HAL open science

From Privacy to Algorithms' Fairness

Chiara Sabelli, Mariachiara Tallacchini

To cite this version:

Chiara Sabelli, Mariachiara Tallacchini. From Privacy to Algorithms' Fairness. Marit Hansen; Eleni Kosta; Igor Nai-Fovino; Simone Fischer-Hübner. Privacy and Identity Management. The Smart Revolution : 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Ispra, Italy, September 4-8, 2017, Revised Selected Papers, AICT-526, Springer International Publishing, pp.86-110, 2018, IFIP Advances in Information and Communication Technology, 978-3-319-92924-8. 10.1007/978-3-319-92925-5_7 . hal-01883619

HAL Id: hal-01883619

https://inria.hal.science/hal-01883619v1

Submitted on 28 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

From privacy to algorithms' fairness

Chiara Sabelli, PhD[1], and Mariachiara Tallacchini, PhD[2]

[1] Freelance science writer. Website: http://chiarasabelli.it/
chiara.sabelli@gmail.it
[2] Facoltà di Economia e Giurisprudenza, Università Cattolica S.C.,
Via Emilia Parmense 84, 29100 Piacenza, Italy.
mariachiara.tallacchini@unicatt.it

Abstract. This article aims to show how the legal and ethical debate – as far as ethics has become an indispensable complementary normative tool within legal frameworks – on the digital world in the United States (US) and the European Union (EU) has significantly opened up to include new dimensions other than privacy, particularly in connection with machine learning algorithms and Big Data. If privacy still remains the main interpretive construct to normatively forge the digital space, increasingly issues of discrimination, equal opportunity, fairness and, more broadly, models of justice, are entering the picture. While offering some examples of the inadequacy of privacy to cover new normative concerns related to Big Data and machine learning, the article also argues that attempts to grant algorithmic fairness represent just the first step in addressing the wider question about what models of digital justice we are willing to apply.

Keywords: privacy, law, ethics, fairness, equal opportunities, models of justice

1 Introduction

The soft and hard normative framework built around the concept of privacy and data protection has been critical in regulating ICT and the Internet development. The right to privacy has been early identified as having a unique potential to represent and solve the new challenges coming from information technologies. Indeed, it has been depicted not only as the major area of concern surrounding the digital world, but also as an incredibly fruitful notion in the attempt to capture and protect several aspects of human life. Due to its flexibility in making sense of a variety of human expressions, privacy has been envisioned as capable, and then charged with the role, of encompassing and responding to most issues related to the information society [25, 43].

Moreover, the attempt to shape and convey the majority of ICT ethical and legal implications through privacy and data protection has appeared (and appealed) as an effective strategy to simplify and normalize normative issues. For a long time this strategy has been successful. Only in the last decade, at first with the problems created by the Internet of Things (IoT) and then even more

with the fast rise and ubiquitous presence of Big Data, the normative landscape has appeared more complex, and new normative issues irreducible to privacy and data protection have started attracting attention.

This has been true especially in the US, where issues of equality, equal opportunity and, more generally, fairness of algorithms have become a major concern. In the EU context, also in conjunction with the broad effort to create an all-encompassing regulatory environment through the GDPR (Regulation (EU) 2016/679), until recently algorithms' fairness has emerged in very few documents.

In the current debate the US (soft law) principle of 'equal opportunity by-design' is pairing with the EU (hard law) principle of 'privacy by-design' as a new main regulatory focus for Big Data and machine learning. While privacy relates to individuals, fairness has both an individual and a collective, social dimension: namely it creates the space for a broad discussion on social choices and change.

This article aims to show how, in the United States (US) and in the European Union (EU), the legal and ethical debate - as far as ethics has become an indispensable complementary normative tool within legal frameworks – after having unified all issues under the umbrella of privacy, has significantly opened up to include new dimensions, particularly in connection with machine learning algorithms and Big Data.

If privacy still remains the main interpretive construct to normatively forge the digital space - especially in the EU –, increasingly, issues of discrimination, equal opportunity, and fairness are calling for a normative response. Some examples are offered of the inadequacy of privacy to cover situations triggered by new digital developments.

While the EU idea of fair processing of data is not entirely adequate to address algorithms' fairness, and the US concept of fairness though providing a better account of the problems involved– is mostly based on soft law measures, an important effort is devoted to designing algorithms more responsive to fairness. The article ends with the suggestion that algorithmic fairness should be coupled with democratic awareness and discussion of models of digital justice as the fundamental value towards change.

The article is structured as follows. Section 2 provides an overview of the historical evolutions of privacy legislation in the US and the EU. Section 3 shows how the notion of privacy as autonomy has become increasingly relevant when dealing with digital data and communications. In Section 4 we illustrate the crucial role of algorithms designed to analysing digital data and how they raise issues related to discrimination and equality of opportunity. In Sections 5 and 6 we look at how the US and the EU have taken into account the concept of fairness in regulating personal data processing. Finally, in Section 7 we argue that, in order to design fair algorithms, more institutional, expert and public debate about a shared model of justice is needed.

2 Legal 'privacies' in the US and the EU

Privacy should be properly referred to in the plural form of 'privacies'. Indeed, it has been widely recognized that privacy meanings and contents are multifaceted and filled with semantic ambiguities and developments that have also led to the splitting of what relates to private individual life, family and correspondence, from what concerns private or sensitive data. Widely diverging meanings concern what is felt as private at individual and collective level as well as what refers to individual and collective identities, ethical systems, and cultures [23].

In building the concepts of privacy in relation with their structures and ways of functioning, legal rules have actually increased the variety of 'privacies', as heterogeneous notions characterize different legal systems. While it is already difficult to generalize in mentioning a single European idea of privacy and while past and current debates on what is privacy and how to protect it in the US are still broadly open, the prism of the legal framework surrounding privacy and data protection allows to capture at least some relevant discrepancies. What follows is a very brief summary of some major differences between the European and US notions of privacy and their current interactions[1].

As known, privacy has been framed as a right in the US in connection with technological development in terms of being let alone [52], namely primarily as a right to autonomy[2]. After a few US Supreme Courts landmark decisions establishing limits and criteria for public authorities to enter the private sphere[3], at the beginning of the 1970s ICT, and later the Internet, shed new light on privacy [16]. Data, and especially personal data, has rapidly become not only a separate, new reality, but also a commodified entity connecting different aspects of social life (from consumption to credit to health). Even more, in the construction of the information society, data has emerged as a new transversal language through different disciplines and technologies, from the life sciences to the sciences of the artificial [6]. These changes have triggered the enactment of some soft law principles, together with some case-by-case binding legislation in sectors where consumers/users' rights appeared more vulnerable[4].

---

[1] As the shift from the Safe Harbor to the Privacy Shield shows.

[2] Warren and Brandeis 1890, 194: "Recent inventions and business methods call attention to the next step which must be taken for the protection of the person, and for securing to the individual what Judge Cooley calls the right 'to be let alone'."

[3] Such as, e.g., Olmstead v. United States 277 U.S. 438 (1928) and Katz v. United States 389 U.S. 347 (1967).

[4] These laws include: Fair Credit Reporting Act, 15 U.S.C. 1681 et seq.; Privacy Act of 1974, 5 U.S.C. 552a, as amended; Family Educational Rights and Privacy Act of 1974, 20 U.S.C. 1232g; Electronic Communications Privacy Act of 1986, 18 U.S.C. 2510-22; Video Privacy Act of 1988, 18 U.S.C. 2710; Children's Online Privacy Protection Act of 1998, 15 U.S.C. 65016506; Gramm-Leach-Bliley Act of 1999, the Financial Services Modernization Act (Public Law 106-102, 106th Congress); the Genetic Information Nondiscrimination Act of 2008 (Public Law 110233, 122 Stat. 881).

The first main policy document on privacy, proposed by the Federal Trade Commission [28] – which content was later endorsed by the OECD [39] – enounced five essential components for privacy protection: consumers should be given notice of an entity's information practices before any personal information is collected from them; consumers should express their choice through consent; they should be able to access information concerning them and to contest data accuracy and completeness; data should be accurate and secure; privacy protection can only be effective if an enforcing mechanism is in place. These principles, known as the Fair Information Practice Principles (FIPPs) became the basis for the Privacy Act of 1974 and widely inspired several legislations.

Though not a constitutional right, privacy has been constitutionally rooted primarily in Amendment 4 of the US Constitution, stating "the right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches;" and also, as to the procedural aspects, in Amendment 14, according to which States will not "deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws."

Notwithstanding these constitutional foundations, a lot of room has been left for self-regulatory measures, voluntary adopted by corporations in their privacy policies; and a widely shared feeling, both within institutions and scholarly literature, surrounds the idea that privacy – which still remains the prevailing way to refer to both privacy and data protection – should not be pervasively regulated. While some documents suggest that "regulatory parsimony" should remain the rule, namely "only as much oversight as is truly necessary to ensure justice, fairness, security, and safety while pursing the public good" [43, p.25]; various new approaches to privacy point at new solutions. Some recommend that data, at least in research, should be regulated under property and autonomy models of privacy [36, 46]. Other approaches, concerned about preserving private life together with public goods, call for more empowerment and control for users in order to promote altruism and to leverage "inequity aversion, reciprocity, and normativity to lessen exploitation among group members", in [26, p.396] and [47]. Others warn about a new "tragedy of the commons," but there is no shared view about which commons are at stake. Some authors think of the tragedy of the social "trust commons", diminished by corporate misbehaviors and requiring new business models [31]; others see it as the data commons, threatened by people removing their data (especially in research) as they do not believe in anonymization and ignore that "the collective benefits derived from the data commons will rapidly degenerate if data subjects opt out to protect themselves" [54, p.4]. In many ways the current US debate on privacy reflects the fluctuations between an individualistic culture and the increasing tendency towards forms of solidarity, sharing, and even a new philanthropic anthropology [4, pp. 2-3][5].

[5] As Benkler has pointed out, "(f)or decades economists, politicians and legislators, business executives and engineers have acted as though all systems and organizations had to be built around incentives, rewards, and punishments in order to get people

Compared to the US context, the European Union regulatory landscape has been much more dependent on formal legislation and on the strong foundation of privacy as a fundamental right. In the immediate aftermath of World War II – and in the light of Hannah Arendt philosophical reflections on privacy as the main tenet for allowing individuals to emerge as unique persons in the world [2] – the Council of Europe (CoE) Convention on Human Rights (1950) has established privacy as everyone's right "to respect for his private and family life, his home and his correspondence" (Article 8). This right was split into two separate dimensions related to the person and their personal data in the Charter of Fundamental Rights of the European Union at Articles 7 and 8. While Article 7 recalls the respect for private and family life, home and communications; Article 8 establishes the protection of personal data as fundamental, and introduces the criteria of fair processing for specified purposes, consent of the person or some other legitimate basis laid down by law.

The CoE Convention of 1981 on the Automatic Processing of Personal Data[6] – currently under amendment[7] – and Directive 1995/46/EC[8], together with other directives, have provided harmonized legislation on data protection in a constantly changing sociotechnical scenario, that the new GDPR is meant to address through new constructs, requirements, sanctions[9].

While the GDPR is built around a bundle of traditional and new principles (the major are lawfulness, fairness and transparency) – interactively used to cover the many detailed aspects of a complex regulation –the principle of Privacy by Design (PbD), or, in the GDPR, Data Protection by Design (DPbD) (Article 25), can be at least metaphorically identified as the unifying symbol that synthesizes the EU regulatory approach. In fact, it refers to protection embedded in the entire cycle of data processing, "to integrate the necessary safeguards into

---

to achieve public, corporate, or community goals (...). And yet all around us we see people cooperating and working in collaboration, doing the right thing, behaving fairly, acting generously, caring about their group or team, and trying to behave like decent people who reciprocate kindness with kindness" [4, p.2-3].

[6] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No.108, Strasbourg, 28/01/1981. Convention 108 was the first and is still today the only binding international legal instrument in the field of data protection.

[7] Parliamentary Assembly of the Council of Europe, Committee on Legal Affairs and Human Rights, Draft Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108) and its Explanatory Report, Doc. 14437, 15 November 2017.

[8] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L 281, 23/11/1995, 31-50.

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, 4.5.2016 L 119/1.

the processing in order to meet the requirements of this Regulation and protect the rights of data subjects" (Article 25).

Within this increasingly articulated European legal framework, populated by a variety of specific directives – and with all the national implementations – for a long time soft law and ethics have been playing a modest role in data legal protection' - the main European expression for citizens' personal information in the digital age. With a remarkably different approach as to the biotechnological domain, law had the primacy over ethics in privacy matters [8, 50, 51]. In the past few years, and primarily in relation to ICT fast developments and pervasive implications for people's lives (e.g. social networks and the Internet of Things), other ethical aspects (e.g. identity, agency, surveillance) have emerged. In 2012 for the first time the European Group on Ethics in Science and New Technologies (EGE) was asked to address the ethical issues raised by ICT, which were essentially identified (again) in privacy and the (traditionally bioethical) protection of individual dignity [25].

Most interpretive and often 'ethical' institutional reflection has been performed through Art.29 Working Party (Art.29 WP) and the European Data Protection Supervisor (EDPS) opinions. Only in 2015 a dedicated ethics committee was established within the EDPS to explore privacy and data protection ethics [22]. The Advisory Group is deemed to help define a "new digital ethics" that should combine the benefits of technology for society and the economy with a reinforcement of individual rights and freedoms. However, a general rise of soft law as an active complements to binding regulation is taking place worldwide [23], "a supplementary approach" where "regulators should encourage businesses to adopt new business models" [45, see also 5 on binding corporate rules]. This 'ethical turn' in privacy matters comes from the awareness that "in today's digital environment adherence to the law is not enough; we have to consider the ethical dimension of data processing" [Buttarelli 2016]. Indeed, also in Europe the impression is growing that development such as Big Data is exceeding the GDPR reform efforts. However, the attempt to show that privacy remains an all-encompassing concept, the main concern and the answer, is still strong. As the EDPS pointed out in 2015, because privacy is now more than ever connected to human dignity, and human dignity is the most fundamental human value and right, privacy and data protection are becoming immediate synonyms for dignity [21, p.4; see also 9].

3 The renewed relevance of privacy as autonomy: the Facebook case and the 'filter bubble'

As said, while in a very distinct way 'data protection' is the main EU construct related to personal information and ICT, in the US 'privacy' has remained the broad term to refer also to individual data protection in the digital domain.

Indeed, even though no 'consensus definition' exists, in the US legal context privacy is multidimensional and includes physical, informational, decisional, proprietary, associational, and intellectual aspects [1]. Therefore, privacy is "a gen-

eral concept that includes confidentiality, secrecy, anonymity, data protection, data security, fair information practices, decisional autonomy, and freedom from unwanted intrusion" [43, p.25].

This is why the right to personal autonomy and self-determination has been protected for a long time through the concept of privacy. In the famous case *Roe v. Wade* (410 U.S. 113 (1973)), where the US Supreme Court had to legally frame abortion, privacy was called in as the foundation for an autonomous private decision. While recognizing that the right to privacy was not explicitly mentioned in the US Constitution[10], the Court stated that "the right of privacy, whether it be founded in the Fourteenth Amendment's concept of personal liberty and restrictions upon state action, as we feel it is, or, as the District Court determined, in the Ninth Amendment's reservation of rights to the people, is broad enough to encompass a woman's decision whether or not to terminate her pregnancy" (at VIII, 153).

Privacy as the entitlement to autonomous decisions, however, is not foreign to the European legal context and had a remarkable application by the European Court of Human Right (ECHR) in 1998. While the ECHR has intervened on the right to privacy in a number of cases, the Case of Guerra and Others v. Italy remains quite unique in the way respect for private life (Article 8) has been interpreted. The controversy followed an industrial accident in Manfredonia where the population, exposed to the effects of toxic chemical substances, accused Italian institutions of having infringed their right to privacy by not providing people with sufficient information about the situation.

In rejecting the arguments of the Italian Government, the Court not only recognized the applicability of Article 8, but also stated that this was not a merely negative right of non-interference, but instead should be seen as a positive right. "(A)lthough the object of Article 8 is essentially that of protecting the individual against arbitrary interference by the public authorities, it does not merely compel the State to abstain from such interference: in addition to this primarily negative undertaking, there may be positive obligations inherent in effective respect for private or family life. (...) The Court holds, therefore, that the respondent State did not fulfil its obligation to secure the applicants' right to respect for their private and family life, in breach of Article 8 of the Convention" (§58; 60).

If the original meaning of the right to privacy as autonomy and integrity of the personal sphere belongs to the US legal landscape, and is not foreign to the European context - even more as a positive right –, this dimension is acquiring a

[10] Both courts and scholars have shown that the US Constitution implicitly recognizes the value of privacy and rights of privacy through provisions guaranteeing: 1) freedom of speech, freedom of religious, political and personal association, and related forms of anonymity (First Amendment); 2) freedom from government appropriation of one's home (Third Amendment); 3) freedom from unreasonable search and seizure of one's body and property (Fourth Amendment); 4) freedom from compulsory self-incrimination (Fifth Amendment); 5) freedom from cruel and unusual punishment, including unnecessarily extreme deprivations of privacy (Eight Amendment); and 6) other personal freedoms (Ninth Amendment).

renewed relevance in the light of the digital interferences in human life. And, in this respect, it is quite interesting that, in their Opinion 7/2015 on the challenges of Big Data, the EDPS has observed that "(t)he right to be let alone is indeed the beginning of all freedom" [21, p.1].

The renewed relevance of privacy as the right to preserve the 'integrity' of the personal and intimate sphere, has become evident, even when personal data are not at stake, as, e.g., in the Facebook-Cornell case of psychological contagion and in what is called the 'filter bubble'.

In 2014 a group of researchers, from the Data Science Team of Facebook Inc. and the Information and Communication Departments of Cornell University, published the results of an experiment conducted on 689,000 subjects using Facebook in English [35]. During one week in January 2012, researchers manipulated the extent to which users were exposed to 'emotional' communication in their News Feed in order to understand whether emotions can be transferred to others without any direct contact. Two separate experiments were conducted, one for positive emotions and the other for negative ones. In each experiment the users were divided into an experimental group and a control group. Each post was analyzed by a text mining algorithm which classified its content in three categories (positive emotion, negative emotion, not emotional) and, with some probability, deleted it if it contained a positive (negative) emotion in the experimental group, whereas it was deleted completely at random for the users in the control group. The researchers found that users with reduced exposure to 'positive' posts produced less 'positive' posts, and in turn wrote more 'negative' posts. The opposite happened for the users with reduced exposure to 'negative' posts.

The participants were completely unaware of having been enrolled in an experiment. When the results appeared in the Proceedings of the National Academy of Science (PNAS) questions were raised about the principles of informed consent and opportunity to opt out, in line with the rules on Human Research Subjects.

PNAS published an Editorial Expression of Concern where, while recognizing that the experiment "involved practices that were not fully consistent" with the relevant principles, it was argued that, being Facebook a private company, the agreement to the Data Use Policy, to which all users agree prior to creating an account on Facebook, constituted informed consent for the research. Moreover, according to PNAS, users' privacy was never violated since the posts were never analyzed by researchers, but processed by a text mining system which kept the whole process blind to humans. However, despite these explanations, it became apparent that, as users' personal sphere of emotions had been violated, privacy was at stake.

Something similar, even outside the research context, is happening constantly with Google search engine and personal accounts on social network platforms. The contents users are exposed to are selected based on their past preferences, with the aim of offering them the most appropriate and enjoyable content. This process gradually encloses everyone inside a 'filter bubble' [41] where they are

mostly exposed to opinions similar to those they already have. Search engines and news feed algorithms are acting as editors, they design the information diet, and this has a big influence on the way users see the world and on the related choices they make.

4 How algorithms learn from the past

The huge quantity of data produced by users, mostly digital data, would be of no use without algorithms capable of analyzing them to extract valuable information. When dealing with large quantity of data coming from different sources, some specific types of algorithms, namely machine learning programs, are needed. Unlike conventional computer programs, these algorithms are only partially pre-programmed. Indeed the most common ones learn from examples offered by historical data. By using previous examples, they produce predictions about the future. This feature of machine learning algorithms is one of the most relevant in this context: algorithms processing historical data detect patterns inside them. These patterns, for example correlation between variables, often incorporate the prejudices and inequalities of our societies. Once the algorithms have learned these patterns, they will use them to make future decisions, thus perpetuating the biases on a much larger scale.

The idea that computer-based decisions would be more objective than human decisions in solving controversial issues has already proved not only fallacious, but often also shortsighted. This is why the attempt to disentangle algorithms and data is neither possible nor meaningful in order to understand how Big Data are changing social life and threatening democratic values.

A simple example of a machine learning algorithm, namely a naive Bayesian filter, may help clarifying what learning 'from historical data' means and implies; and it helps ask whether this use of the term 'learning' is appropriate to what machine learning algorithms do.

4.1 A simple example of machine learning algorithm: the 'anti-spam' filter

Email spam filters represent a specific type of machine learning systems, called *classifiers*. Each time new email message arrives, the classifier decides if it is 'spam' or 'ham', *i.e.* legitimate email. This a very simple case of classifier, which discriminates between only two categories. More complex classifiers are used, for example, to identify the blood type, where the categories are four: 'A', 'B', 'AB' or '0'. The spam filters cannot be pre-programmed, as there is no complete agreement about what an undesired message is and also because the characteristics of spam messages evolve in time. For this reason the program is designed so that it learns by continuously analyzing the user's message flow, helped also by the users when they explicitly mark a message as 'spam' or 'ham'.

Anti-spam filters work with probabilities. The classifier computes the probability of a message being 'spam' and, if it is above some predefined threshold, e.g. 95%, it places the message in the 'spam' folder or deletes it.

Machine learning is used exactly to compute these probabilities, as the case of a naive Bayesan filter shows. After having checked the sender and other attributes of the new incoming message M, the anti-spam filter analyzes the text T with the objective of computing

$$P(\text{M is spam}\big|T) : \text{probability that M is 'spam' if it contains the text T.} \quad (1)$$

In order to evaluate this probability, the filter will look for words frequently used in spam emails. For the sake of simplicity we can interprete the text T of a new incoming message as a list of words: $T = (W_1, W_2, \ldots, W_N)$, and assume that we can factorize the probability above as the product of the single probabilities associated to each word:

$$P(\text{M is spam}\big|T) \sim P(\text{M is spam}\big|W_1)\times P(\text{M is spam}\big|W_2)\times\cdots\times P(\text{M is spam}\big|W_N), \quad (2)$$

where

$$P(\text{M is spam}\big|W_i) : \text{probability that M is 'spam' if it contains the word } W_i \quad i = 1, \ldots, N. \quad (3)$$

It is important to stress here that in this simple example the semantic structures connecting the words inside a text are being neglected. We are oversimplifying the problem with the aim of obtaining simple mathematical expressions whose meaning should appear as clear as possible.

Thanks to the spam reporting activities of all the users of the same email provider, an estimate can be obtained of the probability that W_i appears in a spam message: $P(W_i|\text{M is spam})$. A very simple estimate of this quantity is the frequency of the word W_i appearing in the messages reported as spam by the users:

$$P(W_i|\text{M is spam}) = \frac{\text{number of spam messages containing the word } W_i}{\text{total number of spam messages}} \quad (4)$$

What we are looking for, however, is the probability that the incoming message is 'spam' given that it contains the word W_i, defined in Eq. (3). To do this one can resort to Bayes theorem:

$$P(\text{M is spam}|W_i) = \frac{P(W_i|\text{M is spam}), P(\text{M is spam})}{P(W_i)} \quad (5)$$

where $P(\text{M is spam})$ is the *a priori* probability that an incoming message is 'spam' and $P(W_i)$ is the *a priori* probability that the word W_i appears in email messages. Every time a user reports a message as 'spam', the filter will first compute a new value of $P(W_i|\text{M is spam})$ according to Eq. (4), then it will substitute this new estimate inside Eq. (5). This is the *learning* phase of the algorithm: as soon as new data is recorded, the algorithm will update the probabilities.

After having repeated this procedure for each word in the text, and having combined the results in Eq. (2), the algorithm will finally compute $P(\text{M is spam}|T)$

and, depending on the chosen threshold, it will classify the new incoming message. However, even if $P(\text{M is spam}|T)$ is high, e.g. 98%, the chance still exists for the message to be 'ham'. In this case the classifier will return a so called 'false-positive', a legitimate message marked incorrectly as 'spam'. This simple example shows that machine learning algorithms commit errors, because they work with probability and statistics and, as long as this approach is applied, these errors are implicitly accepted.

4.2 Machine learning in the social realm

Machine learning algorithms are at work in many sectors of our society: credit access, finance, insurance, health, policing, justice, human resources management [10], access to education[11].

Two relevant examples illustrate the potential negative effects that algorithms can have when used to assist decision making on very delicate issues.

More than fifty US Police Corps currently use a *predictive policing* software called PredPol, an algorithm designed by the consultancy company PredPol, founded in 2011 by two researchers at the University of California Los Angeles. The software is adapted from an earthquake prediction model: it takes only three variables about each crime committed in a certain area (location, time, type of the crime). It then aggregates crimes committed in the past and predicts the areas where is more likely that a crime will be committed in the near future. These predictions are used to optimize the patrolling strategy of the police officers.

The overall accuracy of the software has never been evaluated by an independent party, but in 2016 the Human Rights Data Analysis Group published a research [37] on the risk of racial discrimination in software outputs. The study compared the predictions on illegal drug consumption in the city of Oakland, California, made, respectively, by PredPol and by the Department of Health and Human Services (DHHS). While according to Health and Human Services illegal drug consumption is homogeneously distributed among neighbourhoods, PredPol estimated a higher risk in the areas inhabited by African Americans.

The point here is that geography in many American cities is a good proxy for race. Even though ethnicity as an explicit variable is not included in the data about past crimes, the system is able to reconstruct it because it is highly correlated with home address.

In many other countries, including Europe, Police Corps are experimenting these kinds of software despite the lack of rules on how to assess their efficacy, fairness and their economic value. In most cases this software is related to scientific research projects, that start working with institutions in order to obtain real world data to test their systems, and later develop products for the market.

The second example is the evaluation program of public school teachers in Washingto D.C.. Called IMPACT, the program started in 2011 and led, only after the first year, to nearly 200 teachers fired by the Department of Education. The program bases its decisions partly on the feedback given by experts' obser-

[11] Existing literature in the fields illustrates a variety of cases [40, 42].

vation during school classes, partly on the outcome of the *Value Added Model*, in the version devised by the consultancy company Mathematica Policy Research. The model is quite ambitious: it has the objective to estimate how big is the contribution of the teacher to the progress of his/her students from one year to the next one. The algorithm takes as input the outcomes of end-of-year tests which students receive in the different subjects, and some information about their socioeconomic and psychological background. To take into account the students' socioeconomic status, the model considers if he/she is eligible for free meals at school, whereas the psychological condition is approximated by possible learning disabilities.

The outcome of the program have been criticized by many teachers, in particular the case of Sarah Wysoki received extensive media coverage[12]. Wysocki was fired in the summer of 2011, and did not agree with this decision. She tried to understand the functioning of the algorithm, but she did not have the chanche to look at the code or to receive an exahustive explanation, neither by the Department of Education nor by the company which provided the software. This story shows how difficult is in practice to know the logic behind algorithms. This is because on the one hand the institutions that run the software do not know and master its details, and on the other because of the intrinsic complexity of some mathematical models.

Some months later an investigation of the newspaper USA Today[13] revealed that at least 70 schools in the district cheated on the end-of-year test results, thus questioning the robustness of the entire program. This is another dangerous implications of algorithmic decision-making: it depends strongly on the input data and if the subjects of the decision understand this vulnerability they will try to game the system.

5 The limits of de-identification and the differential privacy framework

Since 1990s, a big effort has been devoted by information and data science research to designing privacy preserving algorithms, *i.e.* data analysis processes able to extract valuable knowledge from a database without violating the privacy of the subjects who contributed their personal information. Some examples are provided here below.

[12] Turque, B., "Creative ... motivating and fired", *The Washington Post*, March 6, 2012. https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html?utm_term=.602a90cfd863 (accessed 8 January 2018).

[13] Toppo, G., "Memo warns of rampant cheating in D.C. public schools", *USA Today*, April 11, 2013. https://www.usatoday.com/story/news/nation/2013/04/11/memo-washington-dc-schools-cheating/2074473/ (accessed 8 January 2018).

5.1 The limits of de-identification

The first works in this field have focused on data de-identification, but the approach has quickly revealed its limits.

The case of Netflix Prize is quite famous. In 2006 Netflix launched a contest to improve the design of a 'recommendation system', an algorithm for personalized advice on movies.

The algorithm used by Netflix belongs to the *nearest neighbour* class. When users log in into their account, Netflix will provide them with some recommendations. In order to define these recommendations the system compares your rating history with other Netflix customers' history and identifies which resembles you most. Once the algorithm has identified your *nearest neighbour*, it looks for movies or TV series they rated highest and recommend them to you (these algorithms are also called *collaborative rating algorithms*).

The accuracy of the algorithm is measured by its ability to predict your next rating. The better the software can predict how you will rate different movies, the higher will be your satisfaction and thus your willingness to purchase more products. In 2006 Netflix launched a contest among developers and data scientists to improve the accuracy of its recommendation algorithm. The team that could predict customers' ratings better than the Netflix's recommender system would have been awarded one million dollars.

The training data set included 100,480,507 ratings given by 480,189 users to 17,770 movies. The dataset form's entry was ⟨user, movie, date, rating⟩ and the rating was an integer number from 1 (lowest rating) to 5 (highest rating). Once the competitors had designed the algorithm, they had to train it on the training data set, and then test it on a set containing 1,408,342 entries in the form ⟨user, movie, date, ?⟩. 'Testing' means that the algorithms run over the sample predicting the users' ratings to movies at different points in time. In order to protect privacy, both users and movies were labeled with integer numbers.

A concern for users' privacy was raised some time later, when a scientific article by two data scientists from the University of Texas at Austin [38] showed that anonymized users in the Netflix Prize dataset could be identified through the Internet Movie Database. Following this publication Netflix ended the Netflix prize.

This episode revealed the limits of de-identification procedures: are these strong enough to resist re-identification, given the large amount of redundant data people produce and make available on the Internet?

Another interesting case of privacy violation has been raised in Genome Wide Association Studies (GWAS). Sharing sequencing data sets without identifiers has become a common practice in genomics, even though in 2008 David Craig and collaborators showed that, notwithstanding the fact that GWAS released only summary statistics about their participants, information from an individual's DNA sample could determine if he/she had contributed to the study [32].

A bigger concern was raised by Gymrek and collaborators in 2013 regarding the possibility of re-identifying participants to a GWAS, using information publicly available on the Internet through surname inference [29]. Surnames are

paternally inherited in most human societies and thus one can find a correlation between surnames and Y-chromosome haplotypes (the non-recombining portion of the Y chromosome which is passed, almost unchanged, from father to son). Based on this fact, a number of genetic genealogy projects have flourished, with the aim of reconnecting distant patrilineal relatives. Currently there are at least eight databases, containing hundreds of thousands of surname-haplotype records. These databases are publicly available on the Internet and some are free-of-charge. Using these records researchers have positively matched 12% of male genomes with the exact person they originated from.

5.2 Differential privacy and beyond

The GWAS example shows that, if data controllers release too much statistical information about a sample or if redundant information is publicly available on the same subject, re-identification is possible and the privacy of the subjects who contributed to the study will be violated.

Differential privacy has been proposed to answer the question "How can a curator release only the information about a population which does not compromise participants' privacy?". As Dwork and others have explained, the basic idea in differential privacy is that the statistical information disclosed should not change if a single individual is or is not included in a data set. A survey of the results about differential privacy can be found in [17].

However, as said, even differential privacy falls short when dealing with cases such as illegal drug use in the city of Oakland, where privacy, as protection of private data, cannot prevent discrimination from machine learning algorithms. This is because even if sensible variables are not included, other variables can reconstruct them.

A further and more profound reason exists to say that preserving privacy alone will not solve the problem of unfair algorithms. Cynthia Dwork has made this point very clear[14]. If a clinical study on 1,000 subjects founds that smoking is associated with increased risk of lung cancer, every smoker will be affected by these findings, even if he/she neither participated in the study nor disclosed any personal information. The insurance premium will rise for all smokers because the study finds that they are more exposed to the risk of developing lung cancer. Whenever we can be identified as members of a certain population, we will be affected by the findings of the algorithms analyzing a sample which is considered representative of that population.

[14] Research seminar given in Novembre 2016 at the Institute of Advanced Studies, Princeton during the "Differential Privacy Symposium: Four Facets of Differential Privacy", "The Definition of Differential Privacy", November 12, 2016. https://www.youtube.com/watch?v=lg-VhHlztqo (accessed 17 December 2017).

6 From privacy to fairness: the challenges of machine learning

Already in 1980, in "Do Artifacts have Politics?" Langdon Winner highlighted that all machines, structures and technical systems should be assessed "for the ways in which they can embody specific forms of power and authority" [53, p.21]. These early observations have raised awareness about the choices implicitly embedded, packed, and black-boxed in programs and devices, showing that "architecture matters" [Kroes 2011], namely that ICT structures do not only have ethical and policy impacts, but have built-in values and choices that should be opened-up and unpacked. In other words, digital architectures embody rules and values in their hard and soft structures as "factual normativity," norms written as digital instructions [30].

The renewed relevance of privacy as autonomy and integrity of the personal sphere has revealed that the pervasive virtual' dimensions in human life call for more diversified and complex ways of capturing and humanizing the digital. Big Data and machine learning have strongly impacted the capability of extending privacy towards further meanings and frontiers. While this awareness started quite early in the US, EU institutions, also due to their regulatory effort, only quite recently have become active on the subject. In the US, at both the institutional and academic level, the awareness that some new issues raised by digital transformations could not be framed in terms of privacy began in the early 2010s, with the social applications of algorithms combining Big Data and machine learning. The discriminatory results produced by supposedly neutral artefacts became increasingly apparent, triggering institutional concern, academic conferences[15], and science journalism analyses and responses[16].

In the field of genomic research, in 2012 the Presidential Commission for the Study of Bioethical Issues (PCSBI) added new principles to privacy, namely 1) public beneficence, 2) responsible stewardship, 3) intellectual freedom and responsibility, 4) democratic deliberation, and 5) justice and fairness [43, p.28].

In 2013 Dwork and Mulligan highlighted that both the computer scientists and policy makers, while acknowledging concerns about discrimination, were still maintaining a narrow vision of the issues at stake, tending to position privacy as the dominant problem. However, and regrettably, they noted, "privacy controls and increased transparency fail to address concerns with the classifications and segmentation produced by big data analysis" [19, p.36; see also 18 and 20]. In the past few years the literature on algorithms' justice and fairness has vastly

[15] See, e.g. the series of conferences started in 2013 by NYU, Steinhardt School of Culture, Education, & Human Development, Governing Algorithms: A Conference on Computation, Automation, and Control (May 16-17, 2013), available at http://governingalgorithms.org (accessed 3 January 2017)

[16] Gillespie, T.: Can an Algorithm Be Wrong? Twitter Trends, the Specter of Censorship, and Our Faith in the Algorithms Around Us, Culture Digitally, October 19, (2011), available at http://culturedigitally.org/2011/10/can-an-algorithm-be-wrong (accessed 3 January 2017)

grown, consistently showing that algorithms are already governing our lives and that privacy and transparency are no longer the effective response, e.g. [40, 42].

6.1 The US (Obama) policy framework: big data challenges and 'equal opportunity by design'

The most relevant US initiatives of 'institutional awareness' stemmed from the Council of Advisors on Science and Technology of the Executive Office of President Obama that, between 2014 and 2016, published a series of policy reflections and recommendations [11–14; see also 15].

In analyzing opportunities and challenges of Big Data and machine learning, the reports clearly showed that, while the benefits of digital new developments are manifold, several risks need to be addressed, mostly dealing with the potential for discriminatory treatments and perpetuations of biases affecting decisions in several social, economic, and health sectors.

The challenges are primarily identified at two levels: the data used as inputs to an algorithm; and algorithm design, namely how an algorithm works and how knowable it is by the user - both for computational or proprietary reasons. Problems with data may consist of: poorly selected data; incomplete, incorrect, or outdated data; selection biases (data inputs not representative of a population); unintentional perpetuation and promotion of historical biases. Problems with algorithms' design can refer to: poorly designed matching systems; personalization and recommendation services narrowing instead of expanding user's options; decision-making systems assuming that correlation necessarily implies causation; data sets lacking information or disproportionately representing certain populations.

While data systems should remove inappropriate human biases, the risk exists that use of big data can contribute to systematically disadvantaging certain groups by encoding forms of discrimination into technological systems.

In order to fight against these outcomes, CAST has proposed a principle of 'equal opportunity by design' - somehow the US response to 'privacy by design'. This principle aims at designing data systems that promote fairness and safeguards against discrimination from the first step of the engineering process throughout their lifespan.

According to CAST, the framing of algorithms in the light of equal opportunity should be also accompanied by a number of policy actions, from support to research and the market to inventing better systems, to development of algorithmic auditing and external testing of big data systems, to mechanisms for transparency and accountability, to considering the roles of the government and private sector in setting the rules, to civic participation and education in computer and data science.

What is peculiar to the CAST approach is that it does not focus – at least not primarily – on the remedies and protection offered to users/citizens in dealing with data driven decision-making. Instead, it takes an upstream look at the 'politics of algorithms' as a complex and diversified involvement and recombinations of institutional, corporate, and civil society actors.

6.2 Fairness and data protection in the EU

As said, the European single normative language in dealing with the digital world has been and remains data protection; and, even though the term "fair" has been inhabiting privacy legislation for a long time, it does not seem to adequately apply to algorithmic discrimination, as it primarily refers to "fair data processing". Indeed, only recently fairness, meant as designing technologies that fully respect human rights, has started been taken into account in very few soft law European documents.

Moreover, the term "fair", having a variety of meanings, is very ambiguous.

"Fair" appears only one time in the 1981 CoE Convention at Article 5 ("Quality of data") establishing that data shall be "obtained and processed fairly and lawfully"[17]. In Directive 46/95 the "fair processing" of data is defined at Whereas 28 and 38 through several practices, that the GDPR has further elaborated in a detailed and complex set of requirements, often repeated, overlapping, and mixed with transparency. Indeed, a variety of requirements are listed (primarily at Whereas 39, 60 and 71, and Article 5) that range from the data subject's awareness of the processing and of risks and rights related to it, to the clarity and accessibility of information, the limits to purposes, the storage time, the right to erasure; from the legitimate grounds for data collection and the absence of unjustified adverse effects, to appropriate notice, transparent intention of use, reasonable handling, accuracy of data and rectification of errors; compliance with the law; and more. In many ways, "fair" broadly covers all GDPR provisions, but still leaves some room for further interpretation, as a sort of open interpretive clause.

However, that "fair processing" of individual personal data does not fully respond to the issues raised by Big Data and machine learning algorithms is revealed by Article 22(1)[18], "Automated individual decision-making, including profiling" where "automated individual decision-making" concerns all activities exclusively performed by a machine, while "profiling" (as defined in Article 4) refers to data collection, automated analysis to identify correlations, and the inference from these correlations towards an individual's present or future behavior [3, p.7].[19]. Article 22(1) establishes for the data subject the "right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

A guideline published in late 2017 by Art.29 WP has clarified how to interpret and apply Article 22, also admitting that the new provisions on the risks arising from profiling and automated decision-making still concern, but are "not limited to, privacy" [3, p.4]. According to Art.29 WP, in order to be effective, the right

[17] Directive 46/95 mentions "fair" 5 times and "fairly" 2. In the GDPR "fair" has 14 occurrences, "fairly" 2, "fairness" 1, "unfair" 2.

[18] In Directive 46/95 at Article 15, "Automated individual decisions."

[19] See also: CoE, Recommendation CM/Rec(2010)13 of the Committee of Ministers to member states on the protection of individuals with regard to automatic processing of personal data in the context of profiling (adopted on 23 November 2010).

granted at Article 22 has to translate into several other rights. First of all, it includes the right to be informed (Articles 13(2) (f) and 14(2) (g)), namely to "receive meaningful information about the logic involved," "simple ways to tell the data subject about the rationale behind" the algorithm [3, pp.9;12]. Also, it involves the right to understand the significance of the envisaged consequences for the data subject, the right to obtain human intervention and the right to challenge the decision (Article 22(3)).

The effectiveness of these rights should result from the application of transparency and fairness requirements: greater accountability obligations; specified legal bases for the processing; rights to oppose profiling; and, under certain circumstances, to carry out a data protection impact assessment. Further safeguards should come from other general provisions listed at Article 5(1), namely lawful, fair and transparent processing, purpose limitation, data minimization, accuracy and storage limitation.

The credibility of all this complex narrative, and more simply its feasibility, has been strongly challenged by some commentators who argued that the complexity of algorithms - sometimes opaque even to programmers –, as well as their proprietary protection, make these promises quite unrealistic [24].

While recognizing that other measures introduced by the GDPR - such as the right to erasure, the right to data portability, privacy by design, Data Protection Impact Assessments, certification and privacy seals – can be helpful, Edwards and Veale see restrictions introduced at Article 22 as problematic in many respects: the non-enforceability of statements appearing in the recitals and not in the GDPR text, substantial legal uncertainty, the practical difficulty in knowing when or how decisions are being made, etc. [24, p.21].

Art.29 WP guideline does not convincingly overcome these objections, and keeps suggesting that transparency is the answer, claiming that "the controller should find simple ways to tell the data subject about the rationale behind (...) without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm" [3, pp.9;14]

However, if Article 22 (with all its ramifications) can be hardly seen as the adequate framework towards "more responsible, explicable, and human-centered" algorithms [24, p.19], the very concept of data protection has its own limits: a culture of better algorithms requires taking into account a full range of individual and collective rights. Moreover, within the perspective of the GDPR, most measures concern *ex-post* remedies for individuals who have undergone unfair automated processing, while Big Data and machine learning call for early, upstream analysis of digital architecture and algorithms in terms of fairness.

In this respect, the most open perspective come from some non-legally binding documents issued in 2017 by the Parliamentary Assembly of the Council of Europe and the European Parliament. In its Resolution on the fundamental rights implications of big data[20], the European Parliament, after having pro-

[20] European Parliament Resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement, P8_TA(2017)0076, (2016/2225(INI)).

vided a wide landscape of the digital era and of its challenges to traditional regulatory instruments, has shown that "it is not just a question of data protection"[21]: not only many fundamental rights, but also relevant collective values such as public trust, media freedom and pluralistic information are at stake. In the context of the Council of Europe, the Parliamentary Assembly[22] Recommendation on technological convergence, artificial intelligence and human rights - also touching on machine learning and Big Data – has evoked Article 2 of the Convention on Human Rights and Biomedicine[23], establishing the primacy of the human being "over the sole interest of society or science" as a key right; and has pointed out that "safeguarding human dignity in the 21st century implies developing new forms of governance, new forms of open, informed and adversarial public debate, new legislative mechanisms and above all the establishment of international co-operation" (at point 3).

6.3 Designing fair algorithms

Data scientists who are convinced that data protection will not prevent algorithms from having negative effects on the weaker members of our societies, have gradually moved their research towards *algorithmic fairness*[24]. An example of efforts in this new direction is the "Fairness, Accountability, Transparency in Machine Learning" Conference series started in 2014[25].

A first and very general attempt to define a 'fair algorithm' has indeed been made by Cynthia Dwork and collaborators in 2011 [18]. The basic idea they proposed is to "treat similar individuals similarly", namely to interpret fairness as 'equality'. A fair algorithm is formulated as a constrained optimization problem. The constraint is written in terms of a metric, a mathematical definition of distance. According to Dwork and collaborators (referring to Rawls [44]), the choice of the metric should be made by a regulatory body or a civil rights organization and should be public and open to debate. This framework is designed to enforce

[21] Parliament's rapporteur Ana Gomes, available at http://www.europarl.europa.eu/news/en/press-room/20170314IPR66586/big-data-ep-calls-for-better-protection-of-fundamental-rights-and-privacy (accessed 5 January 2018).

[22] Parliamentary Assembly of the Council of Europe, Recommendation 2102 (2017)1, Technological convergence, artificial intelligence and human rights, adopted by the Assembly on 28 April 2017.

[23] Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine (ETS No. 164, "Oviedo Convention").

[24] This gradual shift is well described by Cynthia Dwork herself in an interview published by Quanta Magazine. Hartnett, K. "How to Force Our Machines to Play Fair", *Quanta Magazine*, November 26, 2016. https://www.quantamagazine.org/making-algorithms-fair-an-interview-with-cynthia-dwork-20161123 (accessed 17 December 2017).

[25] Fairness Accountability and Transparency of Machine Learning: research group website https://www.fatml.org/.

'individual fairness', not 'group fairness'. As seen in previous examples, unfairness is mostly connected to a collective dimension, not only to an individual one. Members of minority groups are more likely to be discriminated by algorithms designed by members of the majority, who are not necessarily 'aware' of their existence and in some cases are not inclined to listen to their needs.

The difficulty in designing algorithms which are fair towards different social groups became apparent in the case of the recidivism risk model called COMPAS, used by many federal courts in the US and sold by the private company NorthPointe. COMPAS takes as inputs the answers given to a standard questionnaire by defendants and policemen and gives as output the risk that the defendant will commit another crime in the near future. This risk estimate helps the judge to decide the length of the sentence, whether the arrested has to be imprisoned until the beginning of the process, and the possible enrollment in some support programs.

In May 2016 ProPublica published a thorough investigation[26] stating that COMPAS fails differently for black and white defendants. In particular in order to assess the accuracy of the algorithm, they studied more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the actual rate on a two-year period. The results showed that the percentage of defendants labeled as high risk who did not commit further crimes was 23.5% among white defendants and 44.0% among black defendants. Similarly, the percentage of defendants labeled as low risk who did commit further crimes was 47.7% and 28.0%, respectively for white and black defendants[27].

The ProPublica investigation has triggered academic attempts to 'fix' the COMPAS algorithm, *i.e.* to remove its discriminatory behavior against the black population. However, for the time being these efforts have proved unsuccessful[28]. The problem is still open. In a recent contribution [55], Zafar and co-authors introduce a notion of fairness based on group's preference for being assigned one

[26] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 7 December 2017).

[27] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks", *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on 7 December 2017.

[28] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say", *ProPublica*, December 30, 2016. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say?utm_source=suggestedarticle\&utm_medium=referral\&utm_campaign=readnext\&utm_content=https\%3A\%2F\%2Fwww.propublica.org\%2Farticle\%2Fbias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say (accessed 7 December 2017).

set of decision outcomes over another, as opposed to the notion of fairness based on parity (equality) of treatment.

The high degree of complexity and specificity of the problems related to algorithmic discrimination has led many to propose the development of some forms of auditing for algorithms. During 2017 the French commission for digital rights[29] has organized a public debate on the ethical aspects of algorithms, called "Éthique Numérique". In the final reccomendations, which originated from the discussions held among professionals of different sectors as well as among citizens, one concerns the "creation of a national platform in order to audit algorithms"[30]. The idea is that not only engineers and programmers, but policy makers and citizens should become more aware of the difficulties involved.

7 Whose vision of fairness?

As a proxy for justice, equity, equality, and appropriateness, fairness opens up an umbrella of meanings and problems even wider than privacy, as visions of justice have animated philosophical, legal, and political debates of different cultural traditions since their origins. What is now at stake is that diverging visions of fairness can be opaquely and disorderly be embedded in all sort of digitalized decision making. The attempts to design fair algorithms, though controversial [27], still represent a promising way forward and may facilitate a better understanding of how to approach fairness and what it involved [49].

However, while new dialogues between the languages of justice and computer and data science need to be framed, some issues should be clarified.

A first point is the following. Attention should be paid to the fact that fair algorithms have already revealed an inclination towards privileging specific visions of fairness, namely those where the model itself (or some of its components) can be more easily quantified and translated into mathematical terms. For instance, even though Rawls's theory remains a powerful vision of justice [44], its adoption in several software mostly depends on the possibility to convert it in algorithms [34]. In other words, the appeal of visions of justice having the potential for an algorithmic definition is higher compared to other lacking this adaptability.

Already in 1979 Nobel laureate Amartya Sen, in discussing the ambiguities of equality (The Equality of What?), noted that "(t)he recognition of the fundamental diversity of human beings does, in fact, have very deep consequences, affecting not merely the utilitarian conception of social good, but (...) even the Rawlsian conception of equality. If human beings are identical, then the application of the prior-principle of universalizability in the form of giving equal weight

[29] Commission nationale de l'informatique et des libertés (CNIL).

[30] "Comment permettre à l'homme de garder la main? Les enjeux thiques des algorithmes et de l'intelligence artificielle", Summary of the public debate organized by the CNIL as established by The Digital Republic Bill (Loi pour une rèpublique numèrique). https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf (accessed 8 January 2017).

to the equal interest of all parties' simplifies enormously" [48, p.202]. And again, in its "The Idea of Justice," Sen has shown how the fundamental axioms about justice are incommensurable. Though all legitimate within their own assumptions, the choices about how to characterize individuals, how to define their similarities and how to prioritize our choices need wider and public reasons to be debated. Who is going to choose the relevant axioms? Principle of justice are plural, criteria need democratic discussion and assessment through the worlds that are generated.

And this leads to a second point. As said, ethics as a non-legally binding regulatory tool has become increasingly relevant in the digital domain, even though there is no agreement about how to implement it. While some authors suggest, for instance, private ethical auditing for algorithms [40], others point at institutional ethics committees (as in the biomedical field) [9], and others, while calling for the "adoption of a normative definition of fairness within the machine learning community", argue in favor of dialogues between machine learning experts and vulnerable populations [49]. Adequate discussion of fair algorithms requires legitimacy and should avoid both the tensions between experts and non-experts and the bureaucracy of ethics committees. The concept of Participatory Design (PD) may be helpful here as an important exploratory tool. Theories of PD, originated in the domain of Human-Computer Interactions (HCI), information systems and socio-informatics, reflect on building digital architectures through participatory procedures, with the aim of making knowledge and values embedded in technological systems more open and democratically shared. In PD users are "co-designers during all stages of the design process;" which "means that decisions about possible design trajectories should .be open to the possibility of change and in ways that enable choices to be unmade or changed" [7, p.3;6].

Indeed, digital architectures should not deterministically impose their own structures, ontologies, mechanisms, explanations over social normativity [36]; instead, in the complex evolution between technoscience and normativity, new spaces for choice and scrutiny in technosocial architectures should be made available to citizens as a matter of democracy and participation [33].

References

1. Allen, A.: Privacy Law and Society. 2nd ed. Thomson Reuters, St. Paul, MN (2011)
2. Arendt, H.: The Human Condition. University of Chicago Press, Chicago, IL (1958)
3. Article 29 (Data Protection Working Party): Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/6799, 17/EN, WP 251, Adopted on 3 October (October 2017) (accessed 20 dicembre 2017).
4. Benkler, Y.: The Penguin and the Leviathan: The Triumph of Cooperation over Self-Interest. Random House, New York, NY (2011)
5. Bowman, J., Gufflet, M.: Meeting the Challenge of a Global GDPR and BCR Programme. European Data Protection Law Review **3**(2) (2017)
6. Boyle, J.: Shamans, Software and Spleens: Law and the Construction of the Information Society. Harvard University Press, Cambridge MA (USA) (1996)

7. Bratteteig, T., Wagner, I.: Disentangling Participation: Power and Decision-making in Participatory Design. Springer, Dordrecht-London (2014)
8. Busby, H., Hervey, T., Mohr, A.: Ethical EU law?: The influence of the European Group on Ethics in Science and New Technologies. European Law Review **33** (2008) 803842
9. Buttarelli, G.: Ethics at the Root of Privacy and as the Future of Data Protection. Address given at event hosted by Berkman Center for Internet and Society at Harvard University and the MIT Internet Policy Initiative and the MIT Media Lab, 19 April (2016) (accessed 23 December 2017).
10. Cantoni, F. and Mangia, G. (eds): Human Resource Management and Digitalization. Routledge, London-New York (2018)
11. CAST (Council of Advisors on Science and Technology of President): Big Data and Privacy: a Technological Perspective (May 2014) (accessed 4 January 2018).
12. CAST: Big Data: seizing opportunities, preserving values (May 2014) (accessed 4 January 2018).
13. CAST: Big Data: Seizing Opportunities and Preserving Values: Interim Progress Report (February 2015) (accessed 4 January 2018).
14. CAST: Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights (May 2016) (accessed 4 January 2018).
15. Council of Economic Advidersr: Big Data and Differential Pricing (February 2015) (accessed 4 January 2018).
16. DHEW (U.S. Department of Health, Education, and Welfare):: Records, Computers, and the Rights of Citizens. Report of the Secretary's Advisory Committee on Automated Personal Data Systems. (July 1973) (accessed 22 December 2017).
17. Dwork, C. In: Differential Privacy: A Survey of Results. *In* Theory and Applications of Models of Computation. Proceedings of the 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Springer, Berlin, Heidelberg (2008) 1–19
18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. arXiv: **1104.3913 [cs.CC]** (2011) (accessed 3 January 2017).
19. Dwork, C., Mulligan, D.K.: It's not privacy and it's not fair. Stanford Law Review Online **66** (September 2013) 35–40
20. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(34) (2014) 211–407
21. EDPS (European Data Protection Supervisor): Meeting the challenges of big data. A call for transparency, user control, data protection by design and accountability. Opinion 7/2015 (November 2015) (accessed 4 January 2018).
22. EDPS: European Data Protection Supervisor Decision fo 3 December 2015 establishing an external advisory group on the ethical dimensions of data protection ('the Ethics Advisory Group'). Brussles (December 2015) (accessed 4 January 2018).
23. EDPS: Debating ethics: Dignity and respect in data driven life. In: 40th International Conference of Data Protection and Privacy Commissioners. (December 2017) (accessed 2 January 2018).
24. Edwards, L., Veale, M.: Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. Duke Law & Technology Review **16** (2017) 18–84
25. EGE (European Group on Ethics in Science and New Technologies): Opinion 26. Ethics of Information and Communication Technologies (2012)
26. Fairfield, J., Engel, C.: Privacy as a Public Good. Duke Law Journal **65**(3) (2015) 385–457

27. Friedler, S., Scheidegger, C., S., V.: On the (im)possibility of fairness. **arXiv:1609.07236v1 [cs.CY]** (2016)
28. FTC (Federal Trade Commission): Fair Information Practice. The United States Federal Trade Commission's Fair Information Practice Principles (FIPPs) (1973)
29. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying Personal Genomes by Surname Inference. Science **339**(6117) (January 2013) 321–324
30. Hildebrandt, M.: Legal and Technological Normativity: More (and less) than Twin Sisters. TECHNE **12**(3) (2008) 169–183
31. Hirsch, D.: Privacy, Public Goods, and the Tragedy of the Trust Commons: A Response to Professors Fairfield and Engel. Duke Law Journal Online **65** (February 2016) 67–93
32. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLOS Genetics **4**(8) (08 2008) 1–9
33. Jasanoff, S.: Science and Public Reason. Routledge, New York (2012)
34. Joseph, M., Kearns, M., Morgenstern, J., Roth, A.: The authority of "fair" in machine learning. **Research Gate** (2016) Presented as a talk at the 2016 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2016).
35. Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences **111**(24) (2014) 8788–8790
36. Lessig, L.: Code – Version 2.0. Basic Books, New York, USA (2006)
37. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5) (2016) 14–19
38. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. SP '08, Washington, DC, USA, IEEE Computer Society (2008) 111–125
39. OECD (Organization for Economic Cooperation and Development): Guidelines on the Protection of Privacy (1980, updated 2013) (accessed 24 December 2017).
40. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY, USA (2016)
41. Pariser, E.: The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Group, New York, NY, USA (2012)
42. Pasquale, F.: The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge, MA, USA (2015)
43. PCSBI (Presidential Commission for the Study of Bioethical Issues): Privacy and Progress in Whole Genome Sequencing, Washington, D.C. (October 2012)
44. Rawls, J.: A theory of justice. Harvard University Press, Cambridge, MA (1971)
45. Rubinstein, I.: Big Data: The End of Privacy or a New Beginning? International Data Privacy Law **1** (2013) (accessed 24 December 2017).
46. Schwartz, P.: Property, Privacy, and Personal Data. Harvard Law Review **117**(7) (2004) 2056–2128
47. Searls, D.: The Intention Economy: When Customers Take Charge. Harvard Business Review Press, Cambridge, MA (2012)
48. Sen, A.: Equality of what? The Tanner Lecture on Human Values. Delivered at Stanford University May 22 (1979)
49. Skirpan, M., Gorelick, M.: The authority of "fair" in machine learning. **arXiv:1706.09976v2 [cs.CY]** (2017) Presented as a talk at the 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017).

50. Tallacchini, M.: From Biobanks to Genetic Digital Networks: Why official pre-identified values may not work. Routledge, London-New York (2015) In: Guimaraes Pereira, A. and Funtowicz, S., "Science, Philosophy and Sustainability. The End of the Cartesian dream", pp.98-111.
51. Tallacchini, M.: To Bind or Not Bind? European Ethics as Soft Law. Routledge, London-New York (2015) In: Hilgartner, S. and Miller, C. and Hagendijk, R., "Science and Democracy. Making Knowledge and Making Power in the Biosciences and Beyond", pp.156-175.
52. Warren, S.D., Brandeis, L.D.: The right to privacy. Harvard Law Review **4**(5) (1890) 193–220
53. Winner, L.: Do Artifacts Have Politics? Daedalus **109**(1) (1980) 121–136
54. Yakowitzas, J.: Tragedy of the Data Commons. Harvard Journal of Law & Technology **25**(1) (2011) 1–67
55. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., Weller, A.: From Parity to Preference-based Notions of Fairness in Classification. **arXiv:1707.00010 [stat.ML]** (2017) To appear in Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017).