

3.3.5. Article 9 GDPR: AI and special categories of data

Article 9 GDPR addresses the so-called sensitive data, namely those personal data whose processing may affect to a larger extent the data subjects, exposing them to severe risks. In this regard AI presents some specific challenges.

The first challenge is connected to re-identifiability. As noted in Section 3.1.1, thanks to AI and big data, pieces of data that apparently are unidentified, not being linked to a specific individual, may be re-identified, and reconnected to the individuals concerned. The re-identification of sensitive data may have serious consequences for the data subject. Consider for instance the case in which de-identified medical records that have been made accessible to the public are re-identified at a later stage, so that the public comes to know the medical conditions of the individuals concerned.

The second challenge is connected to inference. Thanks to AI and big data, it may be possible to link observable behaviour and known features of individuals – online activity, purchases, likes, movements – to non-observable sensitive data on them such as their psychological attitudes, their health condition their sexual orientation, or their political preferences. Such inferences may expose the concerned individuals to discrimination or manipulation.

3.4. AI and transparency

The complexity of AI-based processing, and the fact that such processing cannot be completely anticipated, especially when based on machine learning, makes it particularly difficult to ensure transparency. The issue of transparency can come up at two points in time, when a data subject's information is inputted in an information system that includes AI algorithms (ex-ante transparency), or after the system's algorithmic model has been applied to the data subject, to deliver specific outcomes concerning his or her (ex-post transparency).

3.4.1. Articles 13 and 14 GDPR: Information duties

Transparency at the stage in which personal data are collected or repurposed is addressed in Articles 13 and 14 GDPR, which require that the data subject be informed about

> the purposes of the processing for which the personal data are intended as well as the legal basis for the processing.

Information must also be provided about 'the legitimate interests pursued by the controller or by a third party' where the processing is based on legitimate interest (Article 6(1)(f)). When the data are processed for purposes that could not be foreseen at the time the data were collected – as it is often the case with machine learning applications– the information has to be provided before the new processing, as specified in Article 13(3) and 14(4):

> Where the controller intends to further process the personal data for a purpose other than that for which the personal data were collected, the controller shall provide the data subject prior to that further processing with information on that other purpose and with any relevant further information

The obligation to inform the data subject is waved when compliance is impossible, requires a disproportionate effort or impairs the achievement of the objective of the processing (Article 14(5)(b)):

> [The obligation to provide information to the data subject does not apply when] the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, subject to the conditions and safeguards referred to in Article 89(1) or in so far as the obligation referred to in paragraph 1 of this Article is likely to render impossible or seriously impair the achievement of the objectives of that processing. In such cases the controller shall take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available.

This limitation only applies when the data have not been collected from the data subject. It is hard to understand why this is the case. In fact, the reasons that justify an exception to the information obligation when the data were not obtained from the data subject, should also justify the same exception when the data were collected from him or her.

3.4.2. Information on automated decision-making

Article 13(2)(f) and 14(2)(g) GDPR address a key aspect of AI applications, i.e. automated decision-making. The controller has the obligation to provide:

(a) information on 'the existence of automated decision-making, including profiling, referred to in Article 22(1)' and

(b) '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.'

This provision has been at the centre of a vast debate in the research community, where this legal requirement has been related to the more general, and indeed fundamental issue of explaining AI systems and their outcomes. Indeed, according to the AI4People document,[87] explainability (or explicability) is indeed one of the principles that should inspire the development of AI, along with beneficence, non-maleficence, autonomy and justice. In the current discussion on explainability different perspectives have been put forward.

Computer scientists have focused on the technological possibility of providing understandable models of opaque AI systems (and, in particular, of deep neural networks), i.e., model of the functioning of such systems that can be mastered by human experts. For instance, the following kinds of explanations are at the core of current research on explainable AI:[88]

- *Model explanation,* i.e., the global explanation of an opaque AI system through an interpretable and transparent model that fully captures the logic of the opaque system. This would be obtained for instance, if a decision tree or a set of rules was provided, whose activation exactly (or almost exactly) reproduces the functioning of a neural network.
- *Model inspection,* i.e., a representation that makes it possible to understanding of some specific properties of an opaque model or of its predictions. It may concern the patterns of activation in the system's neural networks, or the system's sensitivity to changes in its input factors (e.g. how a change in the applicant's revenue or age makes a difference in the grant of a loan application).

[87] Floridi et al (2018).

[88] Guidotti et al (2019).

- *Outcome explanation*, i.e., an account of the outcome of an opaque AI in a particular instance. For instance, a special decision concerning an individual can be explained by listing the choices that lead to that conclusions in a decision tree (e.g., the loan was denied because of the applicant's income fell below a certain threshold, his age above a certain threshold, and he did not have enough ownership interest in any real estate available as collateral).

The explanatory techniques and models developed within computer science are intended for technological experts and assume ample access to the system being explained.

Social scientists, on the contrary have focused on the objective of making explanations accessible to lay people, thus addressing the communicative and dialectical dimensions of explanations. For instance, it has been argued that the following approaches are needed.[89]

- Contrastive explanation: specifying what input values made a difference, determining the adoption of a certain decision (e.g., refusing a loan) rather than possible alternatives (granting the loan);
- Selective explanation: focusing on those factors that are most relevant according to human judgement;
- Causal explanation: focusing on causes, rather than on merely statistical correlations (e.g., a refusal of a loan can be causally explained by the financial situation of the applicant, not by the kind of Facebook activity that is common for unreliable borrowers);
- Social explanation: adopting an interactive and conversational approach in which information is tailored according to the recipient's beliefs and comprehension capacities.

While the latter suggestions are useful for the ex-post explanation of specific decisions by a system, they cannot be easily applied ex-ante, at the time of data collection (or repurposing). At that time – i.e., before the user's data are inputted either in the training algorithm, or in the prediction algorithm (using the algorithmic model) – what can be provided to the user is just an indication on the system's general functioning. At this stage, the user should ideally be provided with the following information:

- The input data that the system takes into consideration (e.g., for a loan application, the applicant's income, gender, assets, job, etc.), and whether different data items are favouring or rather disfavouring the outcome that the applicant hopes for;
- The target values that the system is meant to compute (e.g., a level of creditworthiness, and possibly the threshold to be reached in order for the loan to be approved);
- The envisaged consequence of the automated assessment/decision (e.g., the approval or denial of the loan application).

It may also be useful to specify what are the overall purposes that the system is aimed to achieve. In the current practice the information that is provided about AI applications is quite scanty, even when profiling is involved. For example, Airbnb explains its profiling practice by asserting that it will:

> conduct profiling on your characteristics and preferences (based on the information you provide to us, your interactions with the Airbnb Platform, information obtained

[89] Miller (2019). Mittelstadt and Wachter (2019).

> from third parties, and your search and booking history) to send you promotional messages, marketing, advertising and other information that we think may be of interest to you.

The data subject would benefit from more precise and relevant information, especially when important decisions are at stake. In particular, with regard to complex AI systems, the possibility of providing modular information should be explored, i.e., providing bullet points that laypeople can understand, with links to access more detailed information possibly covering technical aspects.

However, it is unlikely that the information that is provided to the general public will be sufficient to gain an understanding that is sufficient for identifying potential problems, dysfunctions, unfairness. This would assume access to the algorithmic model, or at least the possibly of subjecting it to extensive testing, and in the case of machine learning approaches, access to the system's training set.

It has been argued that it would important to enable citizen to engage in 'black box tinkering', i.e., on a limited reverse-engineering exercise that consists in submitting test cases to a system and analysing the system's responses to detect faults and biases.[90] This approach, which involves a distributed and non-systematic attempt at sensitivity analysis, has the advantage of democratising controls but is likely to have a limited success given the complexity of AI applications and the limitations on access to them.

3.5. AI and data subjects' rights

AI is relevant to distinct data protection rights. The GDPR expressly refers to profiling and automated decision-making in connection with the rights to access and the right to object, but AI also raises specific issues relative to other rights such as in particular, the rights to erasure and portability.

3.5.1. Article 15 GDPR: The right to access

A key aspect of transparency (and consequently of accountability) consist in the data subjects' rights to access information about the processing of their data. Data subjects, according to Article 15 GDPR, 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 […] information' [about their processing].

Article 15(1)(f) specifically addresses automated decision-making, requiring the controller to provide, when requested by the data subject, the same information that should have been provided before starting the processing according to 13(2)(f) and 14(2)(g). The mandatory information concerns

> the existence of automated decision-making' and 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

The right to access information is also addressed in Recital 63. The recital first states that the right of access includes the data subject's right to know

> where possible [...] the logic involved in any automatic personal data processing and, at least when based on profiling, the consequences of such processing.

[90] Perel and Elkin-Koren (2017).

The scope of the right to access, or the ways of implementing it are limited by the requirement that but it

> should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software.

This limitation, however, should not entail a complete denial of the right to information:

> [T]he result of these considerations should not be a refusal to provide all information to the data subject. Where the controller processes a large quantity of information concerning the data subject, the controller should be able to request that, before the information is delivered, the data subject specify the information or processing activities to which the request relates'

There has been a wide discussion on whether Article 15 should be read as granting data subjects the right to obtain an individualised explanation of automated assessments and decisions.[91] Unfortunately, the formulation of Article 15 is very ambiguous, and that ambiguity is reflected in Recital 63. In particular it is not specified whether the obligation to provide information on the 'logic involved' only concerns providing general information on the methods adopted in the system, or rather specific information on how these methods where applied to the data subject (i.e., an individual explanation, as we shall see in Section 3.6.5).

3.5.2. Article 17 GDPR: The right to erasure

The right to erasure (or to be forgotten) consists in the data subjects' right to 'obtain from the controller the erasure of personal data concerning him or her without undue delay ', when the conditions for lawful processing no longer obtain (such conditions are forth in Article 17 (1)). An issue may concern whether even inferred personal data or also inferred group data (such as a trained algorithmic model) should be deleted as a consequence of the obligation to erase the collected personal data that have enabled such inferences to be drawn. The answer seems positive in the first case and negative in the second, since the data that are embedded in an algorithmic model are no longer personal. However, erasing the data used for constructing an algorithmic model, may make it difficult or impossible to demonstrate the correctness of that model.

3.5.3. Article 19 GDPR: The right to portability

The data subject has 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 'to transfer the data to other controller'. This right only applies when the processing is based on consent. Thus, the right to portability has a smaller scope that the right to access, which applies to all processing personal data, regardless of the applicable legal basis.

It is not easy to determine the scope of this right with regard to AI-based processing. First, it needs to be determined whether the data 'provided' by the data subject only concern the data entered by the data subject (e.g., keying his or her particulars) or also the data collected by the system when tracking the data subject's activity. Second, it is to be determined whether the right also concerns the data inferred from the collected data about the data subject. A clarification would be useful in this regard.

3.5.4. Article 21 (1): The right to object

The right to object enables data subjects to request (and obtain) that the processing of their data be terminated. This right can be exercised under the following conditions:

[91] Wachter et al (2016), Edwards and Veale (2019).

1. The data subject has grounds relating to his or her particular situation that support the request.
2. The processing is based on the legal basis of Article 6 (3)(e), i.e. necessity of the processing for performing a public task in the public interest or for the exercise of legitimate authority, or on the legal basis of Article 6 (3)(f), i.e., necessity of the processing for purposes of the legitimate interests pursued by the controller or by a third party.
3. The controller fails to demonstrate compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject.

If all these conditions are satisfied, the controller has the obligation to terminate the processing.

The right to object is particularly significant with regard to profiling, since it seems that only in very special cases the controller may have overriding compelling legitimate grounds for continuing to profile a data subject which objects to the profiling on personal grounds.

The right to object does not apply to a processing that is based on the data subject's consent, since in this case the data subject can impede the continuation of the processing just by withdrawing consent (according to Article 7 (3) GDPR).

The GDPR, in regulating the right to object, explicitly refers to profiling, and introduces special norms concerning direct marketing and statistical processings. Such provisions are relevant to AI, given that profiling and statistics are indeed key applications of AI to personal data.

3.5.5. Article 21 (1) and (2): Objecting to profiling and direct marketing

Article 21 (1) specifies that the right to object also applies to profiling:

> 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.

Profiling in the context of direct marketing is addressed in Article 21 (2), which recognises an unconditioned right to object:

> 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.

This means that the data subject does not need to invoke specific grounds when objecting to processing for direct marketing purposes, and that such purposes cannot be 'compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject'.

Given the importance of profiling for marketing purposes, the unconditional right to object to such processing is particularly significant for the self-protection of data subjects. Controllers should be required to provide easy, intuitive and standardised ways to facilitate the exercise of this right.

3.5.6. Article 21 (2). Objecting to processing for research and statistical purposes

The right to object also applies to processing for scientific or historical research purposes and for statistical purposes. In such cases, the objection concerns the inclusion of the data subject information in the input data for the processings at stake (as the result of research and statistics

cannot consist in personal data). The right to object does not apply when the processing is carried out for reasons of public interest (it therefore applies, *a contrario*, when the processing is aimed at private commercial purposes):

> Where personal data are processed for scientific or historical research purposes or statistical purposes pursuant to Article 89(1), the data subject, on grounds relating to his or her particular situation, shall have the right to object to processing of personal data concerning him or her, unless the processing is necessary for the performance of a task carried out for reasons of public interest.

A further limitation is introduced by Article 17(3)(d), which limits the right to erasure when its exercise would make it impossible or would seriously undercut the ability to achieve the objectives of the processing for archiving, research or statistical purposes. This limitation would probably find limited application to big data, since the exclusion of a single records from the processing would likely have little impact on the system's training or, at any rate, on the definition of its algorithmic model.

3.6. Automated decision-making

Article 22, which deals with automated decision-making, is most relevant to AI. As we shall see in what follows, this provision combines a general prohibition on automated decision-making, with broad exceptions.

3.6.1. Article 22(1) GDPR: The prohibition of automated decisions

The first paragraph of Article 22 provides for a general right not to be subject to completely automated decisions significantly affecting the data subject:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Even though this provision refers to a right, it does not provide for a right to object to automated decision-making, namely, it does not assume that automated decision-making is in general permissible as long as the data subject does not object to it. It rather introduces a prohibition upon controllers: automated decisions affecting data subjects are prohibited, unless they fit in one of the exceptions provided in paragraph 2.[92] According to the Article 29 Working Party:

> as a rule, there is a general prohibition on fully automated individual decision-making, including profiling that has a legal or similarly significant effect.[93]

For the application of the prohibition established by Article 22(1), four conditions are needed: a decision must be taken, (2) it must be solely based on automated processing, (3) it must include profiling, (4) it must have legal or anyway significant effect.

The first condition requires that a stance be taken toward a person, and that this stance is likely to be acted upon (as when assigning a credit score).

The second condition requires that humans do not exercise any real influence on the outcome of a decision-making process, even though the final decision is formally ascribed to a person. This condition is not satisfied when the system is only used as a decision-support tool for human beings,

[92] Mendoza and Bygrave (2017).

[93] Article 29, WP251/2017 last revised 2018, 19.

who are responsible for the decision, deliberate on the merit of each case, and autonomously decide whether to accept or reject the system's suggestions.[94]

The third condition requires that the automated processing determining the decision includes profiling. A different interpretation could be suggested by the comma that separates 'processing' and 'including profiling' in Article 22(1), which seems to indicate that profiling only is an optional component of the kind of automated decisions that are in principle prohibited by Article 22(1). However, the first interpretation (the necessity of profiling) is confirmed by Recital (71), according to which the processing at stake in the regulation of automated decision must include profiling:

> Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements.

The fourth condition requires that the decision

> produces legal effects concerning [the data subject] or similarly significantly affects him or her.

Recital (71) mentions the following examples of decision having significant effects: the 'automatic refusal of an online credit application or e-recruiting practices'.[95] It has been argued that such effects cannot be merely emotional, and that usually they are not caused by targeted advertising, unless 'advertising involves blatantly unfair discrimination in the form of web-lining and the discrimination has non-trivial economic consequences (e.g., the data subject must pay a substantially higher price for goods or services than other persons).'[96]

Many decisions made today by AI systems fall under the scope of Article 21(1), as AI algorithms are increasingly deployed in recruitment, lending, access to insurance, health services, social security, education, etc. The use of AI makes it more likely that a decision will be based 'solely' on automated processing. This is due to the fact that humans may not have access to all the information that is used by AI systems, and may not have the ability to analyse and review the way in which this information is used. It may be impossible, or it may take an excessive effort to carry out an effective review – unless the system has been effectively engineered for transparency, which in some cases may be beyond the state of the art. Thus, especially when a large-scale opaque system is deployed, humans are likely to merely execute the automated suggestions by AI, even when they are formally in charge. Moreover, human intervention may be prevented by the costs-and-incentives structure in place: humans are likely not to substantially review automated decision, when the cost of engaging in the review – from an individual or an institutional perspective– exceeds the significance of the decision (according to the decision-maker's perspective).

3.6.2. Article 22(2) GDPR: Exceptions to the prohibition of 22(1)

Paragraph 2 of Article 22 provides for three broad exceptions to Paragraph 1. It states that the prohibition on automated decision-making does not apply when the processing upon which the decision is based

> a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

[94] Article 29, WP251/2017 last revised 2018, 21-22.

[95] For an analysis of legal effects and of similarly relevant effects, see Article 29, WP251/2017 last revised 2018,

[96] Medoza and Bygrave (2017, 89).

b) is authorised by Union or 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.

Based on the broad exception of item (a), automated decision-making is enabled in key areas such as recruitment and lending. However, for the exception to apply, decisions based solely on automated processing must be 'necessary.' Such necessity may depend on the high number of cases to be examined (e.g., a very high number of applications to a job). The necessity of using AI in decision-making may also be connected to AI capacities to outperform human judgement. In this connection we may wonder whether human involvement will still contribute to a stronger protection of data subjects, or whether the better performance of machines – even with regard to the political and legal values at stake, e.g., ensuring 'fair equality of opportunity' for all applicants to a position[97] – will make human intervention redundant or dysfunctional. Outside of the domain of contract and legal authorisation, consent may provide a basis for automated decision-making according to Article 22(2)(c). However, the conditions for valid consent not always obtain, even in cases when automated decision-making seems appropriate. Consider for instance the case in which an NGO uses an automated method for classifying (profiling) applicants to determine their need and consequently allocate certain benefits to them. In such a case, it is very doubtful that an applicant's consent may be viewed as free (as not consenting would entail being excluded from the benefit), but the system seems socially acceptable and beneficial even so.

3.6.3. Article 22(3) GDPR: Safeguard measures

In the cases under Article 22(2)(a) and (c) – i.e. when the automated decision is necessary to contract or explicitly consented – Article 22(3) requires suitable safeguard measures:

> 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.

According to Article 29 Working Party, some of these measures concern risk reduction, Examples are quality assurance checks, algorithmic auditing, data minimisation, and anonymisation or pseudonymisation, and certification mechanisms.[98] Such measures should ensure that the requirements set forth in Recital (71) – concerning acceptability, accuracy and reliability – are respected

> the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, 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

According to the Article 29 Working party, the input data must be shown to not be 'inaccurate or irrelevant, or taken out of context,' and to not violate 'the reasonable expectations of the data subjects', in relation to the purpose for which the data was collected.[99] In approaches based on

[97] Rawls ([1971 1999, 63).

[98] Article 29, WP251/2017 last revised 2018, 32

[99] Article 29, WP251/2017 last revised 2018, 17

machine learning, this should apply not only to the data concerning the person involved in a particular decision, but also to the data in a training set, where the biases built into the training set may affect the learned algorithmic model, and hence the accuracy the system's inferences.

Other measures pertain to the interaction with the data subjects, such the right to obtain human intervention and the right to challenge a decision. For instance, a link could be provided to 'an appeals process at the point the automated decision is delivered to the data subject, with agreed time scales for the review and a named contact point for any queries.'[100] An appeals process is most significant with regard to AI applications, and especially when these applications are 'opaque', i.e., they are unable to provide human-understandable explanations and justifications.

3.6.4. Article 22(4) GDPR: Automated decision-making and sensitive data

Article 22(4) introduces a prohibition, limited by an exception, to ground automated decisions on 'sensitive data', i.e., the special categories set out in Article 9(1):

> Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place

The exception concerns the cases in which the data subject has given explicit consent (Article 9(2)(a)) or processing is necessary for reason of public interest (Article 9(2)(g)). The role of the data subject's consent needs to be clarified since consent does not exclude that the method used for the decision is inacceptable (as when it is discriminatory).

As noted above AI challenges the prohibition of processing sensitive data. First of all, sensitive data can be (probabilistically) inferred from non-sensitive data. For instance, sex orientation can be inferred from a data subject's Internet activity, likes or even facial features. In this case. the inference of sensitive data should count as a processing of sensitive data, and therefore would have to be considered unlawful unless the conditions under Article 9 are met.

Secondly, non-sensitive data can work as proxies for sensitive data correlated to them, even though the latter are not inferred by the system. For instance, the place of residence can act as a proxy for ethnicity. In this case, an unlawful discrimination may take place.

3.6.5. A right to explanation?

To understand the GDPR ambiguous approach to the right to explanation we need to compare two provisions, Recital (71) and Article 22.

According to Recital (71), the safeguards to be provided to data subjects in case of automated decisions include all of the following:

- specific information
- the right to obtain human intervention,
- the right to express his or her point of view,
- the right to obtain an explanation of the decision reached after such assessment
- the right to challenge the decision.

According to Article 22 the suitable safeguards to be provided include 'at least'

[100] Article 29, WP251/2017 last revised 2018, 32

- the right to obtain human intervention,
- the right to express his or her point of view,
- the right to challenge the decision.

Thus, two items are missing in article 22 relative to Recital (71): the provision of 'specific information' and the right to *obtain an explanation of the decision reached after such assessment*'. The first omission may not be very significative, since the obligation to provide information is already established by articles 13, 14 and 15 GDPR, as noted above, even though the requirement that the information be 'specific' is only spelled out in Recital (71). The second omission raises the issue of whether controllers are really required by law to provide an individualised explanation. Two interpretations are possible.

According to the first one, the European legislator, by only including the request for specific explanation in the recitals and omitting it from the articles of the GDPR, intended to convey a double message: to exclude an enforceable legal obligation to provide individual explanations, while recommending that data controllers provide such explanations when convenient, according to their discretionary determinations. Following this interpretation, providing individualised explanation would only be a good practice, and not a legally enforceable requirement.

According to the second interpretation, the European legislator intended on the contrary to establish an enforceable legal obligation to provide individual explanation, though without unduly burdening controllers. This interpretation is hinted at by the qualifier 'at least', which precedes the reference made to a 'right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.' The qualifier seems to suggest that some providers are legally required to adopt further safeguards, possibly including individualised explanations, as indicated in Recital 71. On this second approach, an explanation would be legally needed, whenever it is practically possible, i.e., whenever it is compatible with technologies, costs, and business practices.

Both readings of these provisions – the combination of Article 13, 14, 15 and 22 – seems possible. The reason for this ambiguous language is likely to be that the legislator was unsure as to whether individualised explanations should be made into a legal requirement. As noted by some commentators, the view that data subjects have a right to individualised explanations under the GDPR may in the future be endorsed by data protection authorities and courts, perhaps viewing individualised explanation as a precondition for the data subjects' ability to effectively contest automated decisions.

> A broad reading of Article 22(3), according to which an explanation is required to contest a decision, would strengthen the right to contest. In this case, the argument for a right to explanation of specific decisions could be further buttressed by drawing on the rights to fair trial and effective remedy enshrined in Articles 6 and 13 of the European Convention on Human Rights.[101]

However, we should be cautioned against overemphasising a right to individualised explanations as a general remedy to the biases, malfunctions, and inappropriate applications of AI and big data technologies.[102] A parallel may be drawn between consent and individualised explanation, as both rely on the data subject's informed initiative. It has often been observed that consent provides no effective protection, given the disparity in knowledge and power between controllers and data subjects, and also the limited time and energy available to the latter, and their inability to pool their

[101] Wachter et al (2016).

[102] Edwards and Veal (2019).

interests and resources and coordinate their activities. The same may also apply to the right to an explanation, which is likely to remain underused by the data subjects, given that they may lack a sufficient understanding of technologies and applicable normative standards. Moreover, even when an explanation elicits potential defects, the data subjects may be unable to obtain a new, more satisfactory decision.

3.6.6. What rights to information and explanation?

Our analysis of the right to information and explanation to data subject end up with puzzling results.

Let us summarise the main references in the GDPR:

- According to Article 13 and 14 (on the right to information and Article 15 (on the right to access), the controller should provide information on 'the existence of automated decision-making, including profiling, referred to in Article 22(1)' and 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject'.
- According to Article 22, the data subject has at least the right to obtain human intervention, the right to express his or her point of view, and the right to challenge the decision.
- According to Recital (71), the data subject should also have the right to obtain an explanation of the decision reached after the assessment of his or her circumstances.

We have also observed that according to the European Data Protection Board, controllers should provide data subject, in simple ways, with the 'rationale behind or the criteria relied on in reaching the decision.' This information should be so comprehensive as to 'enable data subjects to understand the reasons for the decision.'[103]

Finally, Article 7(4)(a) of the Directive on Consumer Rights[104] addresses information to be provided to consumers with regard to online offers, which often are based on profiling. It establishes that the supplier should indicate 'the main parameters determining ranking [...] of offers presented to the consumer' as well as 'the relative importance of those parameters as opposed to other parameters".

Based on this set of norms, the obligation to provide information to the profiled data subject can take very different content:

1. information on the existence of profiling, i.e., on the fact that the data subject will be profiled or is already being profiled;
2. general information on the purposes of the profiling and decision-making;
3. general information on the kind of approach and technology that is adopted;
4. general information on what inputs factors (predictors) and outcomes (targets/predictions), of what categories are being considered;
5. general information on the relative importance of such input factors in determining the outcomes;

[103] Guidelines of the European Data Protection Board of 3 October 2017 on Automated individual decision-making and Profiling, p. 25.

[104] Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, as amended by Directive 2019/2161/EU of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernisation of Union consumer protection rules

6. specific information on what data have been collected about the data subject and used for profiling him or her;
7. specific information on what values for the features of the data subject determined the outcome concerning him or her;
8. specific information on what data have been inferred about the data subject;
9. specific information on the inference process through which certain values for the features of the data subject have determined a certain outcome concerning him or her.

In this list, items from (1) to (5) concern information ex ante, to be provided before the data are collected or anyway processed, while items from (5) to (9) concern information to be provided ex post.

With regard to the ex-ante information, it is sure that the controller is required to provide the information under (1) and (2). Information under (3) may also be required, when the adopted technology makes a relevant difference (e.g., it may be inappropriate or lead to errors and biases). Information under (4) should also be provided, as a minimal account of the 'logic' of the processing, at least relative to the categories into which the input factors can be classified. This idea is explicitly adopted in the California Consumer Privacy Act, which at Section 1798.100 (b) requires controllers to 'inform consumers as to the categories of personal information to be collected.' We may wonder whether also some information under (5) should be provided, as an aspect of the information about the 'logic' of the processing, though it may not easy to determine in the abstract (without reference to a specific case) the importance of a certain input factor.

With regard to the ex-post information, all data under (6) should be provided, as they are the object of the right to access. Information about (7) should also be provided, if we assume that there is right to individualised explanation. An individualised explanation may also require information about (8), when the intermediate conclusions by the system play a decisive role. Finally, information about (9) might also be provided, though information on (7) and (8) should generally be sufficient to provide adequate individualised explanations

The information above needs to be complemented with further information in the case of decisions by public authorities, in which case also a reference to the norms being applied and the powers being exercised is needed, based on principles concerning the required justification for administrative acts.

Given the variety of ways in which automated decision-making can take place, it is hard to specify in precise and general terms what information should be provided. What information the controller may be reasonably required to deliver will indeed depend on the importance of the decision, on the space of discretion that is being used, and on technological feasibility. However, it seems that data subjects who did not obtain the decision they hoped for should be provided with the specific information that most matters to them, namely, with the information on what values for their features determined in their case an unfavourable outcome. The relevant causal factors could possibly be identified by looking at the non-normal values that may explain the outcome. Consider for instance the case of person having an average income, and an ongoing mortgage to repay, whose application for an additional mortgage is rejected. Assume both of the following hypotheticals: (a) if the person had had a much higher income her application would have been accepted, regardless of her ongoing mortgage, and (b) if she had had no ongoing mortgage, her application would have been accepted, given her average income. Under such circumstances, we would say that it was the previous mortgage, rather than the average income, the key reason or

cause explaining why the mortgage application was rejected, since it is what explains the departure from the standard outcome for such a case. [105]

3.7. AI and privacy by design

Two different legal perspective, complementary rather than incompatible, may inspire data protection law, a right based and a risk-based approach. Though the focus of the GDPR is on the right-based approach, there are abundant references to the risk prevention in the GDPR that can be used to address AI-related risks. [106]

3.7.1. Right-based and risk-based approaches to data protection

The right-based approach to data protection, which underlies in particular European law, views data protection as a matter of individual rights. These rights are organised in two layers. The top layer includes the fundamental rights to privacy and data protection, which are synergetic to other fundamental rights and principles: dignity, freedom of thought, conscience and religion, freedom of assembly and association, freedom to choose an occupation and right to engage in work, non-discrimination, etc. The lower tier is constituted by the data protection rights granted to individuals by the GDPR, such as the power to consent and withdraw consent (to processing not having other legal bases), the right to information, access, erasure, and the right to object. The focus is on the harm to individuals and on legal measure empowering their initiatives.

The risk-based approach, rather than granting individual entitlements, focuses on creating a sustainable ecology of information, where harm is prevented by appropriate organisational and technological measures. Data protection, when seen from the latter perspective appears to be as a risk-regulation discipline, similar to environmental protection, food safety, or even the regulation of medical devices or financial markets. In these domains the emphasis in on preventive measures, certification, private and public expertise, and on the way in which not only individuals by also society and groups are affected.

3.7.2. A risk-based approach to AI

With regard to AI, both the right-based and the risk-based approaches are meaningful, but the second is particularly significant. It has been noted that in the US a risk-based approach to data protection has emerged in the public sector. A 'Big Data due progress', [107] has been argued for, which requires agencies to educate officers on biases and fallacies of automation, to appoint hearing officers tasked with reviewing automated decisions, to test regularly computer systems, to ensure that audit trails are kept, etc. [108] For instance, it has been argued that the US Federal Trade Commission should play a key role in ensuring fairness and accuracy of credit scoring systems, given the huge impact that a bad credit score may have on people's life. Other suggested remedies include auditing, noticing consumer, and enabling consumers not only to access their data, but also to test the system by submitting hypotheticals. [109]

[105] On the connection between causal explanations and (ab)normality, see Halpern and Hitchcock (2013)

[106] Edwards and Veal (2019).

[107] Edwards and Veal (2019).

[108] Citron (2008).

[109] Citron and Pasquale (2014).

The GDPR also contains a number of provisions that contribute to prevent the misuse of AI, in particular, in connection with the idea of 'privacy by design and by default', namely, with preventive technological and organisational measures.[110]

A serious issue pertaining to risk-prevention and mitigation measures concerns whether the same measures should be required by all controllers engaging in similar processings or whether a differentiated approach is needed, that takes into account the size of controllers and their financial and technical capacity of adopting the most effective precautions. More precisely, should the same standards be applied both to the Internet giants, which have huge assets and powerful technologies and profit of monopolistic rents, and to small start-ups, which are trying to develop innovative solutions with scanty resources. Possibly a solution to this issue can be found by considering that risk prevention and mitigation measures are the object of best effort obligations, having a stringency that is scalable, depending not only on the seriousness of the risk, but also the capacity of the address of the obligation. Thus, more stringent risk preventions measures may be required to the extent that the controller both causes a more serious social risk, by processing a larger quantity of personal data on larger set of individuals and has superior ability to respond to risk in effective and financially sustainable ways.

3.7.3. Article 24 GDPR: Responsibility of the controller

Article 24, on 'Responsibility of the controller', requires the controller to

> implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation'.

Such measures are to be 'reviewed and updated where necessary.' With regard to AI applications, the measures include controls over the adequacy and completeness of training sets, over reasonableness of the inferences, over the existence of causes of bias and unfairness.

3.7.4. Article 25 GDPR: Data protection by design and by default

Article 25 (1) on 'Data protection by design and by default', specifies that both 'at the time of the determination of the means for processing and at the time of the processing' the controller should

> implement appropriate technical and organisational measures which are designed to implement data-protection principles [...] in an effective manner and to integrate the necessary safeguards into the processing.

Article 25(2) addresses data minimisation. It is relevant to AI and big data applications as it requires the implementation of

> appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed.

Such measures should address 'the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility'. Article 25(2) questions the possibility to retain the data in consideration of future still undetermined purposes, unless the scope the future uses is defined (e.g. scientific or market research).

[110] Edwards and Veal 2019.

3.7.5. Article 35 and 36 GDPR: Data protection impact assessment

Article 35 requires that a data protection impact assessment is preventively carried out relatively to processing that is likely to result in a high risk to the rights and freedoms of natural persons. The assessment is required in particular when the processing involves

> 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.

Thus, an impact assessment is usually required when AI-based profiling contributes to automated decision-making affecting individuals, since such profiling is likely to be 'systematic and extensive.'

When the assessment determines that a processing involves 'high risk', according to Article 36 (1) the controller should preventively ask the supervisory authority (the national data protection authority) for advice.

> The controller shall consult the supervisory authority prior to processing where a data protection impact assessment under Article 35 indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk.

The impact assessment must be shared with the supervisory authority. The authority must provide written advice to the controller where

> the supervisory authority is of the opinion that the intended processing referred to in paragraph 1 would infringe this Regulation, in particular where the controller has insufficiently identified or mitigated the risk.

The authority may also use its investigative and corrective powers. In particular it may (article 50(2)(d)):

> order the controller or processor to bring processing operations into compliance with the provisions of this Regulation

The authority may even temporarily or permanently ban the use of the system (article 50(2)(f)).

Articles 35 and 36 are particularly important to the development of data-protection compliant AI application, since may enable cooperation and mutual learning between data protection authorities and controllers.

3.7.6. Article 37 GDPR: Data protection officers

Article 37 requires controllers to designate a data protection officer when they engage in

> processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale, or when they process on a large-scale sensitive data or data concerning criminal convictions.

This provision is relevant to AI, since various AI-based applications are based on data sets collected by the monitoring the behaviour of data subject (e.g., their online behaviour, or their driving behaviour, etc.). A specialised and impartial internal review would arguably be useful in such cases.

3.7.7. Articles 40-43 GPDR: Codes of conduct and certification

Articles 40-43 address codes of conduct and certification. While these provisions do not make explicit reference to AI, codes and conduct and certification procedure may be highly relevant to AI, given the risks involved in AI application, and the limited guidance provided by legal provisions.

Adherence to codes of conduct and certification mechanisms, according to Articles 24 and 25 may contribute to demonstrate compliance with the obligations of the controller and with the requirements of privacy by design. The idea of a certification for AI applications has been endorsed by the European Economic and Social Committee (EESC) which 'calls for the development of a robust certification system based on test procedures that enable companies to state that their AI systems are reliable and safe.' Thus, it suggests developing a 'European trusted-AI Business Certificate based partly on the assessment list put forward by the High-Level Experts' group on AI.' On the other hand, some perplexities on a general framework for certification have also been raised, based on the complexity of AI technologies, their diversity, and their rapid evolution.[111]

Certification and code of conducts could address both algorithms as such (in particular with regard to their technical quality and accuracy) as well as the context of their application (training sets, input data, intended outcomes and their uses). They could enable sectorial approaches and the rapid adaptation to technological and social changes.

On the other hand, it has been observed that 'voluntary self-or co-regulation by privacy seal has had a bad track record in privacy, with recurring issues around regulatory and stakeholder capture.'[112] Certification and codes of conduct – in combination with the requirement to demonstrate compliance, according to accountability – may lead to formalistic practices, rather than to the real protection of the interests of data subject.[113] Much will depend on the extent to which data protection authorities will supervise the adequacy of these soft law instruments, and the effectiveness of their application.

3.7.8. The role of data protection authorities

As shown in the previous sections, there are various references in the GDPR that support a proactive risk-based approach towards AI and big data. It will be up to the creativity of technological and legal experts, in particular those having the role of data protection officers, to provide adequate solutions. An important role can also be played by data protection authorities, in enforcing data protection law, but also in proposing and promoting appropriate standards. The GDPR makes explicit reference both to National data protection authorities and to the European Data Protection Board, to which is confers an important role.

The European Data Protection Board is the continuation or the Article 29 Working Party, established by the 1995 Data Protection Directive. It includes representatives of the Member States' data protection authorities and of the European data protection supervisors is meant to ensure the consistent application of the Regulation. According to Recital (77) the Board is supposed to provide guidance on the implementation of the GDPR through guidelines:

> Guidance on the implementation of appropriate measures and on the demonstration of compliance by the controller or the processor, especially as regards the identification of the risk related to the processing, their assessment in terms of origin, nature, likelihood and severity, and the identification of best practices to mitigate the risk, could be provided in particular by means of approved codes of conduct, approved

[111] AI Now (2018) report

[112] Edwards and Veal (2019, 80).

[113] Edwards and Veal (2019, 80).

> certifications, guidelines provided by the Board or indications provided by a data protection officer.

The Board is entrusted with the task of determining whether certain processing operations do not involve high risks, and of indicating what measures may be appropriate in such cases:

> The Board may also issue guidelines on processing operations that are considered to be unlikely to result in a high risk to the rights and freedoms of natural persons and indicate what measures may be sufficient in such cases to address such risk.

An explicit reference to automated decision-making is contained in Article 70 (1)(f) GDPR, which lists the tasks of Board. With regard to automated decision-making the Board should

> on its own initiative or, where relevant, at the request of the Commission, issue guidelines, recommendations and best practices […] for further specifying the criteria and conditions for decisions based on profiling pursuant to Article 22(2)

3.8. AI, statistical processing and scientific research

AI and big data provide not only risks but also great opportunities. In particular, they offer new avenues to gain knowledge about nature and society that can be used for beneficial purposes. Consider for instance the huge importance of applying AI to medical data, to improve the accuracy of medical tests, to assess connection between symptoms and pathologies, to analyse the effectiveness of therapies. Similar considerations also concern the AI and big data applications to social and economic data, to better plan and optimise private and public activities. As note in Section 2.3.2, big data analytics can lead to unexpected discoveries, which may result from combining data collected for different purposes. Thus, the traditional principles of data protection, such as data minimisation and purpose limitation are challenged, since they may preclude some useful applications and technological development. The problem is aggravated by the fact that many non-European countries seem to offer normative environments that are more facilitative to the full development and deployment of AI systems.

3.8.1. The concept of statistical processing

It has been argued that the way forward, to enable the use of big data analytics also in Europe is to refer to the discipline for scientific and statistical purposes.[114] In particular, Recital (162) GDPR refers to further EU or National law for the regulation of processing for statistical purposes:

> Union or Member State law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality.

In the same Recital, processing for statistical purposes is positively characterised by the objective of producing statistical surveys and results and negatively by the fact that their outcomes are not used for measures or decisions concerning particular individuals:

> Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the

[114] Mayer-Schonberger and Padova 2016.

> personal data are not used in support of measures or decisions regarding any particular natural person.

As it emerges from this characterisation, the meaning of statistical purpose in the GDPR is not narrowly defined and may be constructed as including not only uses for the public interest, but also by private companies for commercial goals.[115]

3.8.2. Article 5(1)(b) GDPR: Repurposing for research and statistical processing

According to Article 5(1)(b) repurposing data for statistical purposes is in principle admissible, as it will 'not be considered to be incompatible with the initial purposes.' Similarly, at 5(1)(e) data retention limits are relaxed with regard to processing for research and statistical purposes. However, processing for research and statistical purposes requires appropriate safeguards, including in particular pseudonymisation. On the other hand, EU or National law may provide for derogation from the data subjects' rights, when needed to achieve scientific or statistical purposes.

3.8.3. Article 89(1,2) GDPR: Safeguards for research of statistical processing

Statistical processing is addressed in Article 89(1), requiring that appropriate safeguards are adopted for processing for archiving, research or statistical purposes and that in particular that the data be pseudonymised or anonymised when these purposes can be achieved in this manner.

> Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject.

The safeguards are linked to data minimisation, though a reference is made not only to anonymisation but also to pseudonymisation (which does not involve a reduction in the amount of personal data).

> Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.

Finally, Article 89 (2) allows for derogations from certain data subjects' rights – to access (Article 15 GDPR), to rectification (16), to restriction of processing (18), to object (21)– in the case of processing for research or statistical purposes.

> Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

It has been argued that the EU and member States have a strong interest in enabling statistical processing, to support economic and technological development. Thus, they may use the provisions above to enable this processing on a large scale, while establishing the required safeguards and derogations. This would provide the opportunity for an EU approach to data analytics, which is compatible with effective data protection:

[115] Mayer-Schoeberger and Padova 2016, 326-7

> GDPR is making small, but noteworthy steps towards enabling Big Data in Europe. It is a peculiar kind of Big Data, though, that European policymakers are facilitating: one that emphasizes reuse and permits some retention of personal data, but that at the same time remains very cautious when collecting data.[116]

The facilitations for scientific and statistical processing, however, may extend beyond reuse and retention: these kinds of processing may also be justified by legitimate interests according to 6(1)(f), as long as the processing is done in such a way as to duly fulfil the that data subjects' data protection interests, including their interests in not being subject to risks because of unauthorised uses of their data.

A difficult issue concerns whether access to the data sets of personal information supporting statistical inferences (e.g., to predict consumer preferences, or market trends) should be limited to the companies or public bodies who have collected the data. On the one hand, allowing, or even requiring, that the original controllers do not make the data accessible to third parties, may affect competition and prevent beneficial uses of the data. On the other hand, requiring the original controllers to make their data sets available to third parties would cause additional data protection risks.

[116] Mayer-Schoenberger and Padova 2016, 331

4. Policy options: How to reconcile AI-based innovation with individual rights & social values, and ensure the adoption of data protection rules and principles

In this section, the main results of the report will be summarised, pointing out the main conclusions reached and proposing some policy indications.

4.1. AI and personal data

In Section 2 the social and legal issues pertaining to the application of AI to personal data have been discussed. First opportunities and risks have been illustrated, and then the key ethical and legal issues have been considered.

4.1.1. Opportunities and risks

First, the concept of AI has been introduced and the development of AI research and applications have been presented, focusing particularly on the recent successes of machine learning based models for narrow AI.

Then, the ways in which AI-based systems may use personal data have been described and the resulting opportunities and risks have been illustrated. It has been observed that personal data can be used to predict human behaviour, to learn the propensities and attitudes of individuals, to exercise influence over behaviour. The feedback relations between AI and big (personal) data have also been considered: the possibility of using AI stimulates the collection of vast sets of personal data, and the availability of big data sets, in its turn, stimulates novel applications of AI.

Benefits and risks concerning the deployment of AI have been examined. The combination of AI and big data offers great opportunities for scientific research, welfare, governance and administration, but it also engenders serious risks for individuals and society: intensified surveillance, control, manipulation, unfairness and discrimination. Even when the processing of data is non-discriminatory and based on reliable technologies, it may lead to unacceptable levels of surveillance, control and nudging, which affect individual autonomy, cause psychological harm, and impair genuine social interactions and the formation of public opinion.

4.1.2. Normative foundations.

The normative foundations of a human-centred regulation of AI have been considered. It has been observed that a framework is emerging, in which traditional ethical ideas, such as respect for human autonomy, prevention of harm, and fairness are combined with specific and somehow technical requirements concerning transparency, explicability, robustness and safety.

Turning from ethics to law, it has been claimed that AI relates to the law at different levels. As a pervasive and multifaceted technology, AI may either enhance or impair the exercise of multiple fundamental rights: privacy and data protection, civil freedoms and social rights. It can also contribute to, or detract from, the realisation of different social values, such as democracy, welfare, or solidarity. Correspondingly, promoting the opportunities of AI and countering its risks falls within the purview of multiple areas of the law, from data protection, to consumer protection, competition law, labour law, constitutional and administrative law. Different interests are at stake: the interests in data protection, in a fair algorithmic treatment, in transparency and accountability, in not being misled or manipulated, in the trustworthiness of AI systems, in algorithmic competition, among others.

4.2. AI in the GDPR

Based on this analysis, the provisions in the GDPR have been analysed to determine to what extent they adequately address AI applications. Does the GDPR contribute make it possible to enjoy the opportunities enabled by AI while preventing the attendant risks, or does it rather fail in this mission, either by establishing barriers to the beneficial deployment of AI, or conversely failing to prevent avoidable risks?

4.2.1. Personal data in re-identification and inferences

First of all, AI raises issues pertaining to the very nature of personal data, concerning in particular the possibility of reconnecting the data subjects with their de-identified data, and the possibility of inferring new personal data from existing data. In this regard the notion of personal data in the GDPR does not provide clear answers. It would be advisable to clarify, possibly in a soft-law instrument, such as an opinion of the Article 29 Working Party, that re-identification consists of a processing of personal data, and indeed can be assimilated to collection of new personal data. Therefore, re-identification is fully subject to all GDPR requirements (including obligations to inform the data subject and the need for a legal basis).

Special considerations apply to the inference of personal data. A possible approach could consist in distinguishing the cases in which an inference of personal data is accomplished without engaging in consequential activities, i.e., the inferred personal data are merely the output of a computation which does not trigger consequential actions, and the cases in which the inferred data are also used as input for making assessment and decisions. In the latter case, the data should definitely count as newly collected personal data.

4.2.2. Profiling

Profiling is at the core of the application of AI to personal data: it consists in inferring new personal data (expanding a person's profile) on the basis of the available personal data. Profiling provides the necessary precondition for automated decision-making, as specifically regulated in the GDPR. A key issue is the extent to which the law may govern and constrain such inferences, and the extent of the data subject's rights in relation to them. This aspect is also not clearly worked out in the GDPR. Neither is the extent to which the data subject may have a right to reasonable automated inferences clear, even when these inferences provide a basis for making assessments or decisions.

4.2.3. Consent

The requirement of specificity, granularity and freedom of consent are difficult to realise in connection with AI applications. Thus, in general, consent will be insufficient to support an AI application, unless it appears that the application pursues a legitimate interest and does not unduly sacrifice the data subject's rights and interests under Article 6 (1)(f). There are, however, cases in which consent by the data subject would be the decisive criterion by which to determine whether his or her interests have been sufficiently taken into consideration by the controller (e.g., consent to profiling in the interest of the data subject).

4.2.4. AI and transparency

The report distinguishes between information to be provided before the data subject's data are processed for the purpose of profiling and automated decision-making (ex-ante information), and the information to be provided after the data have been processed (ex-post information).

Ex-ante information is addressed by the right to information established by Articles 13(2)(f) and 14(2)(g) requiring two kinds of information to be provided: information on the existence of automated decision-making and meaningful information on its logic and envisaged consequences.

There is an uncertainty as to what is meant by the logic and consequences of an automated decision. With regard to complex AI processing, there is a conflict between the need for the information to be concise and understandable on the one hand, and the need for it to be precise and in-depth on the other.

Ex-post information is addressed by Article 15(1), which reiterates the same information requirements in Articles 13 and 14. It remains to be determined whether the controller is required to provide the data subject with only general information or also with an individualised explanation.

4.2.5. The rights to erasure and portability

The GDPR provisions on the rights to erasure and portability do not specifically address AI-based processing. However, some important issues emerge concerning the scope of such rights. With regard to the right to erasure, we may ask whether it may also cover inferred information and with regard to the right to portability, whether it also includes information collected by tracking the individuals concerned. The scope of the right to erasure, as distinguished from the right to object, depends on the extent to which the processing is unlawful. Thus, uncertainties about the unlawfulness of the processing will likely also affect the right to erasure.

4.2.6. The right to object

Article 21 specifically addresses the ability to object to profiling, on personal grounds, when the processing is based on public interests (Article 6 (1)(e)), or on legitimate private interests (Article 6 (1)(f)). Data subjects have an unconditioned right to object to profiling for purposes of direct marketing. Data subjects can also object to profiling for statistical purposes. The right to object should have a vast scope with regard to AI-based processing. The key issue would be to make it easier to exercise this right.

4.2.7. Automated decision-making

Article 22 on automated decision-making is highly relevant to AI, since automated decisions today are indeed taken through AI-based systems. According to the interpretation suggested above, Article 22(1) prohibits any completely automated decisions based on profiling and having legal or significant effects on the data subject. Article 22(2) introduces broad exceptions to the prohibition, allowing for automated decisions to be introduced by contract, law or consent.

This provision raises a number of issues, from determining when a decision is 'based solely on automated processing' to establishing whether its effects 'significantly' affect the data subject, to establishing when exceptions apply. Article 22(3) requires suitable safeguard measures to be adopted, 'at least' concerning the data subject's right to obtain human intervention, to express his or her point of view and to contest the decision. This list omits the safeguard consisting of the right to obtain an individualised explanation, which specifies the reasons why an unfavourable decision has been adopted. It also leaves out the requirement that the decision be 'reasonable,' meaning that its input factors and aims are acceptable and its method reliable (see Section 3.1.2 above). Reasonableness also requires that the extent to which certain input factors influence the decision should be proportionate to the causal or at least predictive importance of such factors relative to the legitimate goals being pursued.

4.2.8. AI and privacy by design

A risk-based approach to data-protection focuses on preventing harm, rather than on providing individual data subjects with legal powers over the processing of their data. A key role in this regard is played by Article 25, which, under the heading 'Data protection by design and by default', requires that technical and organisational measures be adopted to implement data protection principles and integrate safeguards in the processing. With regard to AI, these measures should include controls

over the representativeness of training sets, over the reasonableness of the inferences (including the logical and statistical methods adopted) and over the absence of unfairness and discrimination. Appropriate security measures, such as encryption or pseudonymisation, should also prevent unauthorised uses of the data (Article 32 (1)). High risk processing operations are subject to mandatory data protection assessment (Article 35 (1)), a requirement that applies in particular to the 'systematic and extensive evaluation of personal aspects' for the purpose of automated decision-making including profiling (Article 35 (3)(a)). Article 37 requires that a data protection officer be designated when a 'regular and systematic monitoring of data subjects on a large scale' is envisaged. Articles 40-43, on codes of conduct and certification, although not specifically addressing AI, identify procedures for anticipating and countering risks, and incentivise the adoption of preventive measures that are highly significant to AI.

4.2.9. AI, statistical processing and scientific research

In combination with big data, AI can provide useful results for science and statistical purposes (e.g. in medicine for diagnosis or prognosis, in the social sciences for understanding economic or political behaviour, in business for detecting consumer tastes and trends). These results have a general nature (they are not attached to particular individuals); therefore, they do not count as personal data. However, statistical and scientific processing also affects individuals, by exposing their data to security risks and abuse. Moreover, statistical results may indirectly affect individuals, since they provide information – possibly inaccurate or misleading – concerning the groups to which an individual belongs. The GDPR allows repurposing for scientific and statistical processing (under appropriate safeguards). The permission to engage in scientific and in particular statistical processing may enable beneficial uses of AI and big data in Europe, even though we need to take the implications for data subjects' rights and for competition into account.

4.3. AI and GDPR compatibility

In this section, the main results of the foregoing review will be summarised. It will be argued that policy options exist for ensuring that innovation in the field of AI is not stifled and remains responsible. Guidelines for controllers are needed, though there is no urgent need to make broad changes to the GDPR

4.3.1. No incompatibility between the GDPR and AI and big data

It has been argued that the GDPR would be incompatible with AI and big data, given that the GDPR is based on principles – purpose limitation, data minimisation, the special treatment of 'sensitive data', the limitation on automated decisions – that are incompatible with the extensive use of AI, as applied to big data. As a consequence, the EU would be forced to either renounce application of the GDPR or lose the race against those information-based economies – such as the USA and China – that are able make full use of AI and big data.[117]

Contrary to this opinion, this report shows that it is possible – and indeed likely – that the GDPR will be interpreted in such a way as to reconcile both desiderata: protecting data subjects and enabling useful applications of AI. It is true that the full deployment of the power of AI and big data requires collecting vast quantities of data concerning individuals and their social relations, and that it also requires processing of such data for purposes that were not fully determined at the time the data were collected. However, there are ways to understand and apply the data protection principles that are consistent with the beneficial uses of AI and big data.

[117] Zarsky (2017), Hildebrandt (2015)

The requirement that consent be specific and purpose limitation be respected should be linked to a flexible application of the idea of compatibility, that allows for the reuse of personal data when this is not incompatible with the purpose for which the data were collected. As noted above, the legal basis laid down in Article (6)(1)(f), namely, that the processing should serve a legitimate interest that is not outweighed by the interests of the data subjects, in combination with a compatibility assessment of the new uses, may provide sufficient grounds on which to make reuse permissible. Moreover, as noted above, reuse for statistical purposes is assumed to be compatible, and thus would in general be admissible (unless it involves unacceptable risks for the data subject).

Even the principle of data-minimisation can be understood in such a way as to enable a beneficial application of AI. This may involve in some context reducing the 'personality' of the data, namely the ease with which they can be connected to the individuals concerned, with measures such as pseudonymisation, rather than focusing on the amount of personal data to be preserved. This also applies to re-identification, the possibility of which should not exclude the processing of data which can be re-identified, but rather requires viewing re-identification as the creation of new personal data, which should be subject to all applicable rules, and strictly prohibited unless all conditions for the lawful collection of personal data are met, and should also be subject to the compatibility test.

The information requirements established by the GDPR can also be met with regard to AI-based processing, even though the complexity of AI systems represents a difficult challenge. The information concerning AI-based applications should enable the data subjects to understand the purpose of the processing and its limits, without going into technical details.

The GDPR allows for inferences based on personal data, including profiling, but only under certain conditions and so long as the appropriate safeguards are adopted.

The GDPR does not exclude automated decision-making, as it provides for ample exceptions – contract, law or consent – to the general prohibition set forth in Article 22(1). Uncertainties exist concerning the extent to which an individual explanation should be provided to the data subject. Uncertainties also exist about the extent to which reasonableness criteria may apply to automated decisions.

The GDPR provisions on preventive measures, and in particular those concerning privacy by design and by default should also not hinder the development of AI applications, if correctly designed and implemented, although they may entail some additional costs.

Finally, the possibility of using the data for statistical purposes – with appropriate security measures, proportionate to the risks, which should include at least pseudonymisation – opens wide spaces for the processing of personal data in ways that do not involve the inference of personal data.

4.3.2. GDPR prescriptions are often vague and open-ended

In the previous sections it has been argued that the GDPR allows for the development of AI and big data applications that successfully balance data protection and other social and economic interests. However, this does not mean that such a balance can be found by referring to the GDPR alone. The GDPR rules need to be interpreted and consistently implemented, and appropriate guidance needs to be provided on concrete implication of the GDPR for particular processing activities.

The GDPR indeed abounds in vague clauses and open standards. Among those pertaining to the issues here addressed, the following can be mentioned: the identifiability of the data subject (Article 4 (1)), the freeness of consent (Article (4)(11), the compatibility of further processing with the original (Article 5(1)(c)), the necessity of the data relative to their purpose (Article 5 (1)(c)), the legitimacy of the controller's interests and their non-overridden importance (Article 6(1)(f)), the meaningfulness of the information about the logic involved in automated decision-making (Articles 13(2)(f) and 14 (2)(g)), the suitableness of the safeguard measures to be adopted for

automated decision-making (Article 22 (2)), and the appropriateness of the technical and organisational measures for data protection by design and by default (Article 25).

In various cases, the interpretation of undefined GDPR standards requires balancing competing interests: it requires determination of whether a certain processing activity, and the measures adopted are justified on balance, i.e., whether the controller's interests in processing the data and in (not) adopting certain measures are outweighed by the data subjects' interests in not being subject to the processing or in being protected by additional or stricter measures. These assessments depend on both (a) uncertain normative judgements on the comparative importance of the impacts on the interests at stake and (b) uncertain forecasts concerning potential future risks. In the case of AI and big data applications the uncertainties involved in applying indeterminate concepts and balancing competing interests are aggravated by the novelty of the technologies, their complexities, the broad scope of their individual and social effects.

It is true that the principles of risk-prevention and accountability potentially direct the processing of personal data toward being a 'positive sum' game (where the advantages of the processing, when constrained by appropriate risk-mitigation measures, outweigh its possible disadvantages), and enable experimentation and learning, avoiding the over- and under-inclusiveness issues involved in the applications of strict rules. On the other hand, by requiring controllers to apply these principles, the GDPR offloads the task of establishing how to manage risk and find optimal solutions onto controllers, a task which may be both challenging and costly. The stiff penalties for non-compliance, when combined with the uncertainty as to what is required for compliance, may constitute a novel risk, which, rather than incentivising the adoption of adequate compliance measure, may prevent small companies from engaging in new ventures.

No easy solution is available in the hyper-complex and rapidly evolving domain of AI technologies: rules may fail to enable opportunities and counter risks, but the private implementation of open standard, in the absence of adequate legal guidance, may also be unsatisfactory:

> [Giving] appropriate content to the law often requires effort, whether in analysing a problem, resolving value conflicts, or acquiring empirical knowledge. […] [I]ndividuals contemplating behavior that may be subject to the law will find it more costly to comply with standards, because it generally is more difficult to predict the outcome of a future inquiry (by the adjudicator, into the law's content) than to examine the result of a past inquiry. They must either spend more to be guided properly or act without as much guidance as under rules.[118]

Thus, the way in which the GDPR will affect successful applications of AI and big data in Europe will also depend on what guidance data protection bodies – and more generally the legal system – will be able to provide to controllers and data subjects. This would diminish the cost of legal uncertainty and would direct companies – in particular small ones that mostly need advice – to efficient and data protection-compliant solutions. Appropriate mechanisms may need to be devised, such as an obligation to notify data protection authorities when new applications based on profiling are introduced, but also the possibility to ask for preventive, non-binding, indications on whether and how such applications should be developed, and with what safeguards.

4.3.3. Providing for oversight and enforcement

As noted above, AI applications may affect not only the concerned individuals but also society at large. Even applications based on correct statistical principles, which do not target protected categories, and which adopt appropriate security measures may still impose undue burden on certain categories of citizens, or anyway have negative social impacts. Oversight by competent

[118] Kaplow (1992, 621).

authorities needs to be complemented by the support of civil society. As collective interests, power relations, and societal arrangements are at stake, a broad public debate and the involvement of representative institutions is also needed.

Collective enforcement is also a key issue that is not answered by the GDPR, which still relies on individual action by the concerned data subjects. An important improvement toward an effective protection could consist in enabling collective actions for injunctions and compensation. It has indeed been observed that US courts have been unable so far to deal satisfactorily with privacy harms, since on the one hand they rely on old-fashioned theories requiring compensable harms to be concrete, actual and directly caused by the defendant, and on the other hand they are unable to address a very high numbers of similar claims, each having small monetary value.[119] In Europe, data protection authorities can provide an alternative and easier avenue to enforcement, but nevertheless, the damaged parties have to rely on the judiciary to obtain compensation from privacy harms, which also includes non-material harm (Article 82). Thus, effective protection is dependent on the data subject's ability to engage in lawsuits. The possibility for multiple data subjects to merge similar claims to share cost and engage more effectively with the law is necessary to make legal remedies available to data subjects.

The Court of Justice has recently denied that a consumer can combine his or her individual data protection claim with claims concerning other consumers involved in similar cases.[120] In particular, it has affirmed that Max Schrems could exercise, in the courts of his domicile, only his individual claim against Facebook for data protection violations. He could not bring, before the same court, claims for similar violations that had been assigned to him by other data subjects. Perhaps the proposed directive on collective redress for consumers,[121] currently under interinstitutional negotiation[122], could present an opportunity to enable collective actions in the context of data protection.

4.4. Final considerations: some policy proposals on AI and the GDPR

In the following, the main conclusions of this report on the relations between AI and the processing of personal data are summarised.

- The GDPR generally provides meaningful indications for data protection relative to AI applications.
- The GDPR can be interpreted and applied in such a way that it does not hinder beneficial application of AI to personal data, and that it does not place EU companies at a disadvantage in comparison with non-European competitors.
- Thus, GDPR does not seem to require any major change in order to address AI.

[119] Cohen (2019, Ch. 5).

[120] Judgment in Case C-498/16 *Maximilian Schrems* v *Facebook Ireland Limited*, of 25 January 2018.

[121] Proposal for a directive of the European Parliament and of the Council on representative actions for the protection of the collective interests of consumers, COM(2018) 184 final.

[122] See European Parliament Legislative train schedule, Area of Justice and Fundamental Rights, Representative actions for the protection of the collective interests of consumers - a New deal for consumers at https://www.europarl.europa.eu/legislative-train/theme-area-of-justice-and-fundamental-rights/file-representative-actions-for-consumers

- That said, a number of AI-related data protections issues are not explicitly answered in the GDPR, which may lead to uncertainties and costs, and may needlessly hamper the development of AI applications.
- Controllers and data subjects should be provided with guidance on how AI can be applied to personal data consistently with the GDPR, and on the available technologies for doing so. This can prevent costs linked to legal uncertainty, while enhancing compliance.
- Providing adequate guidance requires a multilevel approach, which involves civil society, representative bodies, specialised agencies, and all stakeholders.
- A broad debate is needed, involving not only political and administrative authorities, but also civil society and academia. This debate needs to address the issues of determining what standards should apply to AI processing of personal data, particularly to ensure the acceptability, fairness and reasonability of decisions on individuals.
- The political debate should also address what applications are to be barred unconditionally, and which may instead be admitted only under specific circumstances. Legally binding rules are needed to this effect, since the GDPR is focused on individual entitlements and does not take the broader social impacts of mass processing into account.
- Discussion of a large set of realistic examples is needed to clarify which AI applications are on balance socially acceptable, under what circumstances and with what constraints. The debate on AI can also provide an opportunity to reconsider in depth, more precisely and concretely, some basic ideas of European law and ethics, such as acceptable and practicable ideas of fairness and non-discrimination.
- Political authorities, such as the European Parliament, the European Commission and the Council could provide general open-ended soft law indications about the values at stake and ways to achieve them.
- Data protection authorities, and in particular the Data Protection Board, should provide controllers with guidance on the many issues for which no precise answer can be found in the GDPR, which could also take the form of soft law instruments designed with a dual legal and technical competence.
- National Data Protection Authorities should also provide guidance, in particular when contacted for advice by controllers, or in response to data subjects' queries.
- The fundamental data protection principles – especially purpose limitation and minimisation – should be interpreted in such a way that they do not exclude the use of personal data for machine learning purposes. They should not preclude forming training sets and building algorithmic models, whenever the resulting AI systems are socially beneficial, and compliant with data protection rights.
- The use of personal data in a training set, for the purpose of learning general correlations and connection, should be distinguished from their use for individual profiling, which is about making assessments of individuals.
- The inference of new personal data, as is done in profiling, should be considered as creation of new personal data, when providing an input for making assessments and decisions. The same should apply to the re-identification of anonymous or

pseudonymous data. Both should be subject to the GDPR constraints on the collection of new data.

- Guidance is needed on profiling and automated decision-making. It seems that an obligation of reasonableness – including normative and reliability aspects – should be imposed on controllers engaging in profiling, mostly, but not only when profiling is aimed at automated decision-making. Controllers should also be under an obligation to provide individual explanations, to the extent that this is possible according to the adopted AI technology and reasonable according to costs and benefits. The explanations may be high-level, but they should still enable users to contest detrimental outcomes.
- It may be useful to establish obligations to notify data protection authorities of applications involving individualised profiling and decision-making, possibly accompanied with the possibility of requesting indications on data-protection compliance.
- The content of the controllers' obligation to provide information (and the corresponding rights of data subjects) about the 'logic' of an AI system need to be specified, with appropriate examples, with regard to different technologies.
- It needs to be ensured that the right to opt out of profiling and data transfers can easily be exercised through appropriate user interfaces, possibly in standardised formats.
- Normative and technological requirement concerning AI by design and by defaults need to be specified.
- The possibility of repurposing data for AI applications that do not involve profiling – scientific and statistical ones – may be broad, as long as appropriate precautions are in place preventing abusive uses of personal data.
- Strong measures need to be adopted against companies and public authorities that intentionally abuse the trust of data subjects by misusing their personal data, to engage in applications that manipulate data subjects against their interests.
- Collective enforcement in the data protection domain should be enabled and facilitated.

In conclusion, controllers engaging in AI-based processing should endorse the values of the GDPR and adopt a responsible and risk-oriented approach, and they should be able to do so in a way that is compatible with the available technologies and with economic profitability (or the sustainable achievement of public interests). However, given the complexity of the matter and the gaps, vagueness and ambiguities present in the GDPR, controllers should not be left alone in this exercise. Institutions need to promote a broad social debate on AI applications, and should provide high level indications. Data protection authorities need to actively engage a dialogue with all stakeholders, including controllers, processors, and civil society, to develop appropriate responses, based on shared values and effective technologies. Consistent application of data protection principles, when combined with the ability to use AI technology efficiently, can contribute to the success of AI applications, by generating trust and preventing risks.

5. References

AI-HLEG, High-Level Expert Group on Artificial Intelligence (2019). *A definition of AI: Main capabilities and scientific disciplines*.

AI-HLEG, *High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI*.

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge University Press.

Balkin, J. M. (2008). The constitution in the national surveillance state. *Minnesota Law Review 93*, 1–25.

Balkin, J. M. (2017). The three laws of robotics in the age of big data. *Ohio State Journal Law Journal 78*, 1217–241.

Barocas, S. and A. D. Selbst (2016). Big data's disparate impact. *California Law Re view 104*, 671–732.

Bayer, J., Bitiukova, N., Bard, P., Szakacs, J., Alemanno, A., and Uszkiewicz, E. (2019). *Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its member states*. Study, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament.

Bhuta, N., S. Beck, R. Geiss, C. Kress, and H. Y. Liu (2015). *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge University Press. Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Bosco, F., Creemers, N., Ferraris, V., Guagnin, D., & Koops, B. J. (2015). *Profiling technologies and fundamental rights and values: regulatory challenges and perspectives from European Data Protection Authorities*. In Reforming European data protection law (pp. 3-33). Springer, Dordrecht.

Brynjolfsson, E. and A. McAfee (2011). *Race Against the Machine*. Digital Frontier Press.

Burr, C. and Cristianini, N. (2019). Can machines read our minds? *Minds and Machines* 29:461–494.

Calo, M. R. (2012). Against notice skepticism in privacy (and elsewhere). *Notre Dame Law Review*, 87:1027–72.

Cate, F. H., P. Cullen, and V. Mayer-Schönberger (2014). *Data Protection Principles for the 21st Century: Revising the 1980 OECD Guidelines*. Oxford Internet Institute.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and Engineering Ethics* 24:505–528.

Cohen, J. D. (2019). *Between Truth and Power. The Legal Constructions of Informational Capitalism*. Oxford University Press.

Cristianini, N. (2016a, 23 November). Intelligence rethought: AIs know us, but don't think like us. *New Scientist* .

Cristianini, N. (2016b, 26 October). The road to artificial intelligence: A case of data over theory. *New Scientist*.

Cristianini, N. and T. Scantamburlo (2019). On social machines for algorithmic regulation. *AI and Society*.

De Hert, P. and Gutwirth, S. (2009). Data protection in the case law of Strasbourg and Luxemburg: Constitutionalisation in action. In Gutwirth, S., Poullet, Y., De Hert, P., de Terwangne, C., and Nouwt, S., editors, *Reinventing Data Protection*? 3–44. Springer.

Edwards, L. and Veale, M. (2019). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review*, 16-84.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena (2018). Ai4people– an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines 28*, 689–707.

Galbraith, J. K. ([1952]1956). *American Capitalism: The Concept of Countervailing Power*. Houghton Mifflin.

Guidotti, R., A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti (2018). A survey of methods for explaining black box models. *ACM Computer Surveys* 51 (5) Article 93, 1–4.

Halpern, J. Y. and Hitchcock, C. (2013). Graded causation and defaults. *The British Journal for the Philosophy of Science*, 1–45.

Harel, D. and Y. Feldman (2004). *Algorithmics: The Spirit of Computing*. Addison-Wesley.

Hildebrandt, M. (2009). Profiling and AML. In Rannenberg, K., Royer, D., and Deuker, A., editors, *The Future of Identity in the Information Society. Challenges and Opportunities*. Springer.

Hildebrandt, M. (2014). Location data, purpose binding and contextual integrity: What's the message? In Floridi, L., editor, *The protection of information and the right to privacy*, 31–62. Springer.

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Edgar.

Jobin, A., Ienca, M., and Vayena, E. (2019). Artificial intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence*, 1: 389–399.

Kahneman, D. (2011). *Thinking: fast and slow*. Allen Lane.

Kamara, I. and De Hert, P. (2019). Understanding the balancing act behind the legitimate interest of the controller ground: A pragmatic approach. In Seligner, E., Polonetsky, J., and Tene, O., editors, *The Cambridge Handbook of Consumer Privacy*. Cambridge University Press.

Kaplow, L. (1992). Rule vs standards: An economical analysis. *Duke Law Journal*, 42: 557–629.

Kleinberg, J., J. Ludwig, S. Mullainathan, and C. R. Sunstein (2018). Discrimination in the age of algorithm. *Journal of Legal Analysis 10*, 113–174.

Kurzweil, R. (1990). *The Age of Intelligent Machines*. MIT. Kurzweil, R. (2012). *How to Create a Mind*. Viking.

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics HFE-1* (March), 4–11.

Lippi, M., P. Palka, G. Contissa, F. Lagioia, H.-W. Micklitz, Y. Panagis, G. Sartor, and P. Torroni (2019). Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*.

Lippi, M., Contissa, G., Jablonowska, A., Lagioia, F., Micklitz, H.-W., Palka, P., Sartor, G., and Torroni, P. (2020). The force awakens: Artificial intelligence for consumer law. *The journal of Artificial Intelligence Research* 67:169 – 190.

Mantelero, A. (2017). Regulating Big Data. The guidelines of the Council of Europe in the context of the European data protection framework. *Computer Law and Security Review 33*, 584–602.

Mayer-Schönberger, V. and K. Cukier (2013). *Big Data*. Harcourt.

Mayer-Schönberger, V. and Y. Padova (2016). Regime change? enabling Big Data through Europe's new data protection regulation. *Columbia Science and Technology Law Review 17*, 315–35.

McAfee, A. and E. Brynjolfsson (2019). *Machine, Platform, Crowd*. Norton.

Marcus, G. and Davis, E. (2019). *Rebooting AI: building artificial intelligence we can trust*. Pantheon Books.

Mindell, D. A. (2015). *Our Robots, Ourselves: Robotics and the Myths of Autonomy*. Penguin.

Nilsson, N. (2010). *The Quest for Artificial Intelligence*. Cambridge University Press. O'Neil, C. (2016). *Weapons of math destruction: how Big Data increases inequality and threatens democracy*. Crown Business. Pariser, E. (2011). *The Filter Bubble*. Penguin.

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown Business.

Pariser, E. (2011). *The Filter Bubble*. Penguin.

Parkin, S. (14 June 2015). Science fiction no more? channel 4's humans and our rogue ai obsessions. *The Guardian*.

Pasquale, F. (2019). The second wave of algorithmic accountability. *Law and Political Economy*.

Pentland, A. (2015). *Social Physics: How Social Networks Can Make Us Smarter*. Penguin.

Polanyi, K. ([1944] 2001). *The Great Transformation*. Beacon Press.

Powles, J. and Nissenbaum, H. (2018). *The seductive diversion of 'solving' bias in artificial intelligence*. Medium.

Prakken, H. and G. Sartor (2015). Law and logic: A review from an argumentation perspective. *Artificial Intelligence 227*, 214–45.

Rawls, J. ([1971] 1999). *A Theory of Justice*. Oxford University Press.

Ruggeri, S., D. Pedreschi, and F. Turini (2010). Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law 18*, 1–43.

Russell, S. J. and P. Norvig (2016). *Artificial Intelligence. A Modern Approach* (3 ed.). Prentice Hall.

Sartor, G. (2017). Human rights and information technologies. In R. Brownsword, E. Scotford, and K. Yeung (Eds.), *The Oxford Handbook on the Law and Regulation of Technology*, pp. 424–450. Oxford University Press.

Stiglitz, J. (2019). *People, Power, and Profits. Progressive Capitalism for an Age of Discontent*. Norton.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton University Press.

Turing, A. M. ([1951] 1996). Intelligent machinery, a heretical theory. *Philosophia Mathematica 4*, 256–60.

van Harmelen, F., V. Lifschitz, and B. Porter (2008). *Handbook of Knowledge Representation*. Elsevier.

Varian, H. R. (2010). Computer mediated transactions. *American Economic Review* (2): 100, 1–10.

Varian, H. R. (2014). Beyond Big Data. *Business Economics* (49), 27–31.

Wachter, S. and B. Mittelstadt (2017). A right to reasonable inferences: Re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*, 1–130.

Wachter, S., B. Mittelstadt, and L. Floridi (2016). *Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation*. *International Data Privacy Law 7*, 76–99.

Yeung, K. (2018). 'Hypernudge': Big data as a mode of regulation by design. *Communication and Society 20*, 118–36.

Zarsky, T. Z. (2017). Incompatible: The GDPR in the age of Big Data. *Seton Hall Law Review*, 47:995–1020.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Hachette.

This study addresses the relation between the EU General Data Protection Regulation (GDPR) and artificial intelligence (AI). It considers challenges and opportunities for individuals and society, and the ways in which risks can be countered and opportunities enabled through law and technology.

The study discusses the tensions and proximities between AI and data protection principles, such as in particular purpose limitation and data minimisation. It makes a thorough analysis of automated decision-making, considering the extent to which it is admissible, the safeguard measures to be adopted, and whether data subjects have a right to individual explanations. The study then considers the extent to which the GDPR provides for a preventive risk-based approach, focused on data protection by design and by default.

OA-02-20-399-EN-N

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

ISBN: 978-92-846-6771-0
doi:10.2861/293
QA-QA-02-20-399-EN-N