Ethics in smart information systems

Chapter · July 2022

DOI: 10.2307/j.ctv2tbwqd5.14

3 authors, including:

Francesca Pratesi
Italian National Research Council
27 PUBLICATIONS **442** CITATIONS
SEE PROFILE

Fosca Giannotti
Italian National Research Council
330 PUBLICATIONS **15,993** CITATIONS
SEE PROFILE

9

Ethics in smart information systems

Francesca Pratesi, Roberto Trasarti and Fosca Giannotti

Introduction

Big data analytics and social mining raise a number of ethical issues: indeed, the scale and ease with which analytics can be conducted today completely change the ethical framework (Uria-Recio, 2018). We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do. Artificial intelligence (AI) is becoming a disruptive technology, and resources for innovation are currently dominated by giant tech companies.

In recent years, we have witnessed different initiatives in Europe aimed at providing environments and infrastructures to share research data and technologies, in accordance with the principles of Open Research Data and Open Science. The general idea behind these initiatives is to provide ecosystems for enhancing scientific collaborations among researchers and practitioners, even those from different disciplines. Examples of recent initiatives in different research fields are: EOSC Pilot (2018) and SoBigData (2015; Forgó et al, 2020) (social sciences), SeaDataCloud (Sea Data Net, 2006) (environmental and earth sciences), and IN-SKA (SKA Organisation, 2011) (physical sciences).

In the field of information and communications technology (ICT), a common goal is to achieve responsible research and innovation (RRI) aimed at providing a platform or ecosystem for ethics-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, or, in other words, social big data science. More and more often, these data regard private aspects of our lives, such as our movements (Inkpen et al, 2018), healthcare (Rodríguez-González et al, 2019), our social interactions and our emotions (Hasan et al, 2019). In this context, it therefore becomes fundamental to take into consideration the legal and ethical aspects of the processing of personal data, especially given the entry into force of the General Data Protection Regulation (GDPR) in May 2018, but also to move forward considering already existing recommendations and exploring the frontier of novel solutions, in accordance with shared societal and moral values. For this reason, it is important that legal requirements and

constraints are complemented by a solid understanding of ethical and legal views and values, such as privacy and data protection.

The rest of this chapter is structured as follows. First of all, we analyse the solution related to data protection, describing the general idea and providing an overview of some technical solutions. Then, we focus on the *right to explanation*, listing the most important properties an explanation should have. Again, we move towards a novel model that aims to cover more ethical aspects in the generation of AI systems. Finally, we conclude the chapter with some general remarks of relevance to making policy in this field.

Privacy and data protection

During the twenty-first century, individual privacy has been one of the most discussed jurisdictional issues in many countries. Indeed, the very fine level of detail of data collected by a variety of organisations comes along with potential issues, such as containing and controlling personal information. Consequently, the opportunities to release the knowledge hidden in data bring an increased risk of privacy violation of the people who are represented in it. The threat includes identification of personal aspects of people's lives, such as their home address, mobility habits and religious or political beliefs. Managing this kind of data is not a trivial task. It is not sufficient to rely only on de-identification (that is, removing the direct identifiers contained in the data) in order to preserve the privacy of the people involved. In fact, many examples of reidentification from supposedly anonymous data have been reported both in the scientific literature and in the media, from GPS trajectories (de Montjoye et al, 2013; Hern, 2014) to movie ratings of on-demand services (Narayanan and Shmatikov, 2008) and, even, from health records (Sweeney, 2002).

Several techniques and technological frameworks have been proposed to counter privacy violations, without losing the benefits of big data analytics technology (Fung et al, 2010a). Unfortunately, no general method exists that is capable of handling both generic personal data and preserving generic analytical results. Nevertheless, big data and privacy are not necessarily opposites: indeed, many practical and impactful services can be designed in such a way that the quality of results can coexist with a high protection of personal data if the Privacy-by-Design (PbD) paradigm is applied (Monreale et al, 2014). The PbD paradigm (Cavoukian, 2009, 2012; Cavoukian and Jonas, 2012), introduced by Cavoukian in the 1990s, aims to protect privacy by inscribing it into the design specifications of information technologies, accountable business practices, and networked infrastructures, from the very start. It represents a profound innovation with respect to traditional methods; the idea is to have a significant shift from a reactive model to a proactive one, that is, preventing privacy issues arising in the first place

instead of remedying them. PbD has raised interest especially in the last few years because an elaboration of this paradigm is explicitly referred to in the new European GDPR (European Parliament and Council, 2016: 118). Indeed, the new regulation states that controllers shall implement appropriate technical measures for ensuring, by default, the protection of personal data.

The problem of protecting individual privacy when disclosing information is not trivial and this makes the problem scientifically attractive. It has been studied extensively also in the data mining community, under the general umbrella of privacy-preserving data mining and data publishing (Monreale, 2011; Pratesi, 2017; Pellungrini, 2020). The aim of the methods proposed in the literature is of assuring the privacy protection of individuals during both the analysis of human data and the publishing of data and extracted knowledge. Two main families of approaches treat the problem of privacy preservation: anonymity by randomisation and anonymity by indistinguishability. More recently, anonymity by encryption has also become popular.

Anonymity by randomisation

Randomisation methods are used to transform data in order to preserve the privacy of sensitive information, perturbing the data using a noise quantity. They were traditionally used for statistical disclosure control (Adam and Wortmann, 1989) and later have been extended to privacy-preserving data mining problems (Agrawal and Srikant, 2000). In the literature, there exist two types of random perturbation techniques: additive random perturbation and multiplicative random perturbation. In the additive random perturbation methodology, the perturbed dataset is obtained drawing independently from the probability distribution (Uniform or Gaussian) some noise quantities and adding them to each record in the original data set. Thus, individual records are not available, while it is possible to obtain distribution describing the behaviour of the original data set. Moreover, from the perturbed data, it is still possible to extract patterns and models, even if there was the need to develop new data mining approaches to work with aggregate distributions of the data in order to obtain mining results (Agrawal and Srikant, 2000; Agrawal and Aggarwal, 2001; Evfimievski et al, 2002; Rizvi and Haritsa, 2002; Zhan et al, 2005; Zhang et al, 2005). For privacy-preserving data mining, multiplicative random perturbation techniques can also be used. The main techniques of multiplicative perturbation are based on the work presented in Johnson and Lindenstrauss (1984).

Unfortunately, the main problem of randomisation methods is that they are not safe in case of attacks with prior knowledge (Kargupta et al, 2003). To overcome this drawback, a relatively new randomisation paradigm was developed: a recent model of randomisation, though based on different

assumptions, is differential privacy. This is a privacy notion introduced by Dwork (Dwork et al, 2006). The key idea is that the privacy risks should not increase for a respondent as a result of occurring in a statistical database. Differential privacy ensures, in fact, that the ability of an adversary to inflict harm should be essentially the same, independently of whether any individual opts in to, or opts out of, the data set. This model is called ε-differential privacy, due to the level of privacy guaranteed ε. It assures a record owner that any privacy breach will not be a result of participating in the database since nothing, or almost nothing, that can be discovered from the database with his record that could not have been discovered from the one without his data (Fung et al, 2010b). Moreover, Dwork (2006) formally proved that ε-differential privacy can provide a guarantee against adversaries with arbitrary background knowledge.

Anonymity by indistinguishability

As already stated, randomisation methods have weaknesses. In some cases, it is better to apply methods that reduce the probability of record identification by public information and that are not data-independent: k-anonymity, l-diversity and t-closeness. The traditional k-anonymity framework (Sweeney, 2000) focuses on relational tables. The basic assumption is that attributes are partitioned in quasi-identifiers and sensitive attributes (Sweeney, 2002). The first kind of attributes can be linked to external information to reidentify the individual to whom the information refers (so-called linking attack); they are available in public such as age, postcode and sex. The second category of attributes instead represents the information to be protected. A data set satisfies the property of k-anonymity if each released record has at least $(k - 1)$ other records also visible in the release whose values are indistinct over the quasi-identifiers. The k-anonymity model usually relies on methods such as generalisation and suppression to reduce the granularity of representation of quasi-identifiers. It is evident that these methods guarantee privacy but also reduce the accuracy of applications on the transformed data. The main problem of k-anonymity is to find the minimum level of generalisation that allows us to guarantee high privacy and good data precision. Indeed, Meyerson and Williams showed that the problem of optimal k-anonymisation is extremely complex to solve (Meyerson and Williams, 2004). Fortunately, many efforts have been done in this field and many heuristic approaches have been designed (see Bayardo and Agrawal, 2005; and LeFevre et al, 2005).

Unfortunately, the k-anonymity framework, in some cases, can be vulnerable (Kifer, 2009). In particular, it is not safe against homogeneity attack and background knowledge attack. The homogeneity attack easily infers the value of the sensitive attributes when a k-anonymous data set contains a group of k entries with the same value for the sensitive attributes.

In a background knowledge attack, instead, an attacker knows information useful to associate some quasi-identifiers with some sensitive attributes. So, he can reduce the number of possible values of the sensitive attributes. Against these two kinds of attack, l-diversity was proposed (Machanavajjhala et al, 2006). The basic idea is to maintain the diversity of sensitive attributes. However, in some cases, the attacker can infer the value of the sensitive attribute knowing the global distribution of the attributes. The t-closeness method (Li et al, 2007) is safe against this kind of attack. It requires that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the overall table has to be bounded by a threshold t, ensuring the two distributions (the original and the sanitised ones) are quite similar.

Anonymity by encryption and cryptography

Many studies have addressed the problem of supporting query execution on encrypted data. One of the most relevant is homomorphic encryption (Gentry, 2009), which supports computations without decrypting the input. This kind of encryption enables the computation of some operations (such as additions, multiplications and quadratic functions) on encrypted data and generates encrypted results, which, conveniently decrypted, correspond to the results of the same operations performed on the plain text. The weak point of this technique is in the efficiency in the query processing. Other methods append indexes (a sort of metadata) to the data and are useful for executing specific queries (Hacigümüş et al, 2002; Ceselli et al, 2005; De Capitani Di Vimercati et al, 2007). In particular, Hacigümüş et al (2002) explain how it is possible to split a query (translating specific query operations) into a server query and a client query. The first query can be executed without having to decrypt the data, while decryption and compensation query are performed at the client site. Ceselli et al (2005) focus on inference exposure, providing a model to evaluate the trade-off between performance degradation and protection ensured. Finally, De Capitani and Di Vimercati et al (2007) concentrate on data outsourcing and present a solution to the enforcement of access control and the management of its evolution.

A possible use of homomorphic encryption, that can be found in Damgård et al (2012), is in Secure Multi-party Computation (SMC) (Yao, 1982; Goldwasser, 1997), which deals with computing a certain function on multiple inputs in a distributed network. The problem in this case is to compute any probabilistic function on inputs that are distributed among the participants in the network, while ensuring independence of the inputs, correctness of the computation, and that no more information is revealed to participants in the computation. The computation can be carried out by a

single participant or by a coalition of participants. As noted in Goldwasser (1997), a trivial centralised solution would be to assume a trusted centre exists, and that all users send their inputs to this trusted centre for the computation of their respective outputs. A preferable option is a solution where trust is distributed. SMC is often used in distributed environments, but regrettably it allows only some kinds of computations.

One of the first techniques is shown in Chaum et al (1988), where participants can share secrets, even if one third of the participants deviate from the protocol (that is based on not leaking secret information and on sending the correct messages). A more recent solution can be found in Gilburd et al (2004), where a new privacy model, k-privacy, is proposed for real-world large-scale distributed systems. They use a relaxed privacy model implementing efficient cryptographically secure primitives that do not require all-to-all communications. Another example is the work of Sanil et al (2004), where they implement a privacy-preserving algorithm of computing regression coefficients, which permits (honest or semi-honest) agencies to obtain the global regression equation as well as to perform rudimentary goodness-of-fit diagnostics without revealing their data.

The right to explanation

The GDPR, in its Recital 71, also mentions the right to explanation, as a suitable safeguard to ensure a fair and transparent processing in respect of data subjects. While privacy and data protection are not novel concepts, and a lot of scientific literature has been explored on these topics, the study of explainability is a new challenge.

So far, the usage of black boxes in AI and machine learning processes implied the possibility of inadvertently making wrong decisions due to a systematic bias in training data collection. Several practical examples have been provided, highlighting the 'bias in, bias out' concept. One of the most famous examples of this concept regards a classification task: the algorithm goal was to distinguish between photos of wolves and Eskimo dogs (huskies) (Ribeiro et al, 2016). Here, the training phase of the process was done with 20 images, hand-selected such that all pictures of wolves had snow in the background, while pictures of huskies did not. This choice was intentional because it was part of a social experiment. In any case, on a collection of additional 60 images, the classifier predicts 'wolf' if there is snow (or light background at the bottom), and 'husky' otherwise, regardless of animal colour, position, pose and so on.

However, one of the most worrisome cases was discovered and published by ProPublica, an independent, non-profit newsroom that produces investigative journalism with moral force. In Angwin et al (2016) and Larson et al (2016), the authors showed how software can actually be racist. In a nutshell, the

authors analysed a tool called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions). COMPAS tries to predict, among other indexes, the recidivism of defendants, who are ranked low, medium or high risk. It was used in many US states (such as New York and Wisconsin), to suggest to judges an appropriate probation or treatment plan for individuals being sentenced. Indeed, the tool was quite accurate (around 70 per cent overall with 16,000 probationers), but ProPublica journalists found that Black defendants were far more likely than White defendants to be incorrectly judged to be at a higher risk of recidivism, while White defendants were more likely than Black defendants to be incorrectly flagged as low risk.

From these examples, it appears evident that explanation technologies can help companies for creating safer, more trustable products, and better managing any possible liability they may have.

The five dimensions of a valid explanation

So far, we analysed the motivation to provide an explanation, which can be both legal and utilitarian. However, from a practical point of view, we need to define some dimensions, useful to understand what makes for a valid explanation. The EU-funded PRO-RES project (aiming to PROmote ethics and integrity in non-medical RESearch) that produced this book hosted a workshop about ethics, social mining and explainable artificial intelligence (ESME 2019). The discussion in this section reports on the group thinking that took place during that event.

First of all, experts tried to define *what is an explanation*, analysing the main characteristics that a good explanation should have:

- *Simplicity*. This is one of the most important properties: the simplest explanation, which requires a minimum cognitive effort to be understood, should be enough. You must be able to reason on the black box model if you are going to understand and to keep all the concepts in your mind.
- *Truth*. This seems trivial, but it must be considered by design: if an explanation is not true, probably there are some biases in the data. However, if you are visualising an advertisement for a wrong reason, it is the classification process that is wrong, while the explanation of why you are visualising the advertisement is still correct.
- *Symbolic*. The explanation should be as general and abstract as possible, and it should possibly imitate human intelligence in the performed reasoning. An example is saliency maps, which are usually a good way to compare algorithms, but they are not symbolic since they only highlight areas or pixels involved in the classification process, without providing any additional information about what that area really represents.

- *High level*, in order to have understandable explanations. This is strictly related to the symbolic property since a more abstract explanation is also generally a higher-level one. Both these characteristics are important because the explanation becomes also simpler to be understood by anyone (see also *Simplicity*, the first bullet point). Consider, as an example, an explanation that is very complex because the explainability model that generated it learned to use too detailed or irrelevant information, generating explanations that are 'overfitted'; in this case, the model can be unable to work properly with new data.
- *Local vs global?* There are different levels to have an explanation: when the explanation is local it is explaining only a single case, while the goal of a global explanation is to recap the overall logic behind a black-box model. The first case is easier for very complex models like neural networks, but to better understand the big picture we probably need something in between, a sort of *sub-global explanation*.
- *Given by causality*, not by correlation, or, even better, by counterfactual analysis or domain adaptation.
- Providing *reasoning and learning* at the same time, taking advantage of multiple data sources (for example, classifying images using both pictures and captions).
- *Actionable*. Indeed, human perception has an element of intuition which is not explainable, or it is very hard to model.
- *Trust*. We must rely on an explanation. Indeed, as highlighted also by Kersting (2020), people are not disposed to forgive a wrong explanation.
- *Stability*. Similar instances should have similar explanation for a given model. A non-deterministic explanation could be easier to provide, but it implies that understanding the model is more difficult; in addition, it violates the property of simplicity and, probably, the trust. Indeed, consistency is a fundamental property also in real life: if a person asks three different doctors for a medical opinion, the opinions must be similar in order for that person to trust them.

Second, the discussion moved on to *how to measure the understanding of an explanation*.

Generally speaking, *we can measure the level of comprehensibility of an explanation as the degree in which humans can replicate the reasoning of the machine*. The measure must be: consistent, trusted, accountable, stable but also monotonic. Indeed, as humans, we accept better explanations which follow a logic: if the measure first grows and then decreases, you will not accept the explanation. The more an explanation is following a certain *monotonicity*, the better it is. The generality of the measure, instead, is not crucial because it depends on the final user of the explanation and on the situation that we are analysing. Regarding the stability, it is preferable that the explanation

does not vary too much if parameters vary. Indeed, with adaptable decision algorithms, a small variation in the inputs may change the decision too.

Third, the discussion focused on what are *interpretable data and interpretable models*. It is easy to agree that there is a need to find a trade-off between accuracy and simplicity: models are often so complex that we need to approximate the flows; however, we need to simplify models being careful not to make them too generic. The gender factor could offer a good explanation: if the majority of women in their 30s visualise an advertisement, explaining to them that they are receiving that advertisement because they are women between 30 and 40 years old, and usually women of that age appreciate the offered product or service, seems to be a good explanation. Of course, data (and how it is integrated into the model) also has its importance; for example, we cannot transform a non-interpretable variable into an interpretable one. Clearly, using such explanations is a generalisation, and it could not capture all the characteristics involved (for example, it could be that age is correlated with other variables).

Then, another crucial point regards the *business perspective* of explanation, that is, the implications for companies that have to guarantee explainable and interpretable systems and models, and whether and how these systems can be actually realised in real applications. A big problem is that companies are forced to provide explanations, but they do not want to reveal how the system is reasoning, in order to preserve business strategies and secrecy. Auditors can solve the problem of checking fairness without compromising trade secrets, *but* controllers may feel that providing too detailed an explanation is against their trade interest: providing a lot of detailed explanations to different individuals may disclose the model. A possibility is that an explanation is personal: if a user requests an explanation, it could be based only on their data (even if this is partially in contrast with the principles of stability, and to being sub-global) and it must be revealed just to the user and cannot be shared. One (not very feasible) alternative is to drop not explainable models and only use intrinsically explainable algorithms, but the possibility to use something that can explain an algorithm is substantial.

However, it seems reasonable that users are interested in knowing an explanation of their own situations, while they are not interested in a super-detailed explanation, so the intellectual property of companies seems not to be at risk. In addition, there is the problem that an explanation could require information that is not directly available on site (for example, the economic system is very complex and if you want to explain the price of some products you need to analyse the whole market); we cannot explain every single decentralised node in the big network, but only treat them as a unique giant black box. It is also very important to clarify for which categories of systems we need to provide an explanation. Social network advertising? Or just for loans, mortgages and health-related systems?

Finally, it is important to consider that the final decision belongs to users, and it must be taken by real persons: automatic decision systems should only support decisions. Thus, another crucial question is who are the *final users*, and how the explanation must adapt to its target. An explanation should surely be human-understandable when decision impacts on legal status; however, humans do not need to understand the full model, just why a certain decision was made on them. *Unfortunately*, very often, different types of people require different explanations (for example, diagnosis explained to a patient or to a doctor).

Indeed, the explanation should change along with the background. If we consider developers as the final user, an explanation allowing a prediction of a system's answer is useful enough. An explanation can be useful for debugging (for example, to find bias in the data). A more difficult objective of explanations is their potential social function, that is, a way to suggest to users how to change their behaviour in order to change a system answer and achieve their goals. Tools are needed for different categories of people and different levels of understanding. Sometimes only one factor among many can be given as an explanation. Maybe a solution could be to provide a system that offers the possibility to go gradually in deep: surely the explanation 'you are receiving this advertisement because you are a woman between 30 and 40 years old' is enough for the majority, but if a woman wants to know more, the system could add some information about her web history, 'and you visited the websites X and Y'. Of course, users still have the option (and the right) to contact the data protection authorities if a received explanation does not satisfy them. However, this path is not followed as often: as an example, it is important to point out that in the first year from the entry in force of the GDPR, the Italian Data Protection Authority received zero requests for an explanation. We need to investigate whether explainability is a right that does not interest people or if the general public is simply not aware of this right.

To conclude, a comment at the ESME 2019 workshop mentioned earlier made by Dino Pedreschi summarises this discussion well: 'Explainability is not a value, it is a tool, and we need to understand how to use it.'

Towards ethics by design and Trustworthy AI

With legal frameworks evolution, ethical concerns and guidelines are changing too. As highlighted in the World Economic Forum (2016), this is reflected by social networks continuing to update privacy policies and settings, by newsrooms making frequent updates to publishing guidelines on how they use material sourced from social media platforms, and by the continuous shifts in what is or is not considered appropriate when individuals post on social media platforms. Moreover, both active and passive data

collections also raise questions. In this context, the World Economic Forum warns both people and organisations, pointing out that people need to be informed about the potential impact of their content being shared widely. On the other hand, organisations must be honest with the user about when and how the content will be used, and whether it will be syndicated to other publishers or organisations.

The World Economic Forum is not the only entity that invokes transparency. Indeed, transparency is one of the pillars in ethics and it is related to several parts of the big data process, such as seeking permission of users, explanation of terms of use, and data usage after the collection. The Organisation for Economic Co-operation and Development (OECD, 2013), UK Cabinet Office (Government Digital Service Cabinet Office, 2016) and Council of Europe (Directorate General of Human Rights and Rule of Law, 2017) state that notice and consent are fundamental tasks in big data management. They also offer other important considerations about ethics. In particular, De Mooy (2017) gives a good excursus on the history of individual control, on cultural differences between Europeans and Americans and a list of key concepts useful for addressing the challenges of privacy management.

These guidelines are: individual empowerment (through education that teaches individual basic technology and data portability), corporate accountability (through a voluntary, self-regulatory risk assessment) and collective accountability (through government-mandated entities that can assess the impact of any big data process). The OECD framework (2013) is presented along with fundamental principles that should be respected in the data usage process: collection limitation (data collected are the minimum necessary and they must be obtained by lawful and fair means), data quality (personal data should be relevant to the purposes for which they are to be used and they must be complete and up to date), purpose specification (purposes should be specified before any data collection), use limitation (data must be used and disclosed only for the specified purpose), security safeguard (data must be protected by reasonable security safeguards), openness (about development, practices and policies), individual participation (individuals should have the right to control, rectify or have their data erased) and accountability (data controllers should be accountable for complying with measures regarding the other principles). In the UK (Government Digital Service Cabinet Office, 2016), we can find a short summarisation, along with some practical examples of good and bad practices, of the six key principles they consider essential to data management: (1) to highlight the users' need and public benefit from the start of the definition of the methods; (2) to use data and tools with the minimum intrusion necessary; (3) to create robust data science models, analysing the representativeness of the data and the presence of potential discrimination features; (4) to be aware of public

perception, understanding how people expect their data to be used; (5) to be clear and open about data, tools and algorithms, providing explanation in plain English; and (6) to keep data secure, following the guidelines provided by the Information Commissioner's Office (ICO, 2017). Finally, the Council of Europe (Directorate General of Human Rights and Rule of Law, 2017) drafted guidelines too. The majority of ethical principles are highly shared among different institutions, and many of them are included in the new EU Regulation. However, in the case of relatively loose regulatory environments, ethical rules are particularly important. Zook et al (2017) listed ten rules for performing ethical research on big data. Some of them are inspired by the concepts already described (for example, inserting ethics directly in the workflow of research or documenting clearly when decisions are made), while others are specifically oriented to research. For example, the importance of debating issues within a group of peers or of sharing data is listed as a fundamental task in some projects, like studies of rare genetic diseases. Last but not least, the 'Ethics guidelines for trustworthy AI' (European Commission, 2019) are a valuable help to researchers. According to the guidelines, Trustworthy AI should be:

1. lawful, respecting all applicable laws and regulations;
2. ethical, respecting ethical principles and values;
3. robust, both from a technical perspective, while taking into account its social environment, since, even with good intentions, AI systems can cause unintentional harm.

One of the most innovative parts of the document is the acknowledgement of potential tensions and the promotion of trade-offs between some ethical imperatives, such as: respect for human autonomy (ensuring respect for the freedom and autonomy of human beings); prevention of harm (and guaranteeing protection of human dignity as well as mental and physical integrity); fairness (both regarding a substantive and a procedural dimension, that is, ensuring equal and just distribution of both benefits and costs and that individuals and groups are free from unfair bias, discrimination and stigmatisation, while seeking effective redress against decisions made by AI systems and by the humans operating them); and explicability (again, processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected). Moreover, some of the authors (Quintarelli, 2020) stressed some other interesting points: (1) the process automation affects the extent and the speed in the data domain, thus, traditional methods are not working anymore; as a consequence, we need to rethink our rules and the way to assign and verify responsibilities; (2) automated systems at some point could perform wrong predictions

or actions, so we need to provide for remedies for such errors, enabling a redress-by-design paradigm; (3) it is not fair to evaluate each single instance of a problem, that is, the personal cases, but we need to evaluate also the benefits of a solution for the whole society.

How to reconcile ethical and industrial objectives

In this subsection, we want to provide an overview of some practical solutions that could help in promoting ethics, even in a business domain. Unless otherwise specified, the discussion reported is again a result of the event ESME 2019, the PRO-RES workshop about ethics, social mining and explainable artificial intelligence.

We start with an example of the application of ethical concerns in everyday life, with particular connection to private companies that affect society, and then we try to figure out some possible general solutions. Applying ethical concerns in everyday life also means that we confront each other also with Web 2.0 and online social networks (OSNs). These relatively new concepts implied clear advantages (reducing distances and democratising the information) but also novel issues. Indeed, several other contemporary problems (fake news is the first among them) are due to the possibility of remaining anonymous. In the real world, actions have different consequences with respect to digital life, and real life often has some form of self-regulatory system (as an example, if I live in a small village and I lie often, I soon become seen as unreliable for other persons). Thus, to lever *individual responsibility* for each action and opinion in the digital world setting too seems to be the right way to proceed. Of course, in a digital work this responsibility level is harder to achieve with regard to the real world, nevertheless, from a technical point of view, mechanisms that regulate this aspect are nearly possible: for example, a certified digital identity can be provided to every online user.

This is particularly important because an OSN is not necessarily a 'bad guy' who tries to break the rules, and it seems unfair that it should be the only entity in charge of supervising the users' contents. In addition, it also seems unfair that the online platform has the responsibility to establish if a certain content is illegal (and several discussions came after, for example, the 2021 US Capitol riot[1]), but an independent authority could help, participating in these disputes; thus, outsourcing of legal decisions might be a solution. Nevertheless, the fact that owners of OSNs are private companies that can and should autonomously decide whether to publish content or not does not remove the responsibility of the role of the company itself in society. Recently, some steps have been made by private companies to enforce control and integrity of published content (Halevy et al, 2020). Of course, a clear drawback in removing anonymity is a possible limitation of freedom of speech and to be too over-blocking, so another aspect that must be also

considered is the trade-off between accountability and the freedoms of expression and information (in some countries, anonymity is fundamental to protect users). However, this dilemma is not new since law usually must balance between opposing rights.

The first problem that we need to face is that laws and ethics do have a certain cultural dependency. For example, in the US, nudity is considered a very serious problem, while in India, hate against castes is a sensitive issue. Again, in Italy, one of the major problems is cyberbullying, while Germany has a law against hate speech and fake news. Moreover, the same problem could have different severity: nudity is a concept that can be different in different countries and even in different locations in any one country! Thus, the *need for global ethical values* contrasts with the fact that each company uses data in different ways and operates in different countries, so common ground could not easily be fairly or equitably established.

Nevertheless, given the international reach of much big data *we need an ethical framework of fairly common standards and values*, where legislation is only the basis. Indeed, the GDPR does not cover all the aspects related to data protection; thus, being GDPR-compliant is only the first step. Other aspects that should be considered are:

- *user-centric model*: we need to work for the individuals;
- *substitute privacy with ethics*: indeed, privacy is only one aspect; transparency is one another pillar of ethics, as we have already discussed;
- provide *examples of business models that are ethics-aware*.

A possibility is represented by the positive-sum model (opposite to the zero-sum one), given by the ethics-by-design paradigm. Ann Cavoukian presents this model (Cavoukian, 2018), an extension of the famous Privacy-by-Design one (thanks to which we became finally aware that we can obtain both privacy and utility in machine learning) that also includes transparency, accountability, algorithmic responsibility and security. Dr Cavoukian, who created the Global Privacy and Security by Design organisation,[2] remarks that investing in prevention is more cost-effective in the long term, and she pointed out the importance of evaluating both algorithm and data in the explainability questions. Companies should understand that ethics is an added value, and, in the long run, this is convenient for companies too. An ethics-aware business leads to more trust from customers; this implies that more users will use the company's product/service (the reputation of a company plays a significant role in the acquisition and retention of clients); more users mean more data and, thus, more money for the company. Indeed, all the participants agree that access data is one of the primary goals of each company, and ethics-by-design can help to gain access to data and to manage it at best.

An *alternative model towards a less profit-centred concept of values* is possible, and the participants identify some necessary ingredients:

- *awareness* of people, both from users and people working in companies;
- encourage *interdisciplinarity*;
- *public incentives*, to overcome the general lack of interest from companies;
- *sustainability*: ethical environment is, in all respects, an environment that we need to protect;
- *ethics-by-design*: ethics is an added value – it is a resource and not merely a cost.

Clearly, some costs are still necessary: to create an interdisciplinary team (for example, legal experts for compliance with the law, social media expert for improving the communication of values, ethical philosophers for analysing the whole aspects, computer scientists for implementing solutions), to implement technological tools that help in explaining the behaviour of a black box and tools for ensuring privacy.

The interest from governments in enforcing ethics is crucial: we cannot wait for private companies creating alternatives, for example, to some overused tools. It is obviously inconceivable and utopian that a company develops, for example, a new ethics search engine able to actually replace Google. As academic institutions, instead, we can contribute to this, if there is a clear direction and effort from the EU government. Indeed, a common European ethical framework might also affect other countries and provide a model that can be adopted worldwide (as the GDPR already did).

As academics, we have the *moral obligation to push toward the creation of new models, and we can contribute by providing practical ideas and solutions* (bringing evidence that they can work), so companies could invest in them.

Conclusion

In this chapter, we described progress and open challenges related to ethics in AI systems and machine learning processes. In particular, we gave an overview of what is mainly done: anonymisation (encryption or removal of personally identifiable information), access control (selective restriction of access to places or resources) and policy enforcement (of rules for the use and handling of resources). We also outline the problems of accountability (the evaluation of compliance with policies and provision of evidence) and data provenance (attesting to the origin and authenticity of information). Then, we talked about transparency (explanation of information collection and processing) and explainability (of algorithms, that is, its reasoning or, at least, a justification of a given decision). In particular, we analysed the main characteristics that a good explanation should have, how to

measure an explanation, and the business perspective of explanation. We argued that we surely need to provide alternative technologies, but, more importantly, we need to find *alternative business models*, that can be applied by private companies. In such a way, we can finally build an economy that 'works for people' and, as advocated by Ursula von der Leyen, this can permit the move from 'need to know' to 'need to share' (von der Leyen, 2019). Finally, we move the discussion to a more complete ethical approach to AI, considering privacy and data governance in the equation, but also human agency and oversight, robustness and safety, transparency and accountability, non-discrimination and fairness, and societal and environmental well-being.

Indeed, AI is a collection of technologies that combine data, algorithms and computing power. On that basis, as stated also by the European Commission (2020), an AI ecosystem can bring the benefits of the technology to the whole of European society and economy:

- *for citizens* to reap new benefits, for example improved healthcare, fewer breakdowns of household machinery, safer and cleaner transport systems, better public services;
- *for business development*, for example a new generation of products and services in areas where Europe is particularly strong; and
- *for services of public interest*, by improving the sustainability of products and by equipping law enforcement authorities with appropriate tools to ensure the security of citizens, with proper safeguards to respect their rights and freedoms.

Given the major impact that AI can have on our society and the need to build trust, it is vital that European AI is grounded in our values and fundamental rights such as human dignity and privacy protection. Furthermore, the impact of AI systems should be considered not only from an individual perspective, but also from the perspective of society as a whole. As policymakers grapple with these new technologies and applications, careful attention needs to be given to their ethical implications.

Acknowledgements

This work was supported by the European Commission through the Horizon 2020 research and innovation project 'PROmoting integrity in the use of RESearch results (PRO-RES)' under grant agreement No 788352, by the H2020-INFRAIA-2018–2020 / H2020-INFRAIA-2019–1 European project 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics' under grant agreement No 871042, and by the H2020 ICT-48 European project 'TAILOR: Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization' under grant

agreement No 952215. The funders had no role in developing the research and writing the manuscript.

The authors need to thank all the people participating in ESME 2019: The PRO-RES workshop about Ethics, Social Mining and Explainable Artificial Intelligence. The complete list of people, who provided fundamental inputs to the discussions reported in this manuscript, can be found at https://kdd.isti.cnr.it/esme2019/

Notes

[1] See: https://www.washingtonpost.com/technology/2021/01/09/trump-twitter-banned-apps/ and https://eu.usatoday.com/story/tech/2021/01/08/twitter-permanently-bans-president-trump/6603578002/ and https://www.ctpost.com/news/slideshow/Q-A-How-can-Twitter-ban-Trump-215406.php

[2] See: https://gpsbydesign.org/

References

Adam, N.R. and Wortmann, J.C. (1989) 'Security-control methods for statistical databases: a comparative study', *ACM Computing Surveys*, 21(4): 515–56.

Agrawal, D. and Aggarwal, C.C. (2001) 'On the design and quantification of privacy preserving data mining algorithms', in P. Buneman (ed), *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 21–23 May, 2001, Santa Barbara, CA*, New York: Association for Computing Machinery, pp 247–55.

Agrawal, R. and Srikant, R. (2000) 'Privacy-preserving data mining', in W. Chen, J.F. Naughton and P.A. Bernstein (eds) *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 15–18, 2000, Dallas TX*, New York: Association for Computing Machinery, pp 439–50.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) 'Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks', ProPublica, 23 May, available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [accessed 23 September 2020].

Bayardo, R.J. and Agrawal, R. (2005) 'Data privacy through optimal k-anonymization', in K. Aberer, M.J. Franklin and S. Nishio *(eds) Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5–8 April 2005, Tokyo, Japan*, Washington, DC: IEEE Computer Society, pp 217–28.

Cavoukian, A. (2009) 'Privacy by design: the 7 foundational principles', revised 2011, Toronto, ON: Information and Privacy Commissioner of Ontario, available from: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf [accessed 2 February 2022].

Cavoukian, A. (2012) 'Privacy by design [leading edge]', *IEEE Technology and Society Magazine*, 31(4): 18–19.

Cavoukian, A. (2018) 'AI ethics by design', Toronto, ON: Ryerson University, available from: https://www.ryerson.ca/content/dam/pbdce/papers/AI_Ethics_by_Design.docx [accessed 23 September 2020].

Cavoukian, A. and Jonas, J. (2012) 'Privacy by design in the age of big data', 8 June, Toronto, ON: Information and Privacy Commissioner of Ontario, available from: https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf [accessed 2 February 2022].

Ceselli, A., Damiani, E., De Capitani Di Vimercati, S., Jajodia, S., Paraboschi, S. and Samarati, P. (2005) 'Modeling and assessing inference exposure in encrypted databases', *ACM Transactions on Information and System Security (TISSEC)*, 8(1): 119–52.

Chaum, D., Crépeau, C. and Damgård, I. (1988) 'Multiparty unconditionally secure protocols', in *STOC '88: Proceedings of the twentieth annual ACM Symposium on Theory of Computing, Chicago, May 2–4, 1988*, New York: Association for Computing Machinery, pp 11–19.

Damgård, I., Pastro, V., Smart, N.P. and Zakarias, S. (2012) 'Multiparty computation from somewhat homomorphic encryption', in R. Safavi-Naini and R. Canetti (eds) *Advances in Cryptology: CRYPTO 2012*, Heidelberg: Springer pp 643–62.

De Capitani Di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S. and Samarati, P. (2007) 'Over-encryption: management of access control evolution on outsourced data', in *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23–27, 2007*, np [US]: VLDB Endowment, pp 123–34.

de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D. (2013) 'Unique in the crowd: the privacy bounds of human mobility', *Scientific Reports*, 3: art 1376, available from: https://doi.org/10.1038/srep01376 [accessed 2 February 2022].

De Mooy, M. (2017) 'Rethinking privacy self-management and data sovereignty in the age of big data', Gütersloh: Bertelsmann Stiftung, available from: https://cdt.org/wp-content/uploads/2017/04/Rethinking-Privacy_2017_final.pdf [accessed 2 February 2022].

Directorate General of Human Rights and Rule of Law (2017) 'Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data', 23 January, Strasbourg: Council of Europe, available from: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a [accessed 2 February 2022].

Dwork, C. (2006) 'Differential privacy', in M. Bugliesi, B. Preneel, V. Sassone and I. Wegener (eds) *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, Berlin: Springer, pp 1–12.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) 'Calibrating noise to sensitivity in private data analysis', in S. Halevi and T. Rabin (eds) *Theory of Cryptography: Proceedings of the Third Theory of Cryptography Conference, TCC 2006, New York, March 4–7, 2006*, Berlin: Springer, pp 265–84.

EOSC Pilot Consortium (2018) The European Open Science Cloud for Research Pilot Project, available from: https://eoscpilot.eu/ [accessed 2 February 2022].

European Commission (2019) 'Ethics guidelines for trustworthy AI', 8 April, available from: https://wayback.archive-it.org/12090/20201227221227/https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [accessed 2 February 2022].

European Commission (2020) 'White paper on artificial intelligence: a European approach to excellence and trust', 19 February, Brussels: European Commission, available from: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en [accessed 2 February 2022].

European Parliament and Council (2016) General Data Protection Regulation, *Official Journal of the European Union*, 4 May, L119, available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL [accessed 2 February 2022].

Evfimievski, A.V., Srikant, R., Agrawal, R. and Gehrke, J. (2002) 'Privacy preserving mining of association rules', in O.R. Zaïane (ed) KDD '02*: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23–26, 2002, Edmonton, AL*, New York: Association for Computing Machinery, pp 217–28.

Forgó, N., Hänold, S., van den Hoven, J., Krügel, T., Lishchuk, I., Mahieu, R., Monreale, A., Pedreschi, D., Pratesi, F. and van Putten, D. (2020) 'An ethico-legal framework for social data science', *International Journal of Data Science and Analytics*, 11(4): 377–90.

Fung, B.C.M., Wang, K., Chen, R. and Yu, P.S. (2010a) 'Privacy-preserving data publishing: a survey of recent developments', *ACM Computing Surveys*, 42(4): art 14, available from: https://dl.acm.org/doi/10.1145/1749603.1749605 [accessed 2 February 2022].

Fung, B.C.M., Wang, K., Fu, A.W.C. and Yu, P.S. (2010b) *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, Boca Raton, FL: CRC Press.

Gentry, C. (2009) 'Fully homomorphic encryption using ideal lattices', in M. Mitzenmacher (ed), *STOC'09: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, May 31–June 2, 2009, Bethesda, MD*, New York: Association for Computing Machinery, pp 169–78.

Gilburd, B., Schuster, A. and Wolff, R. (2004) 'k-TTP: a new privacy model for large-scale distributed environments', in W. Kim and R. Kohavi (eds) *KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle WA, August 22–25, 2004*, New York: Association for Computing Machinery, pp 563–8.

Goldwasser, S. (1997) 'Multi party computations: past and present', in J.E. Burns and H. Attiya (eds) *PODC '97: Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing, Santa Barbara CA, August 21–24, 1997*, New York: Association for Computing Machinery, pp 1–6.

Government Digital Service Cabinet Office (2016) 'Data science ethical framework', 19 May, updated 13 June 2018, available from: https://www.gov.uk/government/publications/data-science-ethical-framework [accessed 2 February 2022].

Hacigümüş, H., Iyer, B., Li, C. and Mehrotra, S. (2002) 'Executing SQL over encrypted data in the database-service-provider model', in B. Moon, D. DeWitt and M.J. Franklin (eds) *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, Madison WI, June 3–6, 2002*, New York: Association for Computing Machinery, pp 216–27.

Halevy, A. Ferrer, C.C. Ma, H., Ozertem, U., Pantel, P., M Saeidi, M., Silvestri, F. and Stoyanov, V. (2020) 'Preserving integrity in online social networks', arXiv, 25 September, art 10311v3, available from: https://arxiv.org/abs/2009.10311 [accessed 3 February 2022].

Hasan, M., Rundensteiner, E. and Agu, E. (2019) 'Automatic emotion detection in text streams by analyzing Twitter data', *International Journal of Data Science and Analytics*, 7(1): 35–51.

Hern, A. (2014) 'New York taxi details can be extracted from anonymised data, researchers say', *The Guardian*, 27 June, available from: https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn [accessed 2 February 2022].

ICO (Information Commissioner's Office) (2017) 'Guide to data protection', updated 29 April 2019, available from: https://ico.org.uk/media/for-organisations/guide-to-data-protection-1-1.pdf [accessed 3 February 2022].

Inkpen, D., Roche, M. and Teisseire, M. (2018) 'Guest editorial: special issue on environmental and geospatial data analytics', *International Journal of Data Science and Analytics*, 5(2/3): 81–2.

Johnson, W.B. and Lindenstrauss, J. (1984) 'Extensions of Lipchitz mapping into Hilbert space', *Contemporary Mathematics*, 26: 189–206.

Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. (2003) 'On the privacy preserving properties of random data perturbation techniques', in *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, New York: IEEE, pp 99–106.

Kersting, K. (2020) 'Making deep neural networks right for the right scientific reasons', workshop on bias and fairness in AI. BIAS 2000, available from: https://sites.google.com/view/bias-2020/recordings [accessed 3 February 2022].

Kifer, D. (2009) 'Attacks on privacy and deFinetti's theorem', in C. Binnig (ed) *SIGMOD '09: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence RI, 29 June 2009–2 July 2009*, New York: Association for Computing Machinery, pp 127–38.

Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016) 'How we analyzed the COMPAS recidivism algorithm', ProPublica, 23 May, available from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [accessed 23 September 2020].

LeFevre, K., DeWitt, D.J. and Ramakrishnan, R. (2005) 'Incognito: efficient full-domain k-anonymity', in F. Ozcan (ed) *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, June 2005*, New York: Association for Computing Machinery, pp 49–60.

Li, N., Li, T. and Venkatasubramanian, S. (2007) 't-closeness: privacy beyond k-anonymity and l-diversity', in R. Chirkova, A. Dogac, M.T. Özsu, and T.K. Sellis (eds) *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, 15–20 April, 2007, Istanbul*, New York: IEEE, pp 106–15.

Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M. (2006) 'l-diversity: privacy beyond k-anonymity', in L. Liu, A. Reuter, K.-Y, Whang and J. Zhang (eds) *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3–8 April 2006, Atlanta, GA*, New York: IEEE, p 24.

Meyerson, A. and Williams, R. (2004) 'On the complexity of optimal k-anonymity', in C. Beeri (ed) *PODS '04: Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, New York: Association for Computing Machinery, pp 223–8.

Monreale, A. (2011) 'Privacy by design in data mining', PhD thesis, University of Pisa.

Monreale, A., Rinzivillo, S., Pratesi, F., Giannotti, F. and Pedreschi, D. (2014) 'Privacy-by-design in big data analytics and social mining', *EPJ Data Science* 3: art 10, available from: https://doi.org/10.1140/epjds/s13688-014-0010-4 [accessed 3 February 2022].

Narayanan, A. and Shmatikov, V. (2008) 'Robust de-anonymization of large sparse datasets', in *2008 IEEE Symposium on Security and Privacy (SP 2008)*, Washington, DC: IEEE, pp 111–25.

OECD (Organisation for Economic Co-operation and Development) (2013) *The OECD Privacy Framework*, available from: https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf [accessed 3 February 2022].

Pellungrini, R. (2020) 'Modeling and Predicting Privacy Risk in Personal Data', PhD thesis, University of Pisa.

Pratesi, F. (2017) 'Privacy Risk Assessment in Big Data Analytics and User-Centric Data Ecosystems', PhD thesis, University of Pisa.

Quintarelli, S. (2020) 'Algorithmic accountability' [Affidabilità e responsabilità degli algoritmi] Fondazione Ugo Bordoni webinar [in Italian], available from: https://youtu.be/jLbSAXdwLrs [accessed 23 September 2020].

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?": explaining the predictions of any classifier', in B. Krishnapuram (ed) *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco August 13–17, 2016*, New York: Association for Computing Machinery, pp 1135–44.

Rizvi, S.J. and Haritsa, J.R. (2002) 'Maintaining data privacy in association rule mining', in *Proceedings of the 28th International Conference on Very Large Data Bases Hong Kong 20–23 August 2002*, np [US]: VLDB Endowment, pp 682–93.

Rodríguez-González, A., Vakali, A., Mayer, M.A., Okumura, T., Menasalvas-Ruiz, E. and Spiliopoulou, M. (2019) 'Introduction to the special issue on social data analytics in medicine and healthcare', *International Journal of Data Science and Analytics*, 8(4): 325–6.

Sanil, A.P., Karr, A.F., Lin, X. and Reiter, J.P. (2004) 'Privacy preserving regression modelling via distributed computation', in W. Kim (ed) *KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: Association for Computing Machinery, pp 677–82.

Sea Data Net (2006) Pan-European Infrastructure for Ocean and Marine Data Management, available from: https://www.seadatanet.org/ [accessed 3 February 2022].

SKA Organisation (2011) Square Kilometre Array: Exploring the Universe with the World's Largest Radio Telescope, available from: https://www.skatelescope.org/science/ [accessed 3 February 2022].

SoBigData (2015) Social Mining and Big Data Ecosystem: A Research Infrastructure (RI), available from: http://project.sobigdata.eu/ [accessed 3 February 2022].

Sweeney, L. (2000) *Uniqueness of Simple Demographics in the U.S. Population*, Technical Report LIDAP-WP4, School of Computer Science, Data Privacy Laboratory, Carnegie Mellon University, Pittsburgh, PA.

Sweeney, L. (2002) 'k-anonymity: a model for protecting privacy', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557–70.

Uria-Recio, P. (2018) '5 principles for big data ethics', Towards Data Science, 14 September, available from: https://web.archive.org/web/20190324111737/https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3 [accessed 3 February 2022].

von der Leyen, U. (2019) 'A union that strives for more: my agenda for Europe: political guidelines for the next European Commission 2019–2024', available from: https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf [accessed 3 February 2022].

World Economic Forum (2016) 'The impact of digital content: opportunities and risks of creating and sharing information online', Geneva: WEF, available from: www3.weforum.org/docs/GAC16/Social_Media_Impact_Digital.pdf [accessed 3 February 2022].

Yao, A.C. (1982) 'Protocols for secure computations', in *23rd Annual Symposium on Foundations of Computer Science (SFCS 1982)*, New York: IEEE, pp 160–4.

Zhan, J.Z., Matwin, S. and Chang, L.W. (2005) 'Privacy-preserving collaborative association rule mining', in S. Jajodia and D. Wijesekera (eds) *Data and Applications Security XIX: Proceedings of the 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Storrs, CT, USA, August 7–10, 2005*, Berlin: Springer, pp 153–65.

Zhang, P., Tong, Y., Tang, S. and Yang, D. (2005) 'Privacy preserving naive Bayes classification', in X. Li, S. Wang and Z.Y. Dong (eds) *Advanced Data Mining and Applications: Proceedings of the First International Conference, ADMA 2005, Wuhan, China, July 22–24, 2005*, Berlin: Springer, pp 744–52.

Zook, M., Barocas, S., boyd, d., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A., Nelson, A. and Pasquale, F. (2017) 'Ten simple rules for responsible big data research', *PLoS Computational Biology*, 13(3): art e1005399, available from: https://doi.org/10.1371/journal.pcbi.1005399 [accessed 3 February 2022].