COUNCIL OF EUROPE

CONSEIL DE L'EUROPE

Strasbourg, 15 October 2018

T-PD(2018)09Rev

CONSULTATIVE COMMITTEE OF THE CONVENTION FOR THE PROTECTION OF INDIVIDUALS WITH REGARD TO AUTOMATIC PROCESSING OF PERSONAL DATA

(Convention 108)

Report on Artificial Intelligence

Artificial Intelligence and Data Protection: Challenges and Possible Remedies

Directorate General of Human Rights and Rule of Law

Report by Alessandro Mantelero, Associate Professor of Private Law at the Polytechnic University of Turin, Department of Management and Production Engineering. The document is an expression of the author's personal viewpoint.

***

Table of contents

Part I – The State of the Art

I.1 Introduction

Defining the field of research of this Report is not an easy matter, since the boundaries of both data protection and Artificial Intelligence (hereinafter AI[1]) are rather uncertain. On the one hand, data-intensive technologies (including AI) represent a challenge to the application of some of the traditional principles of data protection, making them blurrier, less clear-cut or more difficult to apply [CoE 2017; Hildebrandt, 2016; Barocas & Nissenbaum, 2015; Citron & Pasquale, 2014; Mantelero, 2014; Rubinstein, 2013; Boyd & Crawford, 2012; Tene & Polonetsky, 2012]. On the other, AI is a broad field encompassing a variety of approaches that attempt to emulate human cognitive skills [Villani, 2018, 4].

Data protection and AI are by necessity correlated. Leaving aside science fiction scenarios, the rapid evolution of AI applications over recent years has its roots in the progressive process of datafication [Mayer-Schönberger & Cukier, 2013, 78; Lycett, 2013], with the result that personal data have increasingly become both the source and the target of AI applications (e.g. personal assistants, smart home devices etc.).

Against this background, different approaches are emerging in AI development, use and regulation. In reality, AI is largely unregulated and often not grounded on fundamental rights, relying instead mainly on data processing.

Regarding data processing, the global framework offers a range of ways to safeguard fundamental rights and, in particular, the right to the protection of personal data. European active role in the field of data protection may lead to a prominent part to play for this region in addressing the regulatory challenge of AI development.

The adoption of a perspective focused on fundamental rights may also mitigate the envisioned clash between a market- and technology-oriented development of AI and a more inclusive approach. From the perspective of Convention 108 and, more generally, of the Council of Europe's attitude to fundamental rights, a solution to the existing tension may be provided by the regulatory framework and in the jurisprudence of the European Court of Human Rights.

In terms of policy, the foundational nature of fundamental rights has led the Parties to Convention 108 to favour the development of technology grounded on these rights and not

[1] The term Artificial Intelligence was originally coined by John McCarthy, an American computer scientist known as the father of AI. See J. McCarthy, M. L. Minsky, N. Rochester, and C.E. Shannon, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', August 31, 1955. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html accessed 19 June 2018. A definition of AI is available here: https://www.coe.int/en/web/human-rights-rule-of-law/artificial-intelligence/glossary.

merely driven by market forces or high-tech companies. Moreover, the historical roots of European data protection lie in urging policy makers to consider the potentially adverse consequences of data processing technologies.

This rights-based approach necessarily impacts on AI development, which should be consistent with the values expressed in Convention 108 and in the regulations of the Council of Europe. The Parties to the Convention should therefore actively encourage AI developers towards **a value-oriented design** of products and services, and away from vague or overly optimistic views of AI.

At the same time, governments should be the first to use AI in a manner which is centred on safeguarding and promoting data protection and fundamental rights, thereby avoiding the development of AI systems or technologies which constrain individual and collective rights and freedoms.

For these reasons, it is important to extend European regulatory leadership in the field of data protection to a value-oriented regulation of AI [Villani, 2018, 7] based on the following three precepts:

- Values-based approach (encompassing social and ethical values)
- Risk assessment and management
- Participation

The Council of Europe's standpoint is broader than the EU's borders and encompasses a wide variety of legal cultures and regulatory approaches. Despite this, the Council of Europe's legal framework, and Convention 108 itself, provide a uniform background in terms of common values.

The Council of Europe may be one of the best fora to combine attention to fundamental rights and flexibility in technology regulation, adopting **a principles-based approach**. Principles can be broader in scope and interpreted specifically to meet the challenges of a changing world, whereas detailed legislative provisions do not appear to be able to react quickly enough to socio-economic and technological change.

Moreover, principles-based regulations leave room for the peculiarities of each local context. This is even more relevant with regard to AI applications, which can have an impact on contextual legal, ethical and social values [IEEE, 2016].

Of course, data protection per se does not cover all these aspects, which require a broader approach encompassing human rights[2] and societal issues[3] [EDPS, 2018; Mantelero, 2018;

[2] See Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Preamble and Art. 1.

[3] See Consultative Committee of Convention for the protection of individuals with regard to automatic processing of personal data, 'Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data' (hereinafter Guidelines) adopted on 23 January 2017.

Council of Europe, 2017]. However, data protection can strengthen and complement the response to these questions.

Data protection's focus on individuals, an awareness of the social consequences of data use and the link with personality rights may expand the data controller's approach beyond data protection to fundamental rights and collective interests. Regarding its complementary role, data protection helps to reveal the way data are used and the purposes of processing, which represent key elements in a better understanding of the potential consequences for a variety of rights and freedoms.

Finally, AI raises many different sector-specific issues concerning the various AI fields of application (labour, justice administration, crime control, contract relationships, etc.) and the consequences of AI use (e.g. sustainability, environment impact, political impact etc.), which must be addressed separately. Given the focus of Convention 108, these issues are not discussed in this Report, which concerns the common core of all these applications, i.e. data processing. In terms of potential impact, this analysis may therefore provide a contribution to the debate around the issues concerning both AI in general and its specific applications.

I.2 AI development

Over the years, many reports and scientific works have been published on AI and its evolution. It is unnecessary here to trace the uneven trajectory of the scientific and social interest in AI technology that society has shown since the earliest studies [McCulloch & Pitts, 1943; Turing, 1950] to the most recent contributions. Nor is it necessary to describe the increasing variety of AI applications and the results they have achieved.

However, a historical perspective is important to properly understand the present and near-term future for AI. Two questions arise in this regard: why has the policy debate of the last few years focused on AI? And what forms of AI can we reasonably expect in the next few years? The answers to these questions are crucial to addressing AI regulation. Indeed, we need to put the development of AI technology into context and avoid the confusing commercial and media narratives surrounding AI.

To begin with, AI is not mere hype. As occurred in the past with cloud computing, Big Data and IoT, there is a clear tendency of some vendors to magnify the possibilities of AI and the term has become a buzzword in contexts that do not strictly involve this technology. However, there is a basis of truth in this attention to AI concerning the peculiar technological

environment that makes it possible today to achieve results that could only be dreamt of in the past.

Over the past decade, the increasing availability of bandwidth for data transfer, data storage and computational resources – through the new paradigm of cloud computing – and the progressive datafication of large part of our life and environment have created a completely new context. This has led to a breakthrough in AI, enabling new forms of data management to extract more information and create new knowledge.

Big Data analytics and Machine Learning[4] represent the most recent products of this development process [The Norwegian Data Protection Authority, 2018, 5]. The concrete application of these technologies make it possible to envisage the kind of AI that can be reasonably expected in the next few years and shows how we are still very far from so-called General AI [Bostrom, 2016; Executive Office of the President, and National Science and Technology Council - Committee on Technology, 2016, p. 7; The Norwegian Data Protection Authority, 2018; Cummings et al., 2018].

Although "algorithms and artificial intelligence have come to represent new mythologies of our time" [CNIL, 2017], **this report focuses on the existing and near future applications of AI**, leaving aside challenging questions concerning human-like AI, in terms of machine liability and risks for humanity [Bostrom, 2016; Kurzweil, 2016]. Convention 108, both in the original text and in the modernised version, refers to "automated processing" or "automatic processing" and not to autonomous data processing, implicitly highlighting autonomy is a key element of human beings [European Commission, 2018].

This brief summary of the state of the art clearly shows how **AI is unavoidably based on data processing**. AI algorithms necessarily have an impact on personal data use and pose questions about the adequacy of the existing data protection regulations in addressing the issues that these new paradigms raise.

I.3 The perspective adopted

The major threats from AI concern the disputed sets of values adopted by AI developers and users, the latter including both consumers and decision-makers who use AI to support their choices. There is an emerging tendency towards a technocratic and market-driven society,

[4] The difference between these two technologies can be summarised as follows: "patterns and connections. This is where AI can make a difference. While traditional analytical methods need to be programmed to find connections and links, AI learns from all the data it sees. Computer systems can therefore respond continuously to new data and adjust their analyses without human intervention. Thus, AI helps to remove the technical barriers that traditional methods run into when analysing Big Data" [The Norwegian Data Protection Authority, 2018, 5].

which pushes for personal data monetisation, forms of social control and "cheap & fast" decision-making solutions [Spiekermann, 2016, 152-153] on a large (e.g. smart cities) and small (e.g. precision medicine) scale.

As this trend strengthens it challenges and progressively erodes individual self-determination, privacy-focused models, and mindful and cautious decision-making processes. Data bulimia, the complexity of data processing and an extreme data-centred logic may undermine the democratic use of data, supplanting individuals and collective bodies, as well as freedoms and self-determination, with a kind of data dictatorship [O'Neil, 2017] imposed by data scientists insensitive to societal issues.

To prevent the adverse consequences of AI prevailing over the benefits [ITU, 2017; Information Commissioner's Office, 2017, 15-18; World Economic Forum, 2018], it is necessary to stress **the centrality of the human being in technology (and AI) development**. This means reaffirming the predominance of fundamental rights in this field.

In this sense, the right to the protection of personal data can become a stepping stone towards designing a different data society, in which AI development is not driven by pure economic interest or dehumanising algorithmic efficiency.

A broad-ranging debate is needed to reinforce this fundamental rights-based paradigm. We need to critically assess the drive towards the extreme datafication of all aspects of our lives and affirm the importance of individual and collective rights. **Governments and citizens need to recognise the risks of datafication** and the potentially damaging implications of data-driven solutions [Rouvroy, 2016].

As with industrial and product development in the past, **awareness of risk is no a barrier to innovation, but rather an enabler**. Innovation must be developed responsibly, taking the safeguard of fundamental rights as the pre-eminent goal.

This necessarily requires the development of assessment procedures, the adoption of participatory models and supervisory authorities. A human rights-oriented development of technology might increase costs and force developers and business to slow their current time-to-market, as the impact of products and services on individual rights and society have to be assessed in advance. At the same time, in the medium to long-term, this approach will reduce costs and increase efficiency (e.g. more accurate prediction/decision systems, increased trust [World Economic Forum, 2018], fewer complaints). Moreover, businesses and society are mature enough to view **responsibility towards individuals and society as the primary goal in AI development**.

Alternatively, if AI follows a different path – as earlier technologies have done in their early stages – the risk is that it will develop in an unregulated environment, driven purely by technological feasibility, market or political interests, criteria that do not in themselves guarantee respect for human rights.

Data-centric AI development should therefore be based on the principles of Convention 108 as the foundations for a flourishing digital society. The key elements of this approach are:

- **Proportionality** (development of AI should be inspired by the proportionality principle,[5] efficiency should not therefore prevail over individuals' rights and freedoms; individuals have the right not to be subordinated to automated AI systems; legislators should aim to curb AI applications to safeguard individual and societal interests).
- **Responsibility** (which is not merely accountability, but also requires developers and decision-makers to act in a socially responsible manner. It also entails the creation of specific bodies to support and monitor their actions)
- **Risk management** (accountable AI means assessing the potentially adverse consequences of AI applications, and taking appropriate measures to prevent or mitigate such consequences)
- **Participation** (participatory forms of risk assessment are essential to give voice to citizens. At the same time, citizens' participation should not be understood to diminish decision-makers' accountability)
- **Transparency** (despite the current limitations affecting transparency of AI, a certain degree of transparency can help to ensure the effective participation of citizens and more accurately assess the consequences of AI applications).

I.4 Existing framework and principles

The existing regulatory framework applicable to AI and data processing is mainly grounded on Convention 108, although other legal instruments concerning data protection (such as recommendations[6] and guidelines[7]) may also be relevant with regard to specific fields. In this context, the Guidelines on Big Data adopted by the Council of Europe [Council of Europe, 2017] represent the first attempt to address the use of data-intensive solutions for decision-making and are part of a broader wave of documents and resolutions adopted by several European institutions to regulate the impact of algorithms on society [Council of Europe-Committee of experts on internet intermediaries (MSI-NET), 2018; European Data Protection Supervisor - Ethics Advisory Group, 2018; European Parliament, 2017; European Union Agency for Fundamental Rights (FRA), 2018].

[5] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Art. 5.
[6] See, e.g. Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries.
[7] See, e.g., Practical guide on the use of personal data in the police sector (2018); Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data (2017).

The scope of the Guidelines adopted on Big Data was "to contribute to the protection of data subjects regarding the processing of personal data in the Big Data context by spelling out the applicable data protection principles and corresponding practices, with a view to limiting the risks for data subjects' rights. These risks mainly concern the potential bias of data analysis, the underestimation of the legal, social and ethical implications of the use of Big Data for decision-making processes, and the marginalisation of an effective and informed involvement by individuals in these processes".

Although focused on Big Data analytics, these Guidelines cover a variety of questions involving data-intensive and complicated applications for decision making. For this reason, considerations about the potentially positive role of risk assessment (encompassing ethical and societal concerns), testing, data minimisation, expert committees, a precautionary approach[8] and freedom of human decision-makers can be equally applied to AI.

Some of these remedies are discussed further in this report (see below Part II). But concrete applications of AI call for an analysis of new issues (such as the role of transparency and the various values that should underpin AI applications) and suggest new remedies (e.g. a broader data protection impact assessment, potential limitations to AI use). Finally, the approach adopted by existing supervisory bodies (e.g. data protection Supervisory Authorities) may need to be reconsidered in light of the new challenges posed by AI and their potential consequences for society.

In this sense, AI – in a manner analogous[9] to Big Data[10] – represents a challenge for the application of traditional data processing principles[11] and may warrant a search for new applicative solutions to safeguard personal information and fundamental rights.

[8] See also Commission - European Group on, Ethics in Science and, & New Technologies, 2018, 16 ("As the potential misuse of 'autonomous' technologies poses a major challenge, risk awareness and a precautionary approach are crucial").

[9] See also in this sense The Norwegian Data Protection Authority, 2018 ("This report elaborates on the legal opinions and the technologies described in the 2014 report «Big Data – data protection principles under pressure». In this report we will provide greater technical detail in describing artificial intelligence (AI), while also taking a closer look at four relevant AI challenges associated with the data protection principles embodied in the GDPR: Fairness and discrimination, Purpose limitation, Data minimisation, Transparency and the right to information").

[10] See Guidelines, Section II ("Given the nature of Big Data and its uses, the application of some of the traditional principles of data processing (e.g. the principle of data minimisation, purpose limitation, fairness and transparency, and free, specific and informed consent) may be challenging in this technological scenario").

[11] For example, analytics make it hard to identify the specific purpose of data processing at the moment of data collection. Machine learning algorithms, on the other hand, whose purposes are necessarily specified, may not predict and explain how these purposes are to be achieved. In both cases therefore transparency on the purpose and manner of data processing may remain limited.

I.5 Individuals' self-determination in data processing

Over the last few years, privacy scholars have repeatedly pointed out the weakness of data subjects' consent in terms of self-determination. Long and technical data processing notices, social and technical lock-ins, obscure interface design, and a lack of awareness on the part of the data subject are some of the reasons for this weakness.

Moreover, AI-based profiling and hidden nudging practices challenge both the idea of freedom of choice based on contractual agreement and the notion of data subjects' control over their information. Finally, the frequent complexity and obscurity of AI algorithms hamper the chances of obtaining real informed consent.

Legal scholars have addressed these issues by highlighting **the role of transparency** [*ex multis* Edwards & Vale, 2017; Selbst & Powles, 2017; Wachter, Mittelstadt & Floridi, 2017; Burrell, 2016; Rossi, 2016]**, risk assessment** [Guidelines, 2017; Mantelero, 2017] **and more flexible forms of consent**, such as broad consent [Sheehan, 2011] or dynamic consent [Kaye et al., 2015]. Although none of these solutions provides a definitive answer to the problem of individual consent, in certain contexts these solutions, alone or combined, may reinforce self-determination.

Moreover, the notion of self-determination is not circumscribed by a given case of data processing. It can be used in a broad sense to refer to freedom of choice over the use of AI and the right to a non-smart version of AI-equipped devices and services [Office of the Privacy Commissioner of Canada, 2016 ("as smart devices and appliances become more and more normalized, there is an increasing "erosion of choice" for individuals who would have preferred their "non-smart" versions")].[12] This "zero option" for AI goes beyond the individual dimension and also relates to the way in which a community decides **what role AI should play in shaping social dynamics, collective behaviour, and decisions affecting entire groups of individuals** [Asilomar AI Principles, 2017 ("Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives")].

I.6 Minimisation

As with Big Data [Guidelines, 2017], data minimisation[13] poses challenges for AI. While the technologies differ, both Big Data and machine learning AI algorithms need a large amount

[12] See also Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108), Explanatory report, para 40.
[13] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Art. 5.

of data to produce useful results. This means that only a certain degree of minimisation is possible.

Moreover, as mentioned in the previous section on the "zero option", the adoption of solutions other than AI can help reduce the quantity of data collected, limiting the amount of information required (e.g. surveying a sample of the population rather than a large proportion of it).

In addition, some of the Council of Europe Guidelines on Big Data can be extended to AI. The Guidelines contain a principle which can equally be applied to AI: data should be collected and processed in such a way as to "minimise the presence of redundant or marginal data".[14] In the case of AI this primarily concerns training data. The Norwegian Data Protection Authority pointed out that "it would be natural to start with a restricted amount of training data, and then monitor the model's accuracy as it is fed with new data" [The Norwegian Data Protection Authority, 2018]. Moreover, studies could also examine the development of algorithms that gradually delete data using automatic forgetting mechanisms [Gama et al., 2013, 12-13].

Although machine learning necessarily requires large datasets in the training phase, it is important to adopt a design paradigm that **critically assesses the nature and amount of data used**, reducing redundant or marginal data and only gradually increasing the size of the training dataset.[15] Minimisation may also be achieved in training algorithms by using synthetic data[16] [UK Department for Digital, Culture, 2018] originating from a sub-set of personal data and subsequently anonymised [Barse et al., 2003].

I.7 Bias

Although accurate AI systems can reduce or eliminate human bias in decision-making, it is also possible that data-intensive applications are affected by potential bias, as both deterministic and machine learning AI uses data input to extract further information (analytics) or create and train ML models. The bias may concern the data scientists' methods

[14] See Guidelines, Section IV, para 4.2.
[15] See also Guidelines, Section IV, para 4.3 ("When it is technically feasible, controllers and, where applicable, processors should test the adequacy of the by-design solutions adopted on a limited amount of data by means of simulations, before their use on a larger scale").
[16] Synthetic data are generated from a data model built from real data. They should be representative of the original real data. See the definition of synthetic data in OECD. 'Glossary of Statistical Terms'. 2007. http://ec.europa.eu/eurostat/ramon/coded_files/OECD_glossary_stat_terms.pdf ("An approach to confidentiality where instead of disseminating real data, synthetic data that have been generated from one or more population models are released").

(e.g. measurement bias, bias affecting survey methodologies, bias in cleaning and pre-processing stages) [Veale and Binns, 2017], the object of their investigation (e.g. social bias due to historical bias[17] or underrepresentation of some categories) [World Economic Forum, 2018, 8-9], their data sources (e.g. selection bias) or the person responsible for the analysis (e.g. confirmation bias) [UK Department for Digital, Culture, 2018; Information Commissioner's Office, 2017, 43-44; AI Now Institute, 2016].

Biased datasets may adversely affect algorithms, with a higher impact in the case of ML where bias may affect the design and the development (training) of the algorithm. This issue has already been partially addressed by the Council of Europe Guidelines on Big Data, which suggest **a by-design approach to avoid "potential hidden data biases and the risk of discrimination or negative impact on the rights and fundamental freedoms of data subjects, in both the collection and analysis stages**".[18]

Bias may be due to biased datasets [AI now Institute, 2017, 4, 16-17], but may also result from intentional or unintentional decisions by the developers. In this sense, machine predictions and performance "are constrained by human decisions and values, and those who design, develop, and maintain AI systems will shape such systems within their own understanding of the world" [AI Now Institute, 2017, 18]. This is why AI development cannot be left in the hands of AI designers alone: their technical background may mean they are less aware of the societal consequences of their decisions.

Committees of experts from a range of fields (social science, law, ethics, etc.) may represent the best setting in which to discuss and address questions of the impact of AI on individuals and society (see below Section II.3.1), compensating for the limited viewpoint of the AI developers. Multidisciplinary committees might also be able to detect potential bias that depends on the identity of AI developers: e.g. gender bias, ideological bias or under-representation of minorities [AI Now Institute, 2016, 5].

Another way to reduce the chances of AI application bias is through **participatory forms of risk assessment** [Mantelero, 2018] focused not merely on data security and data quality (see below Section II.3.2) but also on the active engagement of the groups potentially affected by AI applications, and who can contribute to the detection and removal of existing bias [AI Now Institute, 2016, 24].

[17] See e.g. Amazon ditched AI recruiting tool that favored men for technical jobs. The Guardian. 2018 October 10. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine ("But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way. That is because Amazon's computer models were trained to vet applicants by observing patterns in résumés submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry").

[18] See Guidelines, Section IV, para 4.2.

This approach, focused on responsible AI design [Guidelines, 2017],[19] aims to prevent the biased conditions that can affect datasets or algorithms. In a context necessarily characterised by a certain degree of obscurity and complexity, prior assessments and responsible design can be more effective than any analysis carried out once a discriminatory result has been discovered [Selbst, 2017, 163 ("Even if the result can be traced to a data quality problem, those problems are often quite complicated to rectify. It might be easy to determine that something is off about the data, but it can be more difficult to figure out what that something is […] Even if all the sources of bias are identified, the magnitude of each source's effect is still likely unknown"); Brauneis et al. 2018, 131].

Attention to potential bias, from the earliest design stage [UK Department for Digital, Culture, 2018], also entails deeper reflection about training datasets and the training phase in general, to curb the negative consequences of historical bias in pre-existing data-sets. On this point, some have suggested tracking "the provenance, development, and use of training datasets throughout their life cycle" [AI Now Institute, 2017].

Accurate testing of the training phase before the deployment of AI algorithms on a large scale could reveal hidden bias. This is why the Guidelines on Big Data highlight the role of simulations [Guidelines Big Data;[20] AI Now Institute, 2017]. Moreover, hidden bias may also involve **machine-generated bias which is different from human bias** [Cummings, 2018, 2 ("Machines and humans have different capabilities, and, equally importantly, make different mistakes based on fundamentally divergent decision-making architectures"); Caruana et al., 2015; Szegedy et al., 2013].

In the AI context, the **assessment of potential bias can also become controversial**, given the multiple variables involved and the classification of people into groups which do not necessarily correspond to the traditional discriminatory categories [Donovan et al., 2018, 5]. Questions regarding **machine bias cannot be deflected by the argument that human decisions are fallible**, and that AI is a way to reduce human error. There are four reasons why this is comparison does not work.

First, AI solutions are designed to be applied serially. As with product liability, poor design (i.e. bias) inevitably affects numerous people in the same or similar circumstances, whereas a human error only affects an individual case.

[19] See Guidelines, Section IV.4.2 ("Controllers and, where applicable, processors should carefully consider the design of their data processing, in order to minimise the presence of redundant or marginal data, avoid potential hidden data biases and the risk of discrimination or negative impact on the rights and fundamental freedoms of data subjects, in both the collection and analysis stages").

[20] See Guidelines, Section IV.4.3 ("When it is technically feasible, controllers and, where applicable, processors should test the adequacy of the by-design solutions adopted on a limited amount of data by means of simulations, before their use on a larger scale. This would make it possible to assess the potential bias of the use of different parameters in analysing data and provide evidence to minimise the use of information and mitigate the potential negative outcomes identified in the risk-assessment process described in Section IV.2").

Second, although there are fields in which error rates for AI are close to, or lower than, the human brain (image labelling, for instance) [Artificial Intelligence Index, 2017], most complicated decision-making tasks have higher error rates[21] [Cummings et al., 2018, 13].

Third, there is a socio-cultural dimension to human error that sets it apart from machine error in terms of social acceptability and exoneration. This necessarily influences the propensity to adopt potentially fallible AI solutions.

Finally, comparing the adverse outcomes of human and AI decisions [e.g. Federal Ministry of Transport and Digital Infrastructure, 2017, 10 "The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks"] is essentially based on the mere numerical comparison of resulting harms (e.g. number of victims of human-driven cars vs. number of fully autonomous AI cars) which is too reductive. In assessing the consequences of AI and human decisions we need to **consider the distribution of the effects** (i.e. individuals adversely affected belonging to different categories, the varying conditions in which the harm occurred, the severity of the consequences, etc.). Moreover, this sort of quantitative approach appears at odds with the precautionary approach [Guidelines, 2017] which requires the adoption of risk prevention policies rather than a mere reduction of harm.

[21] See e.g. Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Computer Science 2:e93 https://doi.org/10.7717/peerj-cs.93

Part II- Challenges and Possible Remedies

II.1 Limitations to AI use

Data protection regulations, as well as Convention 108, provide safeguards that can equally be applied to algorithms (including AI algorithms) used in automated decision-making systems. However, the red line between human and automated decisions cannot be drawn on the basis of the mere existence of a non-human decision-making process. Indeed, the supposedly reliable nature of AI mathematics-based solutions can induce those taking decisions on the basis of algorithms to place trust in the picture of individuals and society that analytics suggest. Moreover, this attitude may be reinforced by the threat of potential sanctions for taking a decision that ignores results produced by analytics. So the presence of a human decision-maker is not *per se* sufficient.

AI algorithms benefit from the allure of mathematical objectivity, which, combined with the complexity of data management and the subordinate position of those taking decisions in an organisation, can make it harder for a human decision-maker to take a decision other than the one suggested by the algorithm.[22]

Against this background, the distinction to be made is between cases where the human decision-maker has effective freedom and those where she does not. Here the Guidelines on Big Data already highlighted **the importance of protecting the effective freedom of the human decision-maker**.[23]

In assessing cases of potential imbalance an important role may be played by expert committees (see below Section II.3.1), which may also facilitate stakeholders' participation in the assessment (see below Section II.3.2).

Where decisions can be delegated to AI-based systems, or when human decision-makers cannot have effective oversight of AI decisions, the broader question arises about whether to adopt these systems rather than human-based methods.[24] This should lead communities

[22] See also Brauneis, Robert, and Ellen P Goodman. 'Algorithmic Transparency for the Smart City'. Yale J. L. & Tech. 20 (2018): 103, 126-127 ("Over time, deference to algorithms may weaken the decision-making capacity of government officials along with their sense of engagement and agency").

[23] See Guidelines, Section IV.7.4 ("On the basis of reasonable arguments, the human decision-maker should be allowed the freedom not to rely on the result of the recommendations provided using Big Data").

[24] See, e.g., ITU, 2017, 34 ("Dr. Margaret Chan, [the now former] Director-General of WHO, observed that "medical decisions are very complex, based on many factors including care and compassion for patients. I doubt a machine can imitate – or act – with compassion. Machines can rationalize and streamline, but AI cannot replace doctors and nurses in their interactions with patients"). See also Article 5 of the Modernised Convention 108 and the Explanatory Report which point out how this principle "is to be respected at all stages of processing, including at the initial stage, i.e. when deciding whether or not to carry out the processing".

or groups potentially affected towards **a participatory discussion on the adoption of AI solutions**, analysing the potential risks (see below risk assessment) and, where they are adopted, monitoring their application (see below vigilance).

II.2 Transparency

In the AI context, transparency[25] can have several different meanings. It may consist in a disclosure on the AI applications used, a description of their logic or access to the structure of the AI algorithms and – where applicable – to the datasets used to train the algorithms. Moreover, transparency can be both an ex ante or an ex post [e.g. Binns et al., 2018] requirement for data-centred decision-making.

Although transparency is important to have a public scrutiny of automated decision-making models [Reisman et al., 2018, 5], a generic statement on the use of AI does little to tackle the risk of unfair or illegitimate data use. On the other hand, accessing the algorithms' structure may make it possible to detect potential bias. However, IP rights and competition issues sometimes restrict this access, and in any case, even if such barriers do not exist, the complexity of the adopted models may represent a major challenge for human cognition [Lipton, 2018, 13]. In addition, in some cases transparency may prevent public bodies from carrying out their duties (e.g. predictive policing systems), or conflict with the data controller's security obligations concerning the personal data of data subjects other than those requesting access [Veale et al., Forthcoming 2018].[26]

For these reasons, a solution focused on disclosing the logic of algorithms may be the better option.[27] Even so, disclosure can be interpreted more or less narrowly. Giving information about the type of input data and the expected output,[28] explaining the variables and their weight, or shining light on the analytics architecture are various forms of transparency regarding the logic of AI algorithms.

Complex analysis processes (e.g. *deep-learning*) are a challenge to this notion of transparency – in terms of explaining the logic of the algorithms [Goodman & Flaxman,

[25] See Article 8 of the Modernised Convention 108.

[26] In any case, algorithms are sometimes harder for human beings to read and understand than mathematical or logical notation or natural language, "hence disclosure of computer code may be the less helpful alternative to easier means of interpretation", see Brauneis, Robert, and Ellen P Goodman. 'Algorithmic Transparency for the Smart City'. Yale J. L. & Tech. 20 (2018): 103, 130.

[27] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Article 9.1.c.

[28] Such information may be provided through 'learning by use' models, giving data subjects the chance to test analytics with different input values. Even in this case, however, there is a danger of misleading identification of the relevant inputs [Diakopoulos, 2013, 18].

2016] and the decisions taken using analytics[29] – and non-deterministic systems make it hard to provide detailed information on the logic behind the data processing.

Furthermore, the dynamic nature of many algorithms is in contrast to the static nature of transparency. Algorithms are continuously updated and changed, whereas a transparency disclosure only concerns the algorithm as it is being used at a given moment.

Finally, access to AI algorithms is not enough to detect potential bias. Resources in terms of time and skills are also required to perform this kind of analysis [Ananny & Crawford, 2016 ("The ideal of transparency places a tremendous burden on individuals to seek out information about a system, to interpret that information, and determine its significance")]. As a result, the deterrent effect of solutions such as auditing [Veale and Binns, 2017] or intervention by human decision-makers is impaired.[30] Research studies are currently trying to develop bias detection methods themselves based on algorithms,[31] though it is hard to see how introducing an algorithmic supervisor for algorithms can reduce the complexity of data governance.

None of these points weakens the argument for increased transparency generally [Burrell, 2016], especially in the public sector,[32] and its role in safeguarding the data subject's self-determination [Edwards & Vale, 2017; Selbst & Powles, 2017; Wachter, Mittelstadt & Floridi, 2017; Rossi, 2016]. If transparency is difficult to achieve with regard to the architecture and logic of algorithms, it may still be helpful in clarifying the reasons behind the decision to use such a complex tool [Burt et al., 2018, 2].

Transparency is only a part of the solution to the challenges of AI and has several limitations that should be fully addressed [Ananny & Crawford, 2016]. Nor should we forget that the algorithms are only one component of the AI application, the other being the datasets used for training or analysis. Biased datasets automatically produce biased results.

[29] See e.g. Article 10, Loi n° 78-17 du 6 janvier 1978 as amended by Loi n°2018-493 du 20 juin 2018. In some cases, it may be impossible to explain the reason for a decision suggested by the algorithm [Burrell, 2016]. Moreover, solutions such as the right to explanation are focused on decisions concerning specific persons, while the collective issues of the use of AI at group level remain unaddressed.

[30] These remedies are possible, but in many cases the auditing process requires a significant effort and human intervention is compromised by the complexity of data processing.

[31] See e.g. Lomas, Natasha. 2018. IBM Launches Cloud Tool to Detect AI Bias and Explain Automated Decisions. TechCrunch (blog). September, 19. Accessed 21 September 2018. http://social.techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions/.

[32] See e.g. Article 10 n. 2, Loi n° 78-17 du 6 janvier 1978 as amended by Loi n°2018-493 du 20 juin 2018. The public sector is known to use algorithms with great attention to the principle of equal treatment and a commitment to transparency and access rights in its administrative processes. On the limitations that may affect algorithmic transparency in the public sector, see Brauneis, Robert, and Ellen P Goodman. 'Algorithmic Transparency for the Smart City'. Yale J. L. & Tech. 20 (2018): 103–176 ("What we learned is that there are three principal impediments to making government use of big data prediction transparent: (1) the absence of appropriate record generation practices around algorithmic processes; (2) insufficient government insistence on appropriate disclosure practices; and (3) the assertion of trade secrecy or other confidential privileges by government contractors. In this article, we investigate each").

Finally, some data-intensive applications focus on de-contextualised data, ignoring the contextual information that often is vital to understand and apply the solution proposed by the AI application. **De-contextualisation** is also a danger in the choice of algorithmic models – where models originally used for one purpose are then re-used in a different context and for a different purpose [Donovan et al., 2018, 7, cite the case of the PredPol algorithm originally designed to predict earthquakes and later used to identify crime hotspots and assign police] – or in using models trained on historical data of a different population [AI Now Institute, 2018].

II.3.1 Risk assessment

Given the limits to transparency and individual self-determination (see above Section I.5), data protection regulations are increasingly stressing the role of risk assessment.[33] Risk assessment by the data controller and a safe AI environment can greatly enhance individuals' trust and their willingness to use AI applications. Users' preferences can be based on effective risk analysis and measures to mitigate risks [The Norwegian Data Protection Authority, 2018, 4], rather than merely relying on marketing campaigns or brand reputation.

The use of algorithms by modern data processing techniques [Council of Europe-Committee of experts on internet intermediaries (MSI-NET), 2018] as well as the trend towards data-intensive technologies [EDPS, 2018] have encouraged some to take a **wider view of the possible adverse outcomes** of data processing [Asilomar AI Principles, 2017 ("Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact")]. Groups of experts and scholars have gone beyond the traditional sphere of data protection [Taylor, Floridi & van der Sloot, 2017] to **consider the impact of data use on fundamental rights and collective social and ethical values** [Mantelero, 2018; Access Now, 2018].

Assessment of compliance with ethical and social values is more complicated than the traditional data protection assessment. Whereas, for example, the values (e.g. data integrity) underlying data security and data management are technologically-based and can thus be generalised across various social contexts, with social and ethical values the situation is different. These are necessarily context-specific and differ from one community to another

[33] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Art. 10.2.

[World Economic Forum, 2018, 12], making it harder to identify a benchmark for this kind of risk assessment.

This point is clearly addressed in the first section of the Guidelines on Big Data [Council of Europe, 2017], which urges both data controllers and data processors to "adequately take into account the likely impact of the intended Big Data processing and its broader ethical and social implications", in order to safeguard human rights and fundamental freedoms, in the light of Convention 108.[34]

The new element in the risk-assessment concerns the range of interests safeguarded and rights protected. The assessment addresses rights that go beyond traditional data protection, like the right to non-discrimination[35] [Barocas & Selbsr, 2016][36], as well as respect for social and ethical values [European Economic and Social Committee, 2017; AI Now Institute, 2017, 34-35 ("In order to achieve ethical AI systems in which their wider implications are addressed, there must be institutional changes to hold power accountable"); Access Now, 2018].

The Guidelines recognise the relative nature of social and ethical values and insist that data uses must not conflict with the "ethical values commonly accepted in the relevant community or communities and should not prejudice societal interests, values and norms".[37] While the Guidelines acknowledge the difficulties in identifying the values to be considered in a broader assessment, they do propose some practical steps towards this end. Following the view of privacy scholars who have examined this issue [Wright, 2011], they suggest that "the common guiding ethical values can be found in international charters of human rights and fundamental freedoms, such as the European Convention on Human Rights".

Given the context-dependent nature of the social and ethical assessment and the fact that international charters may only provide high-level guidance, the Guidelines combine this general suggestion with a more tailored option, represented by "ad hoc ethics committees".[38] If the assessment detects "a high impact of the use of Big Data on ethical

[34] See Guidelines, Section IV, para 1.1.
[35] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Article 6.2.
[36] See also, regarding AI and self-driving cars, Federal Ministry of Transport and Digital Infrastructure. 2017. The Federal Government's Action Plan on the Report by the Ethics Commission on Automated and Connected Driving (Ethical Rules for Self-Driving Computers). http://www.bmvi.de/SharedDocs/EN/publications/action-plan-on-the-report-ethics-commission-acd.pdf?__blob=publicationFile ("In the case of "dilemmatic" situations, in which injury to persons cannot be ruled out, the Commission states that there must be no distinction based on personal features (age, gender, etc.)").
[37] Guidelines, Section IV, para 1.2.
[38] See Guidelines, Section IV, para 1.3 ("the assessment of the likely impact of an intended data processing described in Section IV.2 highlights a high impact of the use of Big Data on ethical values, controllers could establish an ad hoc ethics committee, or rely on existing ones, to identify the specific ethical values to be safeguarded in the use of data").

values," the committees, which in some cases already exist in practice, should identify the specific ethical values to be safeguarded with regard to a given use of data, providing more detailed and context-based guidance for risk assessment.[39]

The "architecture of values" defined by the Guidelines is based on three layers. A first general level is represented by the "common guiding ethical values" of international charters of human rights and fundamental freedoms. The second layer takes into account the context-dependent nature of the social and ethical assessment and focuses on the values and social interests of given communities. Finally, the third layer consists in a more specific set of ethical values identified by ethics committees in relation to a given use of data.

The complexity of this assessment entails the continuous evolution of both the potential risks and the measures to tackle them. In this respect, the data protection supervisory authorities can play a significant role in supporting data controllers, informing them about data security measures and providing detailed guidelines on the risk-assessment process.[40] The Guidelines therefore do not leave the assessment exclusively in the hands of data controllers. In line with the approach adopted in Regulation (EU) 2016/679, if the use of big data "may significantly impact" the rights and fundamental freedoms of data subjects, controllers should consult the supervisory authorities to seek advice on how and to mitigate the risks outlined in the impact assessment. [41]

The Guidelines on Big Data do reach a number of conclusions that can be extended to AI, focussing on the automation of decision-making, which is at the core of the most challenging AI applications.

Finally, the increased burden consequent to a broader assessment is not only justified by the nature of the rights and freedoms potentially affected by AI application, but it also represents an opportunity to achieve competitive advantage. **Fostering public trust[42]** in AI products and services give companies the chance to better respond to the increasing consumers' concern about data use and AI. Similarly, increasing government agencies' accountability about their AI systems increase citizens' trust in public administration and prevent unfair decisions. From this perspective, a significant role can also be played by certifications [IEEE, 2016, 46 "Additionally, we need to develop a certification scheme for AI/AS that ensures that the technologies have been independently assessed as being safe

[39] The same two-layer model, based on general guidelines and tailored guidance provided by *ad hoc* committee, is already adopted in clinical trials. As in the big data context, here the specific application of technology poses context-related questions which must necessarily be addressed depending on the conflicting interests of each case. The results is an 'in the context' assessment of the conflicting interests.

[40] See Guidelines, Section IV, para 2.8.

[41] See Guidelines, Section IV, para 2.8.

[42] See also Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Explanatory Report ("the development and use of innovative technologies should also respect those rights. This will help to build trust in innovation and new technologies and further enable their development").

and ethically sound" but see Brundage et al., 2018, 56, 93], codes of conduct and standards. Those different tools contribute to increase accountability and provide guidance on data and system integrity,[43] encompassing procedures to trace the decision-making process and to prevent any form of manipulation of the generated results.

II.3.2 Ethics committees

In respect to data-intensive applications, ethics committee is attracting increasing attention in AI circles, though there is no a unanimous consensus on its nature and function. Theoretical studies, policy documents and corporate initiatives all offer differing solutions in this regard.

The first difference in approach that emerges concerns the level at which these committees should work [Polonetsky, Tene & Jerome, 2015; Calo, Ryan. 2013; White House, 2015; IEEE, 2016]. Some proposals describe them as national committees [Villani, 2018] which should provide general guidelines on issues of AI development.[44] This is not a completely new idea and resembles the existing national bioethical committees. However, in the case of AI data-intensive applications that use personal information, the interplay between these national committees and the national data protection authorities needs to be examined carefully [Mantelero, 2016], as does the interplay with other national bodies, such as the antitrust or national security authorities. Many countries already have independent watchdogs for supervising specific sectors where AI applications operate or may operate. From a regulatory perspective, it is therefore important to collaborate with these authorities and reconsider their role or strengthen their mutual cooperation [European Data Protection Supervisor, 2016, 3, 15; Conseil national du numérique, 2015, 74].

A different approach would be to introduce, ethics committees at company level, supporting data controllers for specific data applications, focusing on data controllers' operations. They might assume a broader role and act as expert committees not only on ethical issues, but also a broad range of societal issues relating to AI, including the contextual application of

[43] See also Article 7 of the Modernised Convention 108.

[44] See also the UK consultation on the new Centre for Data Ethics and Innovation https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation.

fundamental rights [Mantelero, 2018]. Several companies[45] have already set up internal or external committees to advise on critical projects.

This second solution based on corporate ethics committees, creates fewer difficulties in terms of overlap with existing regulators or supervising bodies but may require a more clearly defined relationship between these committees and the supervisory authorities. National legislators might empower supervisory Authorities to scrutinise these committees when shortcomings in their abilities or decisions affect data processing [Conseil national du numérique, 2015]. As with other types of advisory boards, creating AI ethics committees raises questions about their **independency**, their internal or external status, and the best practices to avoid conflicts of interest.

The make up of these committees will also depend on the complexity of the AI tools and applications. Where societal issues are significant, legal, ethical or sociological expertise, as well as domain-specific knowledge, will be essential.[46]

Such committees may play an even more important role in areas where transparency and stakeholders' engagement are difficult to achieve, such as predictive justice, crime detection or predictive policing.

Ethics committees can provide a valuable support to AI developers in designing rights-based and socially-oriented algorithms. Moreover, dialogue between the developers and

[45] See, in this sense, the increasing propensity of the big data-intensive and high-tech companies to set up their own ethics committees or advisory boards. See, e.g., Natasha Lomas, 'DeepMind now has an AI ethics research unit. We have a few questions for it…' *TechCrunch* (4 October 2017) < http://social.techcrunch.com/2017/10/04/deepmind-now-has-an-ai-ethics-research-unit-we-have-a-few-questions-for-it/> accessed; Axon AI Ethics Board <https://it.axon.com/info/ai-ethics> accessed 9 May 2018; DNA Web Team, 'Google drafting ethical guidelines to guide use of tech after employees protest defence project' *DNA India* (15 April 2018) <http://www.dnaindia.com/technology/report-google-drafting-ethical-guidelines-to-guide-use-of-tech-after-employees-protest-defence-project-2605149> accessed 7 May 2018. See also United Nations, 2011. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. United Nations Human Rights Council (UN Doc. HR/PUB/11/04).

[46] When there is a lower level of technical complexity in terms of the consequences of AI applications, the committee could be replaced by an expert in ethics and society, similar to the DPO's role for data protection. There should also be a mandatory requirement regarding the appointment and quality of the members of the ethics committee. Appointments should be guided by the type of data use and its potential impact on fundamental rights, taking ethical and societal issues as the key criteria. See in this sense IEEE, 2016, 41-42 which recommends "to create roles for senior level marketers, ethicists or lawyers who can pragmatically implement ethically aligned design […] A precedent for this new type of leader can be found in the idea of a Chief Values Officer created by Kay Firth-Butterfield" (CSER Cambridge. Kay Firth-Butterfield: Lucid AI's Ethics Advisory Panel., 2016. https://www.youtube.com/watch?v=w3-wYGbNZU4).

the committee[47] can favour the creation of more transparent data processing procedures and facilitate a clearer definition of their rationale (see also Section II.2).

II.3.3 Participatory assessment

Experts (e.g. ethics committees) can play an important role in detecting the potentially adverse consequences of AI applications and help data controllers to address any critical issue. However, in some cases analysis is impossible without engaging the target communities or groups.

As with the Social Impact Assessment, the societal consequences of AI may arouse an interest in public participation, individual and group empowerment through the assessment process, non-discrimination and equal participation in the assessment. A participatory approach[48] can also be helpful in gaining a better understanding of the various competing interests and ethical and social values.[49]

[47] See also UK Department for Digital, Culture, Media & Sport. 'Data Ethics Framework - GOV.UK', Section 3 (Use data that is proportionate to the user need). Accessed 4 July 2018. https://www.gov.uk/guidance/3-use-data-that-is-proportionate-to-the-user-need.

[48] The role of participatory approaches and stakeholders' engagement is specifically recognised in the context of fundamental rights [The Danish Institute for Human Rights, 2016, 24; Paul De Hert, 'A Human Rights Perspective on Privacy and Data Protection Impact Assessments. In David Wright and Paul De Hert (eds) Privacy Impact Assessment (Springer Dordrecht) 72 ("Further case law is required to clarify the scope of the duty to study the impact of certain technologies and initiatives, also outside the context of environmental health. Regardless of the terms used, one can safely adduce that the current human rights framework requires States to organise solid decision-making procedures that involve the persons affected by technologies")].

[49] Participation of the various stakeholders (e.g. engagement of civil society and the business community in defining sectoral guidelines on values) can be more effective than mere transparency, despite the emphasis on the latter in the recent data processing debate [The Danish Institute for Human Rights, 2016, 10 ("Engagement with rights-holders and other stakeholders are essential in HRIA […] Stakeholder engagement has therefore been situated as the core cross-cutting component")]. See also Walker, 2009, 41 ("participation is not only an end – a right – in itself, it is also a means of empowering communities to influence the policies and projects that affect them, as well as building the capacity of decision-makers to take into account the rights of individuals and communities when formulating and implementing projects and policies"). A more limited form of engagement, based on awareness, was suggested by Council of Europe Committee of experts on internet intermediaries [Council of Europe-Committee of experts on internet intermediaries (MSI-NET), 2018, 45 ("Public awareness and discourse are crucially important. All available means should be used to inform and engage the general public so that users are empowered to critically understand and deal with the logic and operation of algorithms. This can include but is not limited to information and media literacy campaigns. Institutions using algorithmic processes should be encouraged to provide easily accessible explanations with respect to the procedures followed by the algorithms and to how decisions are made. Industries that develop the analytical systems used in algorithmic decision-making and data collection processes have a particular responsibility to create awareness

Stakeholder engagement also **represents a development goal for the assessment** [United Nations Office of the High Commissioner for Human Rights, 2006], since it reduces the risk of under-representing certain groups and may also flag up critical issues that have been underestimated or ignored by the data controller [Wright & Mordini, 2012, 402].

However, stakeholder engagement should not be seen as a way for decision makers (data controllers in this case) to evade their responsibilities as leaders of the entire process [Palm & Hansson, 2006]. Decision-makers must remain committed to achieving the best results in terms of minimising the negative impact of data processing on individuals and society.

Finally, a participatory assessment of the far-reaching effects of algorithmic decision-making [CNIL, 2017, 30] may also drive data controllers to adopt **co-design solutions** for developing AI applications, actively engaging the groups potentially affected by them.

II.4 Liability and vigilance

Liability around AI applications remains an open issue for various reasons. As with product liability, whose principles focused on risk management and uncertainty can be broadly extended to AI, there are a number of applicable regulatory models (strict liability, liability based on fault etc.) and strategies (state intervention, mandatory insurance etc.).

One valuable solution appears to be the extension of the product liability logic to algorithms, channelling all liability to the producer. This would seems to be more workable than the alternative of data protection officer for algorithms [CNIL, 2017, 56 "identifying within each company or authority a team that is responsible for an algorithm's operation the moment this processes the data of humans"], where the pervasiveness of AI applications, the different parts involved and the role of the user make it difficult to disentangle the different aspects of AI liability.

Moreover, liability serves as a sort of closing rule for the system, which is valuable when the various ex ante remedies (such as transparency) have not worked [Asilomar AI Principles, 2017 ("Failure Transparency: If an AI system causes harm, it should be possible to ascertain why")]. However, since tort liability is normally regulated by national legislators, this report needs not discuss the different available solutions.[50]

and understanding, including with respect to the possible biases that may be induced by the design and use of algorithms")].

[50] Liability also assumes different forms in different fields of AI application (e.g. IP liability, decision-making, cars etc.), since liability is quite context specific.

Nevertheless, it is worth pointing out how risk management, transparency and liability can be combined not only at the AI applications development phase, but also in the following stage, when the algorithms are used [Access Now, 2018; ACM, 2018, 2.5]. This could lead supervisory authorities and data controllers to adopt forms of **algorithm vigilance** analogous to pharmacovigilance to react quickly in the event of unexpected and dangerous outcomes (e.g. Microsoft's chatbot Tay[51]) [Commission Nationale de l'Informatique et des Libertés - LINC, 2017].

II. 5 Sector-specific issues

AI has a significant impact on many sectors of our society and economy (e.g. predictive policing, justice, precision medicine, marketing, political propaganda). Sector-specific AI applications are characterised by different challenges and cannot be properly discussed in this report which provides a general overview of the main issues concerning the interplay between data protection and AI. This last section therefore briefly sheds light on two main areas only: public sector and workplace.

AI applications raise a number of specific questions when used in the public sector [Reisman et al., 2018], largely due to the imbalance of power between citizens and the administration and the essential services provided. Moreover, the adoption of complex and obscure AI solutions by governments and their agencies make it more difficult for them to comply with their accountability obligations, not only concerning data processing [Reisman et al., 2018].

This state of affairs would seem to warrant the adoption of tighter safeguards, beyond the remit of ad hoc committees or auditing. The safeguards should also contemplate an evaluation process that critically assess the need for the proposed AI solutions and their suitability to the delivery of services by public agencies or private companies acting on their behalf. This process requires that "at a minimum they [AI applications] should be available for public auditing, testing, and review, and subject to accountability standards" [AI Now Institute, 2017].

To achieve this goal **public procurement procedures may impose specific duties of transparency and prior assessment to AI providers**. Moreover, procurement procedures may also address the issues concerning trade secrets and IP protection, introducing specific contractual exceptions to increase transparency and make AI auditing possible.

[51] See Vincent, James. 'Twitter Taught Microsoft's Friendly AI Chatbot to Be a Racist Asshole in Less than a Day'. The Verge, 24 March 2016. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Regarding the effects of AI on the future of work, leaving aside its impact on the labour market, AI solutions may have an effect on relationships within the workplace.[52] In the first place, they can increase an employer's control over employees, in a situation that is often characterised by an imbalance of power.

Moreover, the use of hidden and unregulated forms of data processing might transform the workplace into an *in vivo* social experiment raising additional important questions about the role of transparency, ethics committees and voluntary participation in data processing.

Finally, devices given to employees by employers may have a dual use. For instance, wearable well-being devices can be worn in the workplace to gather biological data intended to safeguard the employee's health, but employees may also use them outside the work to track their sports fitness. Unless the repercussions for data protection and individual freedom are properly examined, such twin uses may blur the boundaries between work and private and life [AI Now Institute, 2017, 10], raising issues of pervasive control and the right to disconnect.

[52] See also Eur. Court of HR, Bărbulescu v. Romania, judgment of 5 September 2017, application no. 61496/08; Eur. Court of HR, Libert v. France, judgment of 22 February 2018, application no. 588/13.

Part III – Guidelines

The present Guidelines provide a set of baseline measures which governments, AI developers, AI manufacturers, and AI service providers should follow to secure the human dignity and the human rights and fundamental freedoms of every individual, in particular with regard to personal data protection.

Nothing in the present Guidelines shall be interpreted as precluding or limiting the provisions of the European Convention on Human Rights and of Convention 108 as amended ("Convention 108+")[53].

I. General guidance

1. Responsibility towards individuals and society is the corollary of any AI development, taking the safeguard of fundamental rights as an absolute pre-requisite.
2. A fundamental rights-oriented perspective is to be adopted by AI development and AI applications, in particular when AI is used in the context of decision-making processes.
3. AI development relying on personal data must be based on the principles of Convention 108+. The key elements of this approach are: proportionality of data processing, responsibility, transparency and risk management.
4. A risk-aware approach is not a barrier to innovation, but an enabler and the risks of datafication and the potentially adverse implications of data-driven solutions should thus be considered.
5. Individuals and communities should have the right to freely decide what role AI should play in analysing collective behaviour, influencing social dynamics, and in decision-making processes affecting entire groups of individuals.
6. In line with the guidance on risk assessment provided in the Guidelines on Big Data[54], a wider view of the possible outcomes of data processing should be adopted to consider the impact of data use not only on fundamental rights but also on collective social and ethical values.
7. AI development and AI applications cannot diminish or negatively affect the rights of data subjects enshrined in Convention 108+.

[53] Amending Protocol CETS n°223 to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.

[54] Guidelines on big data adopted by the Committee of Convention 108 in January 2017, available at https://rm.coe.int/CoERMPublicCommnSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a.

II. Guidance for AI developers

1. The Committee of Convention 108 actively encourages AI developers to adopt a value-oriented design of their products and services, consistent with Convention 108+ and other relevant instruments of the Council of Europe.
2. AI developers have to assess the adverse consequences of AI applications on the fundamental rights and freedoms of data subjects. Considering such consequences, precautionary approach based on risk prevention policies have to be adopted.
3. AI developers have to adopt a by-design approach to avoid potential unintentional and hidden data biases, and the risk of discrimination or negative impacts on the rights and fundamental freedoms of data subjects, in all phases of the processing, including data collection and analysis stages.
4. In developing AI applications, it is important to adopt a design paradigm that critically assesses the nature and amount of data used. Such design paradigms aim at reducing redundant or marginal data, starting with a restricted amount of training data, and then monitoring the model's accuracy as it is fed with new data. The use of synthetic data can be considered as one of the possible solutions to minimise personal data processed.
5. The risk of de-contextualised data (i.e. ignoring the contextual information characterising the specific situations where the proposed AI-based solutions should be applied) and de-contextualised algorithmic models (i.e. using AI models originally designed for different contexts or purposes) should be adequately considered in developing AI applications.
6. Committees of experts from a range of fields, as well as independent academic institutions, should be involved in AI development to provide a valuable support in designing rights-based and socially-oriented AI and to contribute to detect potential bias. Such committees play an even more important role in areas where transparency and stakeholders' engagement are difficult to achieve, such as for instance in AI designed to be used in a judicial or law enforcement context.
7. Participatory forms of risk assessment, based on the active engagement of the groups potentially affected by AI applications, should be adopted.
8. When it is technically feasible, AI developers should design their products and services in a manner that safeguards users' freedom of choice over the use of AI and provide alternatives to AI-equipped devices and services.
9. Data subjects are entitled to know the AI applications used and the reasoning underlying AI data processing operations, including the consequences of such a reasoning.

III. Guidance for policy makers

1. Public procurement procedures could impose specific duties of transparency and prior assessment of AI systems to service providers.
2. Public trust in AI products and services could benefit from an increased AI developers' accountability and the adoption of risk assessment procedures.
3. Data protection supervisory authorities and data controllers should adopt forms of algorithm vigilance to better ensure compliance with data protection and human rights principles over the entire lifetime of AI applications.
4. Overconfidence in the reliable nature of the solutions provided by AI systems, and fears of potential liability when taking a different decision than the one suggested by AI systems risk altering the autonomy of human intervention in decision-making processes. It is thus crucial that the freedom of human decision makers not to rely on the result of the recommendations provided using AI be preserved.
5. When AI applications may significantly impact on the rights and fundamental freedoms of data subjects, data controllers should consult the supervisory authorities to seek advice to mitigate this potential adverse impact.
6. Countries having established independent bodies supervising specific sectors where AI applications operate or may operate, should strengthen the mutual cooperation between these bodies and their cooperation with data protection supervisory authorities.
7. When committees of experts are created at company level, data protection supervisory authorities should be empowered to scrutinise those committees when shortcomings in their independency, abilities or decisions affect data processing.

References

- Access Now. 2018. The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems. https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-inmachine-learning-systems/.
- ACM. 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics
- AI Now Institute. 2016. The AI Now Report The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term. https://ainowinstitute.org/AI_Now_2016_Report.pdf.
- AI Now Institute. 2017. AI Now 2017 Report. Accessed 26 October 2017. https://assets.contentful.com/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf.
- AI Now Institute. 2018. Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems https://ainowinstitute.org/litigatingalgorithms.pdf.
- Ananny, M. and Crawford, K. 2016. Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. New Media & Society. https://doi.org/10.1177/1461444816676645.
- Artificial Intelligence Index - Annual Report 2017. http://aiindex.org/2017-report.pdf accessed 5 December 2017.
- Asilomar AI Principles. 2017 https://futureoflife.org/ai-principles/.
- Axon AI Ethics Board https://it.axon.com/info/ai-ethics
- Barocas, S. and Nissenbaum, H. 2015. Big Data's End Run around Anonymity and Consent. In Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. (eds), Privacy, big data, and the public good : frameworks for engagement (Cambridge University Press).
- Barocas, S. and, Selbstr, A.D. 2016. Big Data's Disparate Impact. 104 (3) California Law Review 671-732.
- Barse, E. L., H. Kvarnstrom, and E. Jonsson. 'Synthesizing Test Data for Fraud Detection Systems'. In 19th Annual Computer Security Applications Conference, 2003. Proceedings, 384–94, 2003. https://doi.org/10.1109/CSAC.2003.1254343.
- Binns, R. et al. 2018. "It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions. ArXiv:1801.10408 [Cs], 1–14. https://doi.org/10.1145/3173574.3173951.
- Bostrom, N. 2016. Superintelligence paths, dangers, strategies. Oxford, Oxford University Press.

- Boyd, D. and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. 15(5) Information, Communication, & Society 662–679
- Brauneis, R., and Goodman, E.P. 2018. Algorithmic Transparency for the Smart City. Yale J. L. & Tech. 20: 103–76.
- Bray, P. et al. 2015. International differences in ethical standards and in the interpretation of legal frameworks SATORI Deliverable D3.2 http://satoriproject.eu/work_packages/legal-aspects-and-impacts-of-globalization/.
- Brundage, M. et al. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', February 2018. https://maliciousaireport.com/. 56, 93
- Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms 3(1) Big Data & Society https://doi.org/10.1177/2053951715622512.
- Burt, A., Leong, B. and Shirrell, S. 2018. Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models. Future of Privacy Forum.
- Calo, Ryan. 2013. Consumer Subject Review Boards: A Thought Experiment. 66 Stan. L. Rev. Online 97 http://www.stanfordlawreview.org/online/privacy-and-big-data/consumer-subject-review-boards accessed 23 February 2018.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st Annual SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730.
- Citron, D.K. and Pasquale, F. 2014. The Scored Society: Due Process For Automated Predictions. 89 Wash. L. Rev. 1–33
- CNIL - LINC. 2017. La Plateforme d'une Ville Les Données Personnelles Au Coeur de La Fabrique de La Smart City. https://www.cnil.fr/sites/default/files/atoms/files/cnil_cahiers_ip5.pdf.
- CNIL. 2017. How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence. Report on the Public Debate Led by the French Data Protection Authority (CNIL) as Part of the Ethical Discussion Assignment Set by the Digital Republic Bill', December 2017, p. 14. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf .
- CNIL. 2017. How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence. Report on the Public Debate Led by the French Data Protection Authority (CNIL) as Part of the Ethical Discussion Assignment Set by the Digital Republic Bill https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.
- Commission - European Group on, Ethics in Science and, & New Technologies. 2018. Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. Retrieved from https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

- Conseil national du numérique. 2015. Ambition numérique : Pour une politique francaise et europeéenne de la transition numérique http://www.cil.cnrs.fr/CIL/IMG/pdf/CNNum--rapport-ambition-numerique.pdf.
- Council of Europe. 2017. Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a.
- Council of Europe-Committee of experts on internet intermediaries (MSI-NET). 2018. Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications. https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5.
- Cummings, M. L. , Roff, Heather M., Cukier, Kenneth, Parakilas, Jacob and Bryce Hannah. 2018. Chatham House Report. Artificial Intelligence and International Affairs Disruption Anticipated (London: Chatham House. The Royal Institute of International Affairs). https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf.
- Diakopoulos, N. 2013. Algorithmic Accountability Reporting: on the Investigation of Black Boxes (Tow Center for Digital Journalism).
- DNA Web Team, 'Google drafting ethical guidelines to guide use of tech after employees protest defence project' DNA India (15 April 2018) http://www.dnaindia.com/technology/report-google-drafting-ethical-guidelines-to-guide-use-of-tech-after-employees-protest-defence-project-2605149.
- Donovan, J., Matthews, J., Caplan, R. and Hanson, L. 2018. Algorithmic Accountability: A Primer. https://datasociety.net/output/algorithmic-accountability-a-primer/.
- Edwards, L. and Vale, M. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. (2017) 16(1) Duke Law and Technology Review 18-84.
- European Commission - European Group on, Ethics in Science and, & New Technologies. 2018. Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.
- European Commission. 2018. The European Artificial Intelligence Landscape. https://ec.europa.eu/digital-single-market/en/news/european-artificial-intelligence-landscape.

- European Data Protection Supervisor - Ethics Advisory Group, 2018. Towards a digital ethics https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf.
- European Data Protection Supervisor. 2016. Opinion 8/2016. EDPS Opinion on coherent enforcement of fundamental rights in the age of big data.
- European Economic and Social Committee. 2017. The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.
- European Parliament. 2017. European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI)) http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0076+0+DOC+XML+V0//EN&language=EN.
- European Union Agency for Fundamental Rights (FRA). 2018. #BigData: Discrimination in Data-Supported Decision Making http://fra.europa.eu/en/publication/2018/big-data-discrimination.
- Executive Office of the President, and National Science and Technology Council - Committee on Technology. 2016. Preparing for the Future of Artificial Intelligence (Washington D.C.) https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.
- Federal Ministry of Transport and Digital Infrastructure. Ethics Commission Automated and Connected Driving. Accessed 17 July 2018. http://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.
- Gama, J. et al. 2013. A survey on concept drift adaptation. ACM Computing Surveys 1 (1) http://www.win.tue.nl/~mpechen/publications/pubs/Gama_ACMCS_AdaptationCD_accepted.pdf.
- Goodman, B. and Flaxman, S. 2016. EU Regulations on Algorithmic Decision-Making and a "right to Explanation". [2016] arXiv:1606.08813 [cs, stat] http://arxiv.org/abs/1606.08813.
- Hildebrandt, M. 2016. Smart Technologies and the End(s) of Law : Novel Entanglements of Law and Technology (Edward Elgar Publishing).
- IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE,

2016.
http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.IEEE
- Information Commissioner's Office. 'Big Data, Artificial Intelligence, Machine Learning and Data Protection', 2017. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.
- ITU, 2017. AI for Good Global Summit Report 2017 https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf.
- Kaye, J. et al. 2015. Dynamic consent: a patient interface for twenty-first century research networks. 23 (2) European Journal of Human Genetics 141
- Kurzweil, R. 2016. The singularity is near : when humans transcend biology (London : Duckworth, 2016).
- Linnet, T., Floridi, L. and van der Sloot, B. (eds). 2017. Group Privacy: New Challenges of Data Technologies (Springer International Publishing).
- Lipton, Z.C. 2018. The Mythos of Model Interpretability. In Machine Learning, the Concept of Interpretability Is Both Important and Slippery. ACMQueue, 16 (3), https://queue.acm.org/detail.cfm?id=3241340.
- Lomas, N., 'DeepMind now has an AI ethics research unit. We have a few questions for it…' TechCrunch (4 October 2017) http://social.techcrunch.com/2017/10/04/deepmind-now-has-an-ai-ethics-research-unit-we-have-a-few-questions-for-it/ accessed 3 May 2018.
- Lycett, M. 2013. Datafication: making sense of (big) data in a complex world (2013) 22 (4) European Journal of Information Systems 381–386
- Mantelero A. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. Assessment. Computer Law & Security Review (2018), https://doi.org/10.1016/j.clsr.2018.05.017.
- Mantelero, A. 2014. The future of consumer data protection in the E.U. Rethinking the "notice and consent" paradigm in the new era of predictive analytics. Computer Law and Security Review, 30 (6): 643-660.
- Mantelero, A. 2017. Regulating Big Data. The guidelines of the Council of Europe in the Context of the European Data Protection Framework' (2017) 33(5) Computer Law & Sec. Rev. 584-602.
- Mayer-Schönberger, V. and Cukier, K. 2013. Big Data. A Revolution That Will Transform How We Live, Work and Think (London : John Murray).
- McCulloch, W.S. and Pitts, W.H. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5:115-133, 1943.
- Office of the Privacy Commissioner of Canada. 2016. The Internet of Things: An Introduction to Privacy Issues with a Focus on the Retail and Home Environments

https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2016/iot_201602/#heading-0-0-2-15.
- Omer, T. and Polonetsky, J. 2012. Privacy in the Age of Big Data. A Time for Big Decisions. 64 Stan. L. Rev. Online 63–69.
- O'Neil, C. 2017. Weapons of math destruction (London : Penguin Books, 2017).
- Palm, E. and Hansson, S.O. 2006. The case for ethical technology assessment (eTA). 73(5) Technological Forecasting & Social Change 543, 550–551.
- Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability https://ainowinstitute.org/aiareport2018.pdf.
- Rossi, F. 2016. Artificial Intelligence: Potential Benefits d Ethical Considerations' (European Parliament: Policy Department C: Citizens' Rights and Constitutional Affairs 2016) Briefing PE 571.380 http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf.
- Rouvroy, A. 2016. "Of Data and Men": Fundamental Rights and Liberties in a World of Big Data' https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806a6020.
- Rubinstein, I.S. 2013. Big Data: The End of Privacy or a New Beginning?. 3 (2) International Data Privacy Law 74–87.
- Selbst, A.D. 2017. Disparate Impact in Big Data Policing. Georgia Law Review 52 (1), 109–195.Selbst, Andrew D. and Powles, Julia. 2017. Meaningful Information and the Right to Explanation. 7(4) International Data Privacy Law 233–242
- Sheehan, M. 2011. Can Broad Consent be Informed Consent? (3) Public Health Ethics 226–235.
- Speikermann, S. 2016. Ethical IT Innovation. A Value-Based System Design Approach (Boca Raton : CRC Press)
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. 2013. Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199.
- The Danish Institute for Human Rights. 2016. Human rights impact assessment guidance and toolbox (The Danish Institute for Human Rights, 2016) https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-and-toolbox.
- The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

- The Norwegian Data Protection Authority. 2018. Artificial Intelligence and Privacy Report. https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.
- Turing, A. M. 1950. Computing Machinery and Intelligence. 49 Mind 433–460. UK Department for Digital, Culture, Media & Sport. 'Data Ethics Framework - GOV.UK'. https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.
- United Nations Office of the High Commissioner for Human Rights. 2006. Frequently asked questions on a human rights-based approach to development cooperation' (New York and Geneva: United Nations).
- United Nations, 2011. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. United Nations Human Rights Council (UN Doc. HR/PUB/11/04).
- Veale M., Binns R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2):2053951717743530, https://doi.org/10.1177/2053951717743530.
- Veale, M., Binns, R. and Edwards, L. 2018. Algorithms That Remember: Model Inversion Attacks and Data Protection Law. Philosophical Transactions of the Royal Society, Forthcoming 2018, https://doi.org/10.1098/rsta.2018.0083.
- Villani, C. 2018. For a Meaningful Artificial Intelligence towards a French and European Strategy https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.
- Wachter, S., Mittelstadt, B. and Floridi, L. 2017. Why a right to explanation of automated decision - making does not exist in the General Data Protection Regulation. 7(2) International Data Privacy Law 76–99.
- Walker, S.M. 2009. The Future of Human Rights Impact Assessments of Trade Agreements. (Utrecht: G.J. Wiarda Institute for Legal Research) https://dspace.library.uu.nl/bitstream/handle/1874/36620/walker.pdf?sequence=2.
- White House. 2015. Consumer Privacy Bill of Rights. §103(c) (Administration Discussion Draft 2015. https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf.
- Wight, D. and Mordini, E. 2012. Privacy and Ethical Impact Assessment. In Wright, D. and De Hert, P. (eds) Privacy Impact Assessment (Springer Dordrecht) 397–418.
- World Economic Forum. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf.

- Wright, D. and De Hert, P. (eds). 2012. Privacy Impact Assessment (Springer Dordrecht).
- Wright, D. 2011. A framework for the ethical impact assessment of information technology. 13 Ethics Inf. Technol. 199, 201–202.