Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Value-Sensitive Design & Privacy-by-Design technologies for big data analytics 1***
Deliverable No.	***D2.5***
Delivery Date	***31 August 2016***
Authors	***Stefanie Hänold (LUH), Nikolaus Forgó (LUH), Anna Monreale (UNIPI), Salvatore Ruggieri (UNIPI), Jeroen van den Hoven (TUDelft), René Mahieu (TUDelft), David van Putten (TUDelft)***

SoBigData receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
DOCUMENT	
Deliverable No.	D2.5
Deliverable Title	Value-Sensitive Design & Privacy-by-Design technologies for big data analytics
Contractual Delivery Date	31 August 2016
Actual Delivery Date	14 September 2016
Author(s)	Stefanie Hänold (LUH), Nikolaus Forgó (LUH), Anna Monreale (UNIPI), Salvatore Ruggieri (UNIPI), Jeroen van den Hoven (TUDelft), René Mahieu (TUDelft), David van Putten (TUDelft)
Editor(s)	Anna Monreale (UNIPI), René Mahieu (TUDelft), David van Putten (TUDelft), Valerio Grossi (CNR)
Reviewer(s)	Salvatore Ruggieri (UNIPI), Valerio Grossi (CNR)
Contributor(s)	Francesca Pratesi (CNR)
Work Package No.	WP2
Work Package Title	WP2 - NA1_Legal and Ethical Framework
Work Package Leader	TUDelft
Work Package Participants	CNR, UNIPI, LUH, TUDelft
Dissemination	Public
Nature	Report
Version / Revision	V1.0
Draft / Final	Final
Total No. Pages (including cover)	34
Keywords	Privacy, GDR, Value-sensitive Design, applied ethics

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by "big data". Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

This document contains information on SoBigData core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the SoBigData Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

GLOSSARY

ABBREVIATION	DEFINITION
DPD	Data Protection Directive
GDPR	General Data Protection Regulation
IT	Information Technology
VSD	Value-sensitive Design

TABLE OF CONTENT

DELIVERABLE SUMMARY

This deliverable gives a history of design perspectives and ethics applied to technical innovation. It introduces the concept of value-sensitive design and explains how the application of value-sensitive design is essential for implementing new technologies in society in a way that corresponds with our ethical views of good society. And it lists the core features that constitute this approach. It then introduces and defines a notion of responsible innovation and encourages the implementation of responsible innovation through the application of value-sensitive design in the development of ICT.

In the following sections the proposed method is applied to design in the field of big data. In particular, there is a focus on integrating privacy discrimination preventing measures into the designs. One of the main methods for preserving privacy in big data is the application of anonymization. We show, on the basis of current and upcoming European law, that the SoBigData research infrastructure should aim to incorporate the state in the art of anonymization techniques and discrimination-preventing techniques and we provide an overview of the current state of the art.

EXECUTIVE SUMMARY

This deliverable explains the Value-Sensitive Design approach of the SoBigData project. The SoBigData infrastructure is connected to the European Research Infrastructure for Big Data and Social Mining and accumulates various datasets from different sources, including social media content (like tweets, blogs, etc), call graphs from mobile phone call data, networks crawled from many online social networks, including Facebook and Flickr, etc. This mass collection of data raises difficult technical questions on how to design the research infrastructure in such a way that it integrates sensible limitations on the kind of data collected, informed consent, division of legal responsibilities and takes into account possible harms and injustices to the individuals whose information is being collected.

This document contains an overview of the current state of the art in value-sensitive designs for social scientific research with big data. In particular it methodologies for privacy by design and fairness by design.

1 RELEVANCE TO SOBIGDATA

1.1 PURPOSE OF THIS DOCUMENT

The purpose of this document is to specify the basic components of the concept of Value-Sensitive Design, which has the potential to provide a framework for the ethical guidance of SoBigData.

1.2 RELEVANCE TO PROJECT OBJECTIVES

A core objective of the SoBigData Research Infrastructure is to provide an environment where it is possible to reap the benefits of big data analytics while at the same time respecting fundamental humanist values. In order to achieve this goal project partners will be informed regarding the current state of the art in techniques that are available for dealing with big data in an ethical way.

1.3 SOBIGDATA PROJECT DESCRIPTION

SoBigData serves the wide cross-disciplinary community of data scientists, i.e., researchers studying all aspects of societal complexity from a data- and model-driven perspective, including data and text miners, visual analytics researchers, socio-economic scientists, network scientists, political scientists, humanities researchers, and more.

1.4 RELATION TO OTHER WORKPACKAGES

This deliverable gives an overview of techniques that can be used in the development of the research infrastructure. For WP 4 it can be used as a basis for the curriculum of value centred approach of using big data. Making sure that the next generation of big data researchers is aware of and trained in the best available techniques.

For Work Packages 8, 9 and 10 this work package can help to design the infrastructures in such way that best practices (see D2.2) will be encouraged.

1.5 STRUCTURE OF THE DOCUMENT

The rest of this deliverable is organised into an ethical and a technical part. Section 2 describes the potential of applying value-sensitive design as a way to overcome the difficulty of keeping ethical control in an environment of fast paced change. Section 3 contains an overview of the current state of the art in privacy enhancing and discrimination preventing techniques in the field of big data analytics.

2 VALUE-SENSITIVE DESIGN IN IT INFRASTRUCTURES

In the middle of the 20th century, scholars in the social sciences and humanities have reflected on how the telegraph, the telephone and TV have shaped our societies[1]. In the last 30 years, researchers in a variety of disciplines such as technology assessment, computer ethics, information and library science, science and technology studies and cultural and media studies have conducted research into the way new media, computers and mobile phones have turned a wired society into a full fledged digital society. In the last 10 years we have entered a new phase of the digital shaping of society. We are trying to come to grips with artificial intelligence, big data, social media, smartphones, robotics, the Internet of Things, apps and bots, self-driving cars, deep learning and brain interfaces. New digital technologies have now given rise to a hyper-connected society. IT is not only getting in between people, but it is also getting under our skin and into our heads – often literally.

Our standard ways of keeping tabs on technology by means of information technology assessment, tech policy and regulation, soft law, ethical codes for IT professionals, ethical review boards (ERBs) for computer science research, standards and software maturity models and combinations thereof, are no longer sufficient to lead us to a responsible digital future. Our attempts to shape our technologies are often too late and too slow (e.g. by means of black letter law) or too little or too weak (e.g. codes of conduct). The field of privacy and data protection is an example of both. Data protection lawyers are constantly trying to catch up with the latest in big data analysis, the Internet of things, deep learning and sensor and cloud technology. On any given day, we often find ourselves trying to regulate the technology of tomorrow with legal regimes of yesterday. This gives rise to the question 'How should we make our ethics bear upon high impact and dynamical digital phenomena?'

2.1 HOW TO DO ETHICS OF IT NOW?

The first thing we need to realize is that the technologies we end up using are consolidated sets of choices that were made in their design, development and implementation. These choices are about e.g. interfaces, infrastructures, algorithms, ontologies, code, protocols, integrity constraints, architectures, governance arrangements, identity management systems, authorization matrices, procedures, regulations, incentive structures, monitoring and inspection and quality control regimes. We build a social world that is shaped by the algorithms that determine how far our messages reach into our networks, what is recommended to us on the basis of what the system has learned about our search history and preferences, what is filtered out and how our reputation is built. We inhabit parametrized life worlds with properties and dynamics that we are mostly unaware of, or that we hardly understand, in case we are aware of them. Digital technology determines how we interact with each other, what we end up seeing and what we end up thinking. The filter bubbles described by Eli Pariser (2011) and Epstein and Robertson's description of Search Engine Manipulation effects (2015) provide examples. The technology that we are using is thus not neutral, since its design is informed by the world views and values of its makers. Once their ideas, values and assumptions have been embedded or expressed in digital artefacts, they start to influence the options, behavior and thinking of users. The digital technologies, our tablets, laptops, and smartphones with their software, apps, user interfaces and default settings form our 'choice architectures' (Sunnstein and Thaler 2010) and our

[1] A good example is the work of Ithiel de Sola Pool in the mid 20th century. See, for example, *Politics in Wired Nations*, Selected Writings, Transaction Publishers, London/New York.

'extended minds' (Clark and Chalmers 1998). The dilemmas and ethical problems we confront are a function of the programs that are running.

Recent thinking about ethics of IT and computer science has therefore focused on how to develop pragmatic methodologies and frameworks that assist us in making moral and ethical values integral parts of research and development and innovation processes at a stage in which they can still make a difference. These approaches seek to broaden the criteria for judging the quality of information technology to include a range of moral and human values and ethical considerations. Moral values and moral considerations are construed as requirements for design. This interest for the ethical design of IT arises at a point in time where we are at a crossroads of two developments: first, "a value turn in engineering design" and on the other hand "a design turn in thinking about values"

First, when the computer was introduced around the middle of the twentieth century, scholarly attention was mainly focused on the technology itself. The computer was developed without too much thought about (1) the use and application in real life, or (2) the social, organizational and political changes it would require to function properly or the impact it would have on society. Computers were a new and fascinating technology: solutions looking for problems. The technology initially appeared to be 'context-free', 'context-independent' and neutral. In the seventies and eighties, attention was increasingly drawn to the context of the technology, i.e. real organizations, (human) user needs and requirements, work conditions, etc. The social and behavioral sciences became increasingly involved with information technology (IT) in the form of (i) human-computer interaction, (ii) participatory design and (iii) social informatics. However, these efforts and commitments were initially mainly focused on a limited set of values, such as user-friendliness and worker-safety. Furthermore, the social and organizational context was often taken into account only as a way to identify potential barriers to the successful implementation of systems and to prevent failed investments. In the first decade of the 21st century, the successful application of information technology is increasingly seen as being dependent on its capacity to accommodate human values. Human beings, whether in their role as employers, consumers, citizens, or patients, have moral values, moral preferences and moral ideals. Information technology cannot and ought not to be at odds with them, and preferably should support and express them. In every society, there are ongoing moral and public debates about liability, equality, property, privacy, autonomy and accountability. Successful implementation is more and more construed in terms of how and to what extent values are taken into account in the design and architecture of systems. Values may even become driving factors in the development of IT instead of being an impediment in the design of information technology. We seem to have entered a third phase in the development of IT that we would like to refer to as "The Value Turn in IT", where the needs and values of human users, as citizens, or patients, are considered in their own right and not simply as a side constraint on successful implementation.

Secondly, simultaneous to the development of the views on technology and society, a development in ethics occurred during the course of the last century. From a predominantly meta-ethical enterprise in the beginning of the 20th century, where the focus was on questions concerning the meaning of ethical terms such as "good" and "ought" and on the cognitive content and truth of moral propositions, the philosophical climate changed in the sixties and ethics witnessed an "Applied Turn". Moral philosophers started to study problems and practices in the professions, issues in public policy and public debate. In the USA, especially, a notable development took place, as philosophers gradually started to realize that philosophy could contribute to social and political debates by clarifying terms and structuring arguments, e.g. concerning the Vietnam War and civil rights, abortion, environmental issues, animal rights, and euthanasia. The focus at this point was on the application of normative ethical theory, utilitarianism or Kantianism, for instance, to practical problems. There often remained a considerable gap with the real world between the prescriptions derived from general theories and the results of the prescriptions in the world of policy making and the

professional practice. However, in the last decade, applied ethics has developed into an even more practical discipline as emphasis is now being placed by some authors on the design of institutions, infrastructure and technology, as the shaping factors in our lives and in society.

If ethics wants (to help or to contribute to) real and desirable moral changes in a digital world then digital systems, institutions, infrastructures and applications themselves need to be designed to be demonstrably in accordance with our shared moral values. This design perspective does not only apply to digital technology, but also to other fields of engineering and other sectors in society. Ethicist will have to devote a good part of their attention to design in order to be relevant in the 21st century. This notable shift in perspective in practical ethics might be termed "The Design Turn in Applied Ethics" (Van den Hoven et al. 2016, forthcoming).

This has given rise to a different and pragmatic approach to ethics of IT; that goes by different names, but focuses on design and design for values as moral requirements early in the development of new functionality.

2.2 VALUE-SENSITIVE DESIGN

As a strong proponent of private transport, famous architect and urban planner Robert Moses designed low overpasses on New York parkways, so that cars could easily access e.g. Jones Beach, while at the same preventing buses to pass under. This turned out to have severe social and political implications, as Langdon Winner (1980) pointed out, as the poor and (mainly) colored population – who are largely dependent on public transport –, were prevented from accessing Jones Beach. Indirectly, the overpass functioned as a border-mechanism separating the wealthy from the poor with respect to the area that lies behind. Even if it is still contested whether Moses' design was consciously intended to have the implication of 'natural' or even racial selection as it did, according to Winner it is nevertheless a clear-cut illustration of the political dimensions that artifacts may have. With his account of "The Politics of Artifacts", he was one of the first to point to the political and social ideologies, values and biases our technologies have embedded in them.

Other studies into the philosophy and sociology of technology have also revealed numerous illustrations of the fact that social and political biases and values are incorporated in technical artifacts, systems and infrastructures (see, for example, Cowan 1985[G1] , Lansing 1991, Latour 1992, Mumford 1964). The examples in these studies illustrate how technologies tend to promote certain ideologies, while obscuring others. Batya Friedman, Helen Nissenbaum, Jeff Bowker and other scholars in ethics of information technology have extended this research into questions of how information technologies specifically can carry values and contain biases. The presumption here is that technology is not neutral with respect to values. Value-Sensitive Design (VSD) recognizes that the design of technologies bears "directly and systematically on the realization, or suppression, of particular configurations of social, ethical, and political values" (Flanagan et al. 2008).

The idea of making social and moral values central to the design and development of new technology originated at Stanford in the 1970's, where it was a central subject of study in Computer Science. It has now been adopted by many research groups and is often referred to as Value-Sensitive Design (VSD). Various groups in the world are now working on this theme. Batya Friedman (Friedman 1997; 2002; 2004) was one of the first to formulate this idea of VSD, others have followed with similar approaches, e.g. 'Values in

Design' at University of California (Bowker; Gregory) at Irvine and NYU (Nissenbaum 2001) and 'Values for Design' (Van den Hoven 2007; Brey 2001; Friedman 1999; Friedman et al. 2002; Camp; Flanagan et al. 2005; Flanagan et al. 2008; Van den Hoven 2009, 2015)[2].

These approaches share the following features:

First, there is the claim that values can be expressed and embedded in technology. In the way that Moses' racist preferences were expressed in the low hanging overpasses[3]. Values and moral considerations can, through their incorporation in technology, shape the space of action of future users, i.e. they can affect the set of affordances and constraints of users. A road from A to B allows one to drive to B, but not to C. Large concrete walls without doors make it necessary to take a detour. Architects and town planners have known this for quite some time. If values can be imparted to technology and shape the space of actions of human beings, then we need to learn to explicitly and transparently incorporate and express shared values in the things we design and make. And what is more we need to accept accountability for the process to all who are directly or indirectly affected.

Secondly, values and choice made by some will have real effects (often not obvious) on those who are directly or indirectly affected. A good example of how this works can be found in the recent work of Cass Sunstein entitled *Nudge*, which construes the task of applied ethicists and public policy makers as a matter of 'choice architecture' (Sunnstein and Thaler 2010, Van den Hoven, 2016, forthcoming). Think for example of the person who arranges the food in your university lunchroom. That person is your choice architect insofar as he is arranges the things from which you can choose, and by doing so makes some of your choice more likely than others. For example by placing the deep fried stuff almost beyond reach and the healthy fruit and veggies in front, the consumer is invited (not forced) to go for the healthy stuff (the nudge). Speed bumps and the 'fly' in men's urinals are other examples of persuasion and nudging by technology. Digital technologies, in the form of computer interfaces, apps, menus, webpages, search engines provide paradigm cases of choice architectures that have real impact on how people choose, act and think.

Thirdly, there is the claim that conscious and explicit thinking about the values that are imparted to our inventions is morally significant. Churchill famously observed in front of the House of Commons: "first we shape our dwellings and then our dwellings start to shape us". Technology and innovation are formidable shapers of human lives and society. It is therefore very important to think about what we are doing to ourselves and to each other by means of technology.

A final feature of the value-design approach is that moral considerations need to be articulated early on in the process, at the moment of the design and development, when value considerations can still make a difference. This sounds easier that it in fact is. This desideratum runs into the so-called 'Collingridge dilemma', which states that early in the process of development of a technology, the degrees of freedom for design are significant, but information that could inform design is relatively scarce, while later on in the

[2] See for an overview (Alina Huldtgren's overview, Design for values in IT) in Van den Hoven, Vermaas and Van de Poel, Springer, 2015). In 2015 an international workshop was held to map out the challenges of Value Sensitive Design in the next decade, see https://www.researchgate.net/publication/283435670_Charting_the_Next_Decade_for_Value_Sensitive_Design In 2016 A follow up international Lorentz workshop is held https://www.lorentzcenter.nl/lc/web/2016/852/description.php3?wsid=852&venue=Oort

[3] There is some controversy over the true motives of Robert Moses, but Winner's example has become paradigmatic in this context and there are a panoply of examples to the same effect.

development of the technology, as information starts to become available, the degrees of freedom in design have diminished.

According to this design approach to ethics of technology ethical analysis and moral deliberation should not be construed as abstract and relatively isolated exercises resulting in considerations situated at a great distance from science and technology, but that instead they should be utilized at the early stages of the research and development. Moreover, they should be construed as non-functional or supra-functional requirements on a par with functional requirements that are used in design. Moral considerations deriving from fundamental moral values (e.g. equity, justice, privacy, security, responsibility) should be decomposed to the point that they can be used alongside other functional requirements to inform design at an early stage. The gradual functional decomposition of supra functional requirements results in the moral specifications.

2.3 RESPONSIBLE INNOVATION[4]

The division of Google concerned with innovations, Google X, has worked on Google Glass which they tested in 2014 and they stopped as a project in 2015. The glasses allowed one to have voice controlled internet access and augmented reality features. Although Google promised not to make face recognition features available for this wearable platform, there were many privacy, safety and security concerns. The idea that large numbers of people would be looking at each other through the lens of a Google device and that people would constantly be checking out things and other people in fairly inconspicuous ways, while surreptitiously taking pictures and capturing data, met with too much public resistance to continue the innovation project. Assuming that Google did not expect upfront to be out of touch with the ethics of society, this seems like an example of an innovation that was discontinued as a result of a failure to deal with the relevant moral considerations.

The Netherlands has learned similar interesting lessons about ethics and digital innovation in the first decade of the 21st century. A first instructive case was the attempt to introduce smart electricity meters nationwide. In order to make the electricity grids more efficient and meet the EU CO2 reduction targets by 2020, every household in The Netherlands would have to be transformed into an intelligent node in the electricity network. Each household could thus provide detailed information about electricity consumption and help electricity companies to predict peaks and learn how to "shave off" the peaks in consumption patterns. After some years of R&D, a plan to equip every Dutch household with a smart meter was proposed to parliament. In the meantime however, opposition to the proposal by privacy groups had gradually increased over the years (Abdulkarim 2011). The meter was now seen as a 'spying device' and a threat to the personal sphere of life, because it could take snapshots of electricity consumption every 7 seconds, store data in a database of the electricity companies for data mining, and provide the most wonderful information about what was going on inside the homes of Dutch citizens. With some effort, it could even help to tell which movie someone had been watching on a given night. By the time the proposal was brought to the upper house of the Dutch parliament for approval, public concern about the privacy aspects had become very prominent and the upper house rejected the plan on data protection grounds. The European Commission, being devoted to the development of smart electricity grids in its member states, feared that the Dutch reaction to this type of innovation would set an example for other

[4] This draws upon Van den Hoven's discussion of the relation between Responsible Innovation and Value Sensitive Design in Richard Owen's Responsible Innovation (2013).

countries and would jeopardize the EU wide adoption of sustainable and energy saving solutions in an EU market for electricity (Abdulkarim 2009).

Another story – not very different from that of the smart meter – is the introduction of a nation-wide electronic patient record system in The Netherlands. After 10 years of R&D and preparations, lobbying, stakeholder consultation and debates – and last but not least an estimated investment of 300 million Euro – the proposal was rejected by the upper house in parliament on the basis of privacy and security considerations (Tange 2008; Van Twist 2010).

Clearly these innovations in the electricity system and health care system could have helped The Netherlands to achieve cost reduction, greater efficiency, sustainability goals, and in the case of the electronic Patient Record System, higher levels of patient safety. In both cases, however, privacy considerations were not sufficiently incorporated in the plans so as to make them acceptable. If the engineers had taken privacy more seriously right from the start and if they had made greater efforts to incorporate and express the value of privacy into the architecture at all levels of the system, transparently and demonstrably, then these problems would probably not have arisen.

The important lesson to learn from these cases is that values and moral considerations (i.e. privacy considerations) should have been taken into account as "non-functional requirements" at a very early stage of the development of the system, alongside the functional requirements, e.g. storage capacity, speed, bandwidth, compliance with technical standards and protocols. A real innovative design for an Electronic Patient Record System or a truly smart electricity meter, would thus have anticipated or pre-empted the main moral concerns and accommodated them into its design, reconciling efficiency, privacy, sustainability and safety. Value-sensitive thinking at the early stages of development at least might have helped engineers to do a better job in this respect. There is a range of fine-grained design features that could have been considered and that could have been presented as choices for consumers. A smart meter is not a given, it is to a large extent what we design and make it to be. Respect for privacy can be built in (Garcia and Jacobs 2011; Jawurek et al. 2011). There are several objections against this suggestion. The first is that of moralism, another is that of relativism. Should values be built-in at all and if so which values should be 'built-in' and with which justification? There seem to such a great variety of values. Empirical research even seems to indicate that there is no coherent and stable set of European values, let alone global values. Both objections, it seems, can be addressed satisfactorily. No technology is ever value neutral (Van den Hoven 2012b). It is always possible that a particular technology, application or service, favors or accommodates a particular conception of the good life, at the expense of another, whether this was intended or not. There is therefore virtue in making particular values at play explicit and evaluate how their implementations works out in practice and adjust our thinking accordingly. Being overly impressed in the field of technology by objections of moralism and relativism and as a result would abstain from working with values in an explicit and reflective way, we would run the risk that commercial forces, routine, bad intentions would reign free and impose technology with values that were not discussed and reflected upon by relevant parties.

Serious attention to moral considerations in design and R&D may not only have good moral outcomes, but may also lead to good economic outcomes. Consider the case of so-called 'privacy enhancing technologies'. The emphasis on data protection and the protection of the personal sphere of life is reflected in demanding EU data protection laws and regulation. The rest of the world has always considered the preoccupation with privacy as a typically European political issue. As a result of the sustained and systematic attention to data protection and privacy, Europe has become an important cradle of new products and services in the field of Privacy by Design or Privacy Enhancing Technologies. Now, the Big Data society is on our doorstep and many computer users – also outside of Europe – are starting to appreciate products and services that can accommodate user preferences and values concerning privacy, security and identity, Europe has a

competitive advantage and is turning out to be an important commercial player in this branch of the IT industry.

Innovation can thus take the shape of (engineering) design solutions to situations of moral overload (Van den Hoven et al. 2012). One is morally overloaded when one is burdened by conflicting obligations or conflicting values, which cannot be realized at the same time. But as we saw above, conflicts of privacy and national security seem amenable to resolution by design and innovation in the form of privacy enhancing technologies. Conflicts between economic growth and sustainability were resolved by sustainability technology. Some think of these solutions as mere "technical fixes" and do not construe them as genuine solutions to moral problems. We do not take a stance on this issue here. We just want to point out that in such cases it seems to me that we have an obligation to bring about the required change by design or innovation (Ibidem). The principle that seems to be operative can be formulated as follows: 'If a contingent state of the world at time t1 does not allow us to satisfy two or more of our moral values or moral obligations at the same time, but we can bring about change by innovation in the world at t1 that allows us to satisfy them all together at a later time t2, then we have a moral obligation at t1 to innovate'(Van den Hoven 2013).

This is an important part of what responsibility implies in the context of innovation. It construes innovation as a second order moral obligation: the obligation to bring about a change in the world that allows us to make more of our first order moral obligations (e.g. for security and privacy, for economic growth and sustainability, safety and security) than we could have done without the innovation. Normally, the principle that 'ought' implies 'can' holds, but a noteworthy feature of this second-order obligation to innovate is that it does not imply 'can'. This means that we may be under the obligation to come up with an innovation that solves our problem, although success is not guaranteed.

It may seem fairly obvious to claim that we have a higher order moral obligation to innovate when it leads to moral progress, but it requires a considerable shift in our thinking about innovation. We need to learn to think – as argued above - of ethical considerations and moral values in terms of requirements in design and research and development at an early stage. Value discourse should therefore not be left on an abstract level, but needs to be operationalized or 'functionally decomposed', as is often done with high level and abstract requirements in engineering and design work. The process of functional decomposition eventually leads to a level of detail that points to quite specific design features of the system, the 'moral specs'. This requires engineers to be value-focused in their thinking and capable of articulating the values at play with different stakeholders (Pommeranz 2012b).

If some innovative organization or process would be praised in virtue of its being "responsible", this would imply, among other things, that those who initiated it and were involved in it must have been accommodated as moral and responsible agents, i.e. they must have been enabled:

(A) to obtain – as much as possible – the relevant knowledge on (i) the consequences of the outcomes of their actions and on (ii) the range of options open to them and

(B) to evaluate both outcomes and options effectively in terms of relevant moral values (including, but not limited to well-being, justice, equality, privacy, autonomy, safety, security, sustainability, accountability, democracy and efficiency).

In light of (I) and (II) above, we suggest that another implication of the notion of Responsible Innovation is the capability of relevant moral agents

(C) to use these considerations (under A and B) as requirements for design and development of new technology, products and services leading to moral improvement

On the basis of this characterization of innovation and the implications (A), (B) and (C) we may characterize Responsible Innovation in summary as follows:

(III) Responsible Innovation is an activity or process which may give rise to previously unknown designs either pertaining to the physical world (e.g. designs of buildings and infrastructure), the conceptual world (e.g. conceptual frameworks, mathematics, logic, theory, software), the institutional world (social and legal institutions, procedures and organization) or combinations of these, which – when implemented – expand the set of relevant feasible options regarding solving a set of moral problems''.

2.4 CONCLUSION: VALUE-SENSITIVE DESIGN

If given a choice we would prefer a situation where only those digital technologies would gain social acceptance that were morally acceptable. We would prefer a situation where technologies deemed morally unacceptable would also not be socially accepted and gain currency for precisely that reason. In order to bring about this ideal situation, it would be helpful if the digital products and services, our systems and software, could be made to wear the index of moral acceptability on their sleeves and could be made in such a way that they send honest signals to users about their moral quality and the values that have been used to shape them. In order to achieve this level of accountability and transparency Ethics of IT in the 21st century will have to be developed along design lines sketched out here.

3 PRIVACY- AND FAIRNESS-BY-DESIGN IN BIG DATA ANALYTICS

The big data originating from the digital breadcrumbs of human activities, sensed as a by-product of the ICT systems that we use everyday, record the multiple dimensions of social life: automated payment systems record the tracks of our purchases; search engines record the logs of our queries on the web; wireless networks and mobile devices record the traces of our movements. These big data describing human activities are at the heart of the idea of a "knowledge society", where the understanding of social phenomena is sustained by the knowledge extracted from the miners of big data across the various social dimensions by using social mining technologies. Thus, the analysis of our digital traces can create new opportunities to understand complex aspects, such as mobility behaviors, economic and financial crises, the spread of epidemics, the diffusion of opinions and so on. However, the remarkable opportunities of discovering interesting patterns from these data can be outweighed due to the high risks of ethical issues in data processing and analysis and ethical consequences of their suggestions and predictions. Important ethical risks are: (i) *privacy violations*, when uncontrolled intrusion into the personal data of the subjects occurs, and (ii) *discrimination*, when the discovered knowledge is unfairly used in making discriminatory decisions about the (possibly unaware) people who are classified, or profiled.

Nevertheless, big data analytics and ethics are not necessary enemies. In the literature, some works have shown that many practical and impactful services based on big data analytics can be designed in such a way that the quality of results can coexist while enforcing ethical requirements. The secret is to develop big data analytics technologies that *by-design* enforce ethical value requirements to offer safeguards of fairness.

3.1 PRIVACY-BY-DESIGN IN BIG DATA

Despite the benefits of big data analytics, it cannot be accepted that big data comes at a cost for privacy. Moreover, technology and innovation cannot be stopped because of privacy issues. It is, thus, necessary to find a solution to balance between making use of big data technologies and protecting individuals' privacy and personal data. *We want to point out is that there is no big data without privacy*. If privacy principles are not respected, big data will fail to meet individuals' needs; if privacy enforcement ignores the potential of big data, individuals will not be adequately protected. Therefore, all stakeholders should work together in addressing the new challenges and highlighting privacy as a core value of big data. In the 1990s, Ann Cavoukian introduced the principle of *Privacy-by-design* [Cav09], based on the idea to embed privacy measures and privacy enhancing technologies directly into the design of information technologies and systems. The privacy-by-design paradigm represents a significant innovation with respect to the traditional approaches of privacy protection because it requires a significant shift from a reactive model to proactive one. In other words, the idea is preventing privacy issues instead of remedying them.

3.1.1 LEGAL ASPECTS OF PRIVACY BY DESIGN

The current legal framework to assess big data analytics in the EU is the DPD and the implementing national laws. From May 2018 the Directive will be replaced by the General Data Protection Regulation. As explained in D.2.2 the European framework for data protection only is applicable if personal data is processed. Anonymization of data is not just a measure to prevent the application of data protection regulations. It is also required by the principle of data minimization to de-identify the data as much as the purpose for the

processing allows it. However, it has proven difficult to create a truly anonymous dataset whilst retaining as much of the underlying information as required for the task.[1]

It is to a great extent the steadily evolving technological developments that make it very difficult to determine whether the data that is envisaged for processing is effectively anonymized. This is all the more the case in the field of big data/ smart data analysis as huge amounts of data are combined and for every combination an assessment must be proceeded. It is important to note that this test is a dynamic one and should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed.

Although the Regulation upholds the traditional concept of personal and non-personal data, it also introduces the concept of pseudonymous data on the European level which may be interpreted as a reaction by the European Legislator to the uncertainties that come along with anonymization techniques. For the case that data is pseudonymized, it is explicitly stated now in the Regulation that the data qualifies as personal data and therefore the framework set up by the Regulation will apply. Pseudonymous data in the sense of the Regulation is personal data that has been processed in such a manner that it can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (Art. 4 (5) GDPR). However, the explicit introduction of pseudonymization in the Regulation is not intended to preclude any other measures of data protection. The Regulation allows controllers who apply appropriate safeguards, such as encryption or pseudonymization, some leeway to process the data for a secondary purpose.

The implementation of appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access and against all other unlawful forms of processing is one obligation data controller is carrying. Having regard to the state of the art and the cost of their implementation, such measures shall ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected (Art. 17 (1) DPD). Similarly, the Regulation provides in Art. 32 that the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons. In assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed. The Regulation also list examples that can be used as appropriate, for instance: the pseudonymisation and encryption of personal data the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services; the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident; a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

The GDPR introduces also in Art. 25 the concept of “data protection by design” which intends that privacy should be a feature of the development of a product, rather than something that is tacked on later. The controller shall at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects. In doing so he must consider the state of the art, the

cost of implementation and the nature, scope and context and purposes of processing as well as the risks for the data subject. The data contoller shall also provide appropriate technical and organisational measures for ensuring that, by default, that only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

The Regulation also obliges data controllers in cases where a type of processing poses a high risk to the rights and freedoms of the data subject to carry out , prior to the processing, an assessment of the impact of the envisaged processing operations on the protection of personal data. As paragraph Art. 35 (3) GDPR shows that scientific research may not give reason to create such an obligation out of Art. 35 (1) GGDPR, especially where there is no systematic and extensive evaluation of personal aspects of the data subject and no further use of the information is made for making decisions that affect him or her. The controller shall consult the supervisory authority prior to processing where a data protection impact assessment under Article 35 indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk (Art. 36 (1) GDPR).

3.1.2 EXISTING PRIVACY STRATEGIES

[ENI15] explores the privacy-by-design strategies in the different phases of the big data value chain while Monreale et al. [MRP+14] proposes the instantiation of the privacy-by-design paradigm to the designing of big data analytical services.

[ENI15] classifies privacy models in two main families. The first family includes *k-anonymity* [PS98] *and its extensions* taking care of attribute disclosure, like p-sensitive k-anonymity [TV06], l-diversity [MKA+07], t-closeness [NTV07], (n,t)-closeness [NTS10], and others. The second family is built around the notion of *ε-differential privacy* [D06], along with some variants like crowd-blending privacy [GHL+12] or BlowFish [MK15]. Some of these privacy models are also reviewed in [A2914].

It is important to the note that there is a fundamental difference between k-anonymity and differential privacy: the first one (and its variants) is focusing on anonymizing a data set before its release for further analysis while the second one is about running queries on the data, following a predefined type of analysis, in a way that the answers do not violate individuals' privacy. Although it has been argued that the differential privacy's query-based approach is superior to the "release and forget" approach of k-anonymity, its practical implementation (taking into account the utility-privacy trade-off) is not possible in every data analytics scenario. Thus, k-anonymity still has a dominant role, especially on data releases (i.e. when the query-based model is not applicable).

In many contexts it is not clear what means applying the privacy- by-design principle and which is the best way to apply it for obtaining the desired result. The idea presented in [MRP+14] is to inscribe privacy protection into any analytical process by design, so that the analysis incorporates the relevant privacy requirements from the very start, evoking the concept of privacy-by-design discussed above. The articulation of the general "by design" principle in the big data analytics domain is that higher protection and quality can be better achieved in a goal- oriented approach. In such an approach, the data analytical process is designed with assumptions about:

- the sensitive personal data subject of the analysis;

- the attack model, i.e., the knowledge and purpose of adversary that has an interest in discovering the sensitive data of certain individuals;
- the category of analytical queries that are to be answered with the data.

These assumptions are fundamental for the design of a privacy-aware technology.

Under the above assumptions, it is conceivable to design a privacy- aware analytical process that can:

- transform the data into an anonymous version with a quantifiable privacy guarantee - i.e., the probability that the malicious attack fails;
- guarantee that a category of analytical queries can be answered correctly, within a quantifiable approximation that specifies the data utility, using the transformed data instead of the original ones.

This methodology was applied to guarantee privacy in the following fields.

Privacy in Data Publishing. Monreale et al. [MAA+10] designed a method for the privacy-aware publication of movement data enabling clustering analysis useful for understanding human mobility behavior in specific urban areas. The released trajectories are made anonymous by a suitable process that realizes a generalized version of the original trajectories. The results obtained with the application of this framework show how trajectories can be anonymized to a high level of protection against re-identification while preserving the possibility of mining clusters of trajectories, which enables novel powerful analytic services for info-mobility or location-based services. Similarly, [MPP+14] proposed a privacy-preserving method for anonymizing sequence data such as query logs, sequence of locations and so on, while preserving the quality of the results of sequential pattern mining.

Privacy in Data Mining Outsourcing. Giannotti et al. in [GLM+13] designed a method for a privacy-aware outsourcing of the pattern mining task; in particular, the results show how a company can outsource the transaction data to a third party and obtain a data mining service in a privacy-preserving manner. In this setting, not only the underlying data but also the mined results (the strategic information) are not intended for sharing and must remain private. A survey that outlines a variety of techniques and approaches that address the privacy issues in the context of data mining outsourcing can be found in [MW16] that identify four types of approaches:

1) k-anonymity based approaches which typically transform the original database by adding fake items and/or fake transactions after hiding the meanings of items in the transaction database by replacing items with integers [WCH+07, TYC10, GLM+13, THC13, CBM15].
2) randomization based approaches, which before outsource data add noise in such a way that it is hard to estimate the original values from the perturbed data, while information critical to data mining are still preserved [CL05,CSL07,L13,LCC15,QLW08].
3) differential privacy based approaches which can be *interactive* and *non-interactive*. In the interactive setting, users send queries to the service provider. Queries and/or their responses are modified by the DP mechanism in order to protect privacy [McS09, BLS+10]. While in the non-interactive setting, a data publisher computes and releases a sanitized version of a database, possibly a synthetic database, to the public (including the cloud) for future data mining and analysis [BX13, MCF+11,VSB+13].

4) encryption based approaches, which adopt cryptographic techniques to protect privacy [LXG+12,WCK+09].

Privacy in Distributed Analytical Systems. Monreale et al. in [MWP+13] proposed a method for a privacy-aware distributed mobility data analytics, where we have a untrusted central station that collects some aggregate statistics computed by each individual node that observes a stream of mobility data. The central station stores the received statistical information and computes a summary of the traffic conditions of the whole territory, based on the information collected from data collectors.

3.2 FAIRNESS-BY-DESIGN IN BIG DATA

Social discrimination refers to an unjustified difference in treatment of individuals on the basis of any physical or cultural trait, such as sex, ethnic origin, religion or political opinions. The problems of assessing the presence, extent, nature, and trends of social discrimination (*discrimination discovery*) and of preventing discrimination in automated decision making based on profiling and predictive modeling (*discrimination prevention*) are thus of primary importance. In the last fifty years, such problems have been investigated from a social [New08], legal [BS16,CS14], economic [Arr71,CG11], and, recently, from a data mining perspective [RR14,Zlio15]. The ease of data storage and retention, the ever increasing computing power, and the development of intelligent data analysis and mining techniques make it possible to apply "in-the-large" and to improve classical statistical and econometric techniques for discrimination discovery. On the other hand, the same big data technologies, when applied without any value-sensitive concern, may discover traditional prejudices that are endemic in reality, or patterns of lower performances, skills or capacities of protected-by-law social groups. Decision-making algorithms for profiling and predictive modeling (classifiers, recommender systems, filtering and screening systems, etc.) may then assign the status of general rules to such practices, with the result that these rules may be deeply hidden within obscure models. *Fairness* goes beyond the legal obligations of non-discrimination, and accounts for bias-free predictive models. Another related concept is *algorithm accountability and transparency* [Dia16], which requires mechanisms for disclosure of decision motivation, and for model interpretability and auditability. We survey here the scientific literature on measurement of fairness and on methods for achieving (a certain degree of) fairness. There is yet no general theory of fairness-by-design[2], as in the case of privacy-by-design. Also, there is no shared analytical process of how to achieve fair data mining models. Nevertheless, there are examples case studies on how to apply/challenge the proposed methods (e.g. [RRT13, WH16]). The incorporation of anti-discrimination law requirements in specific application fields, such as credit scoring systems, has been also considered in the econometric literature [LW10].

3.2.1 DEFINITIONS AND MEASUREMENTS OF FAIRNESS

Quantitative definitions of fairness are the building block of approaches for reasoning about the issue. Proposed definitions can be categorized into *group fairness* and *individual fairness*. In group fairness (also called, *statistical parity*) demographics of the individuals receiving any decision outcome are the same as demographics of the underlying population. In *individual fairness*, individuals who are similar receive similar outcomes. Quantitative measures of these two notions are surveyed in [Zlio15], which categorizes them as follows:

Statistical tests check how likely the difference between groups is due to chance. They answer the question: "is there discrimination?"

Absolute measures express the absolute difference between groups, quantifying the magnitude of discrimination.

Conditional measures express how much of the difference between groups cannot be explained by other attributes, also quantifying the magnitude of discrimination.

Structural measures quantify the number of individuals impacted by direct discrimination. They answer the question: "how widespread is discrimination?"

3.2.2 METHODS FOR TESTING/CERTIFYING FAIRNESS OF PREDICTIVE MODELS

Discrimination has been identified in law and social study literature as either direct or indirect. Direct discrimination (also called systematic discrimination or disparate treatment) consists of rules or practices explicitly treat one person less favorably on forbidden grounds than another is, has been or would be treated in a comparable situation. Predictive models can directly discriminate by learning rules against protected social groups. The naive approach of deleting attributes that denote protected groups from the dataset used to train predictive models does not prevent them to discriminate since other attributes (sometimes called redundant encodings) that are strongly correlated with gender, race, etc. could be used as proxies by the learning algorithm. Indirect discrimination (also called disparate impact), is an apparently neutral provision, criterion or practice which results in an unfair treatment of a protected group. An example is redlining, which is the practice of banks of denying credits based on the residence of the applicant. Since residence is a proxy for race, especially in the highly segregated urban cities in the U.S., such a practice actually hides discrimination. The computer science literature tackles the problem of discovering/testing/certifying direct and indirect discrimination in: (1) a dataset of historical decision records; and in (2) predictive models extracted from (big) data.

Approaches for (1) can be applied on the output of a predictive model for checking whether automated decision contain discrimination. Such approaches are based on association rule mining [HD13,RPT10] (tackling group discrimination), k-NN [LRT11] (tackling individual discrimination), bayesian network [MC14], probabilistic causation [BHMR]. Methods for indirect discrimination assume some external knowledge to find out proxy attributes [RPT10,RHK+14].

Approaches for (2) consist of auditing a prediction model to the purpose of understanding why it makes certain decisions. Model creators can build interpretable models, either by explicitly using interpretable structures like decision trees, or by building shadow models that match model outputs in an interpretable way [RSG16]. Outside auditors cannot access/modify the model (which is typically proprietary), and hence rely on black-box methods. Such methods can test for direct discrimination [DSZ16, HPB+14,TAG+15] and indirect discrimination [FFM+15,AFF+16].

3.2.3 METHODS FOR ACHIEVING FAIRNESS

We follow the categorization by [RR14] into four non mutually-exclusive strategies for achieving fairness of data mining models.

Data sanitization. It consists of a controlled distortion or generalization of the dataset at hand. If one before releasing it or before using it to train data mining models, such pre-processing approaches are independent of the learning model and algorithm at hand. State of the art proposals for discrimination-sanitization include approaches using: generalization and randomization of predictive values [DHPRZ12,Rug14]; and goal-directed perturbation of decision values [HD13,KC12,KZC13,LRT11]. A different approach is taken by [FFM+15,AFF+16], where an auditing method is devised to obscure direct and indirect influence of protected attributes over predictions of a classifier.

Algorithm modification. Extensions of machine learning algorithms are devised that take into account fairness constraints. [CW10] considers naıve Bayes models, and proposes training a separate model for each protected group; and, adding a latent variable to model the class value in the absence of discrimination. [KCM10] modifies the entropy-based splitting criterion in decision tree induction to account for attributes denoting protected groups. [KAS12] measures the indirect causal effect of variables modeling grounds of discrimination on the independent variable in a classification model by their mutual information. Then they apply a regularization (i.e., a change in the objective minimization function) to probabilistic discriminative models, such as logistic regression.

Knowledge sanitization. This approach acts on the output of the data mining algorithms by modifiying/regularizing the extracted model in order to remove learnt discriminatory rules. Proposed approaches tackle classification rules [PRT09], Bayesian models [CW10], decision trees [KCM10].

Prediction correction. It consists of the correction of predictions/decisions made by a model on-the-fly. [KKZ12] proposes correcting predictions of probabilistic classifiers that are close to the decision boundary, given that discrimination may occur when there is no clear feature supporting a positive or a negative decision.

A number of tools available online that implement the above techniques is referenced next:

- [AFF+16] https://github.com/cfalk/BlackBoxAuditing
- [HPB+14] https://bitbucket.org/aheneliu/goldeneye
- [LRT11,PRT10,Rug14] http://pages.di.unipi.it/ruggieri/software
- [TAG+15] is integrated in Scipy (https://scipy.org/)

3.2.4 LEGAL ASPECTS OF FAIRNESS (GDPR ARTICLES ON AUTOMATED DECISION MAKING)

The Data Protection Directive does not provide any rules specifically addressing the issue of profiling, but it considers automated individual decisions by granting the right to every person not to be subject to a decision which produces legal effects concerning him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him (Art. 15 (1) DPD).

The Directive contains the possibility to derogate from this principal rule if an automated decision is taken in the course of entering into or performance of a contract, provided the request for the entering into or the performance of the contract, lodged by the data subject, has been satisfied or that there are suitable safeguards for the legitimate interests of the data subject, e.g. arrangements that allow him to give his opinion. In case automated decisions are authorized by law measures to safeguard the data subject's legitimate interest are required (Art. 15 (2) DPD).

The General Data Protection Regulation contains some regulation on automated decision making including profiling[5]. Principally data subjects have the right to avoid being subject to a decision based solely on automated processing, which produces legal effects concerning him or her or significantly affecting him or her (Art. 22 (1) GDPR).

The Regulation sets in Art. 22 (2) a bundle of rules under which circumstances automated decision making is allowed. It is allowed if it

- is necessary for entering into, or performance of, a contract between the data subject and a data controller;
- is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
- is based on the data subject's explicit consent.

In cases automated decision making is allowed the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests. As a minimum the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision must be provided (Art. 22 (3) GDPR).

For special categories of personal data the Regulation is more restrictive when it comes to automated decision making. Special categories of data are personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation (Art. 9 (1) GDPR). Automated decision making must not be based on special categories of data unless the data subject has given explicit consent according to Art. 9 (2a) GDPR or if the processing is necessary for the processing of substantial public interest on the

[5] Profiling is defined as any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements (Art. 4 (4) GDPR).

basis of Union or Member State law (Art. 9 (2g) GDPR) and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

It is a crucial element to implement suitable measures to safeguard the data subject's right, freedoms and legitimate interests. The Recitals of the Regulation provide some guidance. Thereafter such measures should include:

- specific information to the data subject[6] and
- the right to obtain human intervention,
- to express his or her point of view,
- to obtain an explanation of the decision reached after such assessment and
- to challenge the decision.

The controller should also use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect (Recital 71 GDPR). Apart from that the basic rules of the Regulation governing the processing of personal data, such as the legal grounds for processing or data protection principles apply (Recital 72 GDPR) --> D2.2.

It is presumed that the European Data Protection board will provide guidelines to better understand under which circumstances automated decisions are permissible for special categories of data (Recital 72 GDPR.

The Regulation also contains rules on the right to object of the data subject relating to profiling in Art. 21.

The data controller must take his responsibilities seriously. According to Art. 24 GDPR the controller, taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary.

[6] Specific information duties are laid down in Art. 13 (2f) which requires that the controller shall provide the data subject with information about the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

REFERENCES

Section 2:

Al Abdulkarim LO and Lukszo Z (2009) Smart Metering for the Future Energy Systems in the Netherlands. In: Proceedings of the Fourth International Conference on Critical Infrastructures. At: Linköping University, Sweden. [s.l.]: March 27-April 30, 2009 and 28-30 April, 2009.

Brey P (2001) Disclosive Computer Ethics. In: Spinello, RA, Tavani, HT (eds.) Readings in Cyberethics, Jones and Bartlett Publishers Inc., Massachusetts, p. 51-62.

Brey P (2000) Method in Computer Ethics: towards a multi-level interdisciplinary approach. Ethics and Information Technology, 2:3, 1-5.

Camp LJ (n.d.) Design for Values, Design for Trust. Retrieved September 18, 2007, from http://www.ljean.com/design.html

Clark A and Chalmers DJ (1998) The Extended Mind. Analysis 58:10-23.

Cowan, R (1983) More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave. Basic Books. 1983

Epstein R and Robertson RE (2015) The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In: Proceedings of the National Academy of Sciences 112 (33): E4512–E4521. doi:10.1073/pnas.1419828112. ISSN 0027-8424. PMC 4547273. PMID 26243876.

Flanagan M et al. (2005) Values at play: Design tradeoffs in socially-oriented game design. Conference on Human Factors in Computing Systems, 751-760.

Flanagan, M et al. (in press) Values in Design: Theory and Practice. In: Van den Hoven MJ and Weckert J (eds), Information Technology and Moral Philosophy. Cambridge University Press, New York.

Friedman B (1997) Human values and the design of computer technology (CSLI lecture notes; no. 72). Cambridge University Press, New York.

Friedman B (1999) Value-Sensitive Design: A Research Agenda for Information Technology. National Science Foundation, Contract No: SBR-9729633.

Friedman B (2004) Value Sensitive Design, in: Bainbridge WS (ed.) Berkshire Encyclopedia of Human-Computer Interaction. Berkshire Publishing Group, Massachusetts.

Friedman B and Kahn P (2000) New directions: a value-sensitive design approach to augmented reality. Proceedings of DARE 2000 on Designing augmented reality environments, 163-164.

Friedman B and Kahn Jr. PH (2003) Human Values, Ethics, and Design, in: Jacko, J.A. and Sears A. (eds) The Human-Computer Interaction Handbook, Lawrence Erlbaum Associates, 1177-1201.

Friedman B and Kahn Jr. PH and Borning A (2006) Human-Computer Interaction and Management Information Systems - Foundations, in: Zhang, P. and Galetta D. (eds.) Advances in Management Information Systems, Volume 4, M.E. Sharpe, Inc.

Friedman B and Kahn Jr. PH (in press) A Value-Sensitive Design Approach to Augmented Reality, in: Mackay, W.E. (ed.) Design of Augmented Reality Environments, The MIT Press, Massachusetts.

Friedman B et al (2002) Value Sensitive Design: Theory and Methods. (Technical Report 02-12-01).

Garcia,F and Jacobs B (2011) Privacy Friendly Energy Metering via Homomorphic Encryption. Lecture Notes in Computer Science, Vol. 6710, 226-238.

Jawurek M et al. (2011) Plug-in Privacy for smart metering billing. Lecture Notes in Computer Science, 2011, Vol. 6794. 192-210

Lansing JS (1991) Priests and programmers: Technologies of power in the engineered landscape of Bali. Princeton University Press, Princeton, N.J.

Latour B (1992) Where are the missing masses? In: Shaping Technology-Building Society. Studies in Sociotechnical Change, Wiebe Bijker and John Law (eds), MIT Press, Cambridge Mass. p. 225-259.

Mumford L (1964) Authoritarian and Democratic Technics. Technology and Culture, 5(1), 1-8.

Nissenbaum H (2001) How computer systems embody values. IEEE Computer, 34(3), 118-120.

Pariser E (2011) The Filter Bubble: What the Internet is Hiding from You. Penguin: London.

Pommeranz, A (2012b). Designing Human Centered Systems for Reflective Decision Making. Dissertation. TUDelft, Delft.

Tange H (2008) Electronic patient records in the Netherlands. From: http://hpm.org/en/Surveys/BEOZ_Maastricht_-_Netherlands/12/Electronic_patient_records_in_the_Netherlands.html

Thaler RH and Sunstein CR (2008) Nudge. Yale University Press, London/New Haven.

Van den Hoven MJ (2005) Design for values and values for design. Information Age, 7(2), 4-7.

Van den Hoven MJ (2007) Moral Methodology and Information Technology. In: Tavani HT and Himma K (eds). Handbook of Computer Ethics. Wiley, New Jersey.

Van den Hoven MJ (2007) ICT and Value Sensitive Design. In: Goujon P et al. (eds.) The Information Society: Innovation, Legitimacy, Ethics and Democracy, Springer, Dordrecht, p. 67-73.

Van den Hoven MJ (2010) The use of normative theories in computer ethics. In: The Cambridge Handbook of Information and Computer Ethics, Cambridge University Press, New York.

Van den Hoven MJ (2012b) Neutrality and Technology: Ortega Y Gasset on the Good Life. In: Brey P et al. (eds) The Good Life in a Technological Age. Routledge, London, 2012. p. 327-339.

Van den Hoven MJ et al. (2012) Engineering and the Problem of Moral Overload. Science and Engineering Ethics, 2012.

Van den Hoven MJ (2013) Value Sensitive Design and Responsible Innovation. In: Richard Owen, e.a. (eds) Responsible Innovation, Wiley, Chichester, 2013. Pp. 75-85.

Van den Hoven MJ et al. (2016, forthcoming). Designing in Ethics, Cambridge University Press, Cambridge.

Van Twist M (2010) URL: www.rijksoverheid.nl/bestanden/.../rapport-het-epd-voorbij.pdf

Winner L (1980) Do artifacts have politics? Daedalus, 109(1), 121-136

Section 3:

[A2914] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques. 2014.

[AFF+16] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian. Auditing Black-box Models by Obscuring Features. arXiv:1602.07043v1, 2016.

[Arr71] K. J. Arrow. The theory of discrimination. In O. Ashenfelter and A. Rees, editors, Discrimination in Labor Markets, pages 3–33. Princeton University Press, 1971.

[BHMR] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti. Exposing the Probabilistic Causal Structure of Discrimination. arXiv preprint arXiv:1510.00552, 2015.

[BLS+10] R. Bhaskar, S. Laxman, A. Smith, A. Thakurta. Discovering frequent patterns in sensitive data. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Know. Discovery and Data Mining*, 503–512, 2010.

[BS16] S. Barocas, A. D. Selbst. Big data's disparate impact. California Law Review, 104, 2016. Available at SSRN: http://ssrn.com/abstract=2477899.

[BX13] L. Bonomi, L. Xiong. A two-phase algorithm for mining sequential patterns with differential privacy. In *Proc. of the 22nd ACM Int. Conf on Information & Know. Management*, 269–278, 2013.

[Cav09] A.Cavoukian. Privacy-By-Design, 2009.
https://www.ipc.on.ca/images/resources/privacybydesign.pdf

[CBM15] I. Chandrasekharan, P. K. Baruah, and R. Mukkamala. Privacy- preserving frequent itemset mining in outsourced transaction databases. In *Proc. of Int. Conf. on Advances in Computing, Communications and Informatics,* 787–793, 2015.

[CG11] K. K. Charles, J. Guryan. Studying discrimination: Fundamental challenges and recent progress. *Annual Review of Economics*, 3:479–511, 2011.

[CL05] Keke Chen, Ling Liu. Privacy preserving data classification with rotation perturbation. In *Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2005),* 589–592, 2005.

[CS14] K. Crawford, J. Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review,* 55: 93-128, 2014.

[CSL07] K. Chen, G. Sun, L.Liu. Towards attack-resilient geometric data perturbation. In *Proc. of the Seventh SIAM Int. Conf. on Data Mining,* 78–89, 2007.

[CW10] T. Calders, S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[D06] C. Dwork. Differential Privacy. Proc. of the 33rd International Colloquium (ICALP), 1-12, 2006.

[DHPRZ12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. Proc. of the Int. Conf. on Innovations in Theoretical Computer Science (ITCS 2012), 214–226, 2012.

[Dia16] N. Diakopoulos. Accountability in Algorithmic Decision Making. *Comm. of the ACM*. Feb. 2016

[DSZ16] A. Datta, S. Sen, Y. Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. Proc. of the IEEE Symposium on Security and Privacy, 2016.

[ENI15] European Union Agency For Network And Information Security (ENISA). Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics. 2015.https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/big-data-protection/

[FFM+15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian. Certifying and Removing Disparate Impact. In Proc. of the ACM SIGKDD 2015, 259-268, ACM, 2015.

[GHL+12] J. Gehrke, M. Hay, E. Lui, and R. Pass. Crowd-Blending Privacy. Proc. of the 32nd Annual Cryptology Conference, 479-496, 2012.

[GLM+13] F. Giannotti, L.V.S. Lakshmanan, A. Monreale, D. Pedreschi, W.H. Wang. Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases. *IEEE Systems Journal* 7(3): 385-395, 2013.

[HD13] S. Hajian, J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering, 25(7):1445–1459, 2013.

[HPB+14] A. Henelius, K. Puolamäki, H. Boström, L. Asker, P. Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. Data Min. Knowl. Discov. 28 (5-6): 1503-1529, 2014.

[KAS12] T. Kamishima, S. Akaho, J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Proc. of the Eur. Conf. on Machine Learning and on Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2012), volume 7524 of LNCS, 35–50. Springer, 2012.

[KC12] F. Kamiran, T. Calders. Data preprocessing techniques for classification without discrimination. Knowledge and Inf. Systems, 33(1):1–33, 2012.

[KCM10] F. Kamiran, T. Calders, M. Pechenizkiy. Discrimination aware decision tree learning. In Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2010), 869–874. IEEE Computer Society, 2010.

[KKZ12] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. Proc. of the IEEE Int. Conf. on Data Mining (ICDM 2012), 924–929, 2012.

[KZC13] F. Kamiran, I. Zliobaite, T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and Inf. Systems, 35(3):613–644, 2013.

[L13] K.-P. Lin. Privacy-preserving kernel k-means outsourcing with randomized kernels. In 13th IEEE International Conference on Data Mining Workshops, 860–866, 2013.

[LCC15] K.-P. Lin, Y.-W. Chang, Ming-Syan Chen. Secure support vector machines outsourcing with random linear transformation. *Knowledge and Information Systems*, 44(1):147–176, 2015.

[LRT11] B. T. Luong, S. Ruggieri, F. Turini. k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. 17th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011): 502-510. ACM, 2011.

[LXG+12] Q. Lu, Y. Xiong, X. Gong, W. Huang. Secure collaborative outsourced data mining with multi-owner in cloud computing. In *IEEE 11th Int. Conf. on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 100–108, 2012.

[LW10] R. P. Lieli and H. White. The construction of empirical credit scoring rules based on maximization principles. *Journal of Econometrics*, 157(1):110–119, 2010.

[MAA+10] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, S. Wrobel. Movement Data Anonymity through Generalization. Trans. Data Privacy 3(2): 91-121, 2010.

[McS09] F. D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proc. of the 2009 ACM SIGMOD Int. Conf. on Management of data*, 19– 30, 2009.

[MC14] K. Mancuhan and C. Clifton. Combating discrimination using bayesian networks. In Artificial Intelligence and Law, 22(2), 2014.

[MCF+11] N. Mohammed, R. Chen, B. Fung, P. S Yu. Differentially private data release for data mining. In *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 493–501, 2011.

[MKA+07] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1:1, 2007.

[MK15] A. Machanavajjhala, D. Kifer. Designing statistical privacy for your data. *Communications of the ACM*, 58:3 58-67, 2015.

[MPP14] A. Monreale, D. Pedreschi, R. G. Pensa, F. Pinelli. Anonymity preserving sequential pattern mining. *Artificial Intelligence Law* 22(2): 141-173, 2014.

[MRP+14] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, D. Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 2014:10, 2014.

[MWP+13] A. Monreale, W.H. Wang, F. Pratesi, S. Rinzivillo, D. Pedreschi, G. Andrienko, N. Andrienko. Privacy-preserving Distributed Movement Data Aggregation. Proc. of the 16th AGILE Conference on Geographic Information Science,225-245, 2013.

[New08] D. M. Newman. *Sociology: Exploring the Architecture of Everyday Life*. Pine Forge Press, 2008.

[NTS10] L. Ninghui, L. Tiancheng, S. Venkatasubramanian. Closeness: A New Privacy Measure for Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*,.22 (7) 943-956, 2010.

[NTV07] L. Ninghui, L. Tiancheng, S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. Proc. of the IEEE 23rd Int. Conf. on Data Engineering, 106-115, 2007.

[PRT09] D. Pedreschi, S. Ruggieri, F. Turini. Measuring discrimination in socially-sensitive decision records. In Proc. of the SIAM Int. Conf. on Data Mining (SDM 2009), 581–592. SIAM, 2009.

[PS98] P. Samarati, L. Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. 1998. https:/epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf

[QLW08] Ling Qiu, Yingjiu Li, Xintao Wu. Protecting business intelligence and customer privacy while outsourcing data mining tasks. *Knowledge and Information Systems*, 17(1):99–120, 2008.

[RHK+14] S. Ruggieri, S. Hajian, F. Kamiran, and X. Zhang. Anti-discrimination Analysis Using Privacy Attack Strategies. Proc. of ECML-PKDD 2014, Part II: 694-710. Vol. 8725 of LNCS, Springer, 2014.

[RPT10] S. Ruggieri, D. Pedreschi, F. Turini. Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data. Vol. 4, Issue 2, May 2010, Article 9.

[RR14] A. Romei, S. Ruggieri. A multidisciplinary survey on discrimination analysis. The Knowledge Eng. Review. 29 (5): 582-638, 2014.

[RRT13] A. Romei, S. Ruggieri, F. Turini. Discrimination discovery in scientific project evaluation: A case study. Expert Systems with Applications 40 (15): 6064–6079, 2013.

[RSG16] M. T. Ribeiro, S. Singh, C. Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. Proc. of the ACM SIG KDD 2016, 2016.

[Rug14] S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy, 7(2): 99-129, 2014.

[SAM16] J. Stoyanovich, S. Abiteboul, G. Miklau. Data, Responsibly: Fairness, Neutrality and Transparency in Data Analysis. Tutorial at the Int. Conf. on Extending Database Technology (EDBT 2016), Mar 2016.

[TAG+15] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.P. Hubaux, M. Humbert, A. Juels, and H. Lin. Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit. arXiv preprint arXiv:1510.02377, 2015.

[THC13] C.H. Tai, J.W. Huang, M.H. Chung. Privacy preserving frequent pattern mining on multi-cloud environment. In *Proc. of Int. Symp. on Biometrics and Security Technologies,* 235–240, 2013.

[TV06] T. M. Truta, B. Vinay. Privacy Protection: p-Sensitive k-Anonymity Property. Proc. of the 22nd Int. Conf. on Data Engineering Workshops, Atlanta, 2006.

[TYC10] C.H. Tai, P.S. Yu, M.S. Chen. k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 473–482, 2010.

[VSB+13] J. Vaidya, B. Shafiq, A.Basu, Y. Hong. Differentially private naive bayes classification. In *IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 1, 571–576, 2013.

[WCH+07] W. Kit Wong, D. W. Cheung, E. Hung, B. Kao, N. Mamoulis. Security in outsourcing of association rule mining. In *Proc. of the 33rd Int. Conf. on Very Large Data Bases*, 111–122, 2007.

[WCK+09] W. K. Wong, D. Wai-lok Cheung, B. Kao, N. Mamoulis. Secure knn computation on encrypted databases. In *Proc. of the 2009 ACM SIGMOD Int. Conf. on Management of Data*, 139–152, 2009.

[WH16] Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. White House Report, May 2016.

[Zlio15] I. Zliobaitye. A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148v1, 2015. http://arxiv.org/abs/1511.00148