D5.2
Recommendations and conclusions

Ethical and Societal Implications of Data Sciences

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731873

e-SIDES – Ethical and Societal Implications of Data Sciences

Data-driven innovation is deeply transforming society and the economy. Although there are potentially enormous economic and social benefits this innovation also brings new challenges for individual and collective privacy, security, as well as democracy and participation. The main objective of the CSA e-SIDES is to complement the research on privacy-preserving big data technologies, by analyzing, mapping and clearly identifying the main societal and ethical challenges emerging from the adoption of big data technologies, conforming to the principles of responsible research and innovation; setting up and organizing a sustainable dialogue between industry, research and social actors, as well as networking with the main Research and Innovation Actions and Large Scale Pilots and other framework program projects interested in these issues. It will investigate stakeholders' concerns, and collect their input, framing these results in a clear conceptual framework showing the potential trade-offs between conflicting needs and providing a basis to validate privacy-preserving technologies. It will prepare and widely disseminate community shared conclusions and recommendations highlighting the best way to ultimately build confidence of citizens and businesses towards big data and the data economy.

D5.2 Recommendations and conclusions

Work package	WP 5 – Validation framework
Lead author	Daniel Bachlechner (Fraunhofer ISI)
Contributing authors	Michael Friedewald (Fraunhofer ISI) Karolina La Fors (Leiden University) Alan M. Sears (Leiden University) Bart Custers (Leiden University)
Internal review	Gabriella Cattaneo (IDC) Richard Stevens (IDC)
Due Date	M33 (September 2019)
Date	17 December 2019
Version	1.0
Type	Report
Dissemination level	Public

This document is Deliverable 5.2 of Work Package 5 of the e-SIDES project on Ethical and Societal Implications of Data Science. e-SIDES is an EU funded Coordination and Support Action (CSA) that complements Research and Innovation Actions (RIAs) on privacy-preserving big data technologies by exploring the societal and ethical implications of big data technologies and providing a broad basis and wider context to validate privacy-preserving technologies. All interested stakeholders are invited to look for further information about the e-SIDES results and initiatives at www.e-sides.eu.

Grant Agreement number: 731873

Executive Summary

The primary aim of this document is to translate the insights gained during the implementation of the e-SIDES project into recommendations that help bringing different groups of actors as well as the field as a whole forward. With respect to groups of actors, the recommendations target developers and operators of big data solutions, developers of privacy-preserving technologies, policy makers dealing with relevant issues and civil society (organisations).

The recommendations are motivated by a couple of success stories that show that benefitting from data and taking responsibility seriously go well together. The stories that were selected for this purpose do not only address the storage, exchange and synthetisation of personal data but also authentication and signature.

The recommendations are generally derived from the work conducted in e-SIDES. To make sure that they are related to the ongoing debate on guidelines for responsible data-driven innovation, this document includes a short review of related previous work. Neither the recommendations themselves nor the review can be considered comprehensive. Nevertheless, they address a broad range of aspects.

What makes the recommendations particularly relevant and unique is that the selection is the result of an intensive 3-year process involving numerous stakeholders with different backgrounds and assuming different roles in the big data life cycle.

Developers and operators of big data solutions are recommended to

- Comply with laws and corporate policies
- Define and fill a C-level position in charge of data
- Publish a declaration of their data ethics policies
- Maintain dialogue with other stakeholders
- Perform impact assessments
- Adhere to privacy by design principles
- Protect privacy by default
- Implement appropriate security controls
- Perform regular ethics reviews

While some of the recommendations are more relevant to developers of big data solutions, others are more relevant to operators.

Developers of privacy-preserving solutions are recommended to

- Maintain dialogue with other stakeholders
- Make the provenance of data traceable
- Support attractive business models
- Empower users through supportive interfaces

Policy makers dealing with relevant issues are recommended to

- Foster data literacy

- Require protective measures where necessary
- Enforce regulations effectively
- Promote the establishment of strong bodies of oversight
- Link co-financing to ethical behaviour
- Link public sector procurement to ethical behaviour

Civil society organisations are recommended to

- Inform individuals about risks
- Maintain dialogue with other stakeholders
- Foster adherence to professional standards and codes of conduct
- Promote the idea of a data ethics oath

Grant Agreement number: 731873

Contents

Grant Agreement number: 731873

Figures

Abbreviations

EU	European Union
PIA	Privacy impact assessment
PII	Personally identifiable information
PEP	Polymorphic encryption and pseudonymisation
RD	Responsbile Data

Grant Agreement number: 731873

1. Introduction

This section outlines the background, the methodology and the structure of this document.

1.1. Background

This report is Deliverable 5.2 of the e-SIDES project. In this project, the ethical, legal, societal and economic implications of big data applications are examined in order to complement the research on privacy-preserving big data technologies (mainly carried out by ICT-18-2016 projects) and data-driven innovation (carried out, for instance, by ICT-14-2016-2017 and ICT-15-2016-2017 projects).

This deliverable provides recommendations on how to overcome common implementation barriers. The specific requirements of developers and operators of big data solutions, developers of privacy-preserving technologies, policy makers and the civil society are taken into account.

1.2. Methodology

This deliverable is based on desk research and a reflection of previous work carried out within the scope of the project.

1.3. Structure

This deliverable is structured as follows:

- Section 1 outlines the background, the methodology and the structure of the deliverable.
- Section 2 introduces success stories showing that progress towards responsible data-driven innovation is being made.
- Section 3 describes related previous work on recommendations for responsible data-driven innovation.
- Section 4 makes recommendations based on the work carried out within the scope of the e-SIDES project.

Grant Agreement number: 731873

2. Success stories

This section tells success stories that show that benefitting from data and taking responsibility seriously go well together. The stories not only address the storage, exchange and synthetisation of personal data but also authentication and signature.

2.1. Storing personal data

There are numerous approaches regarding personal data stores. Two that come from the US and Europe, respectively, and have received quite some attention are:

- Tim Berners-Lee's project Solid[1]
- digi.me[2]

The project Solid pursues a new way to personal data stores. The project aims to change the way web applications work today by reversing the way companies use personal data. People become the servers and companies the clients. The main objectives are true data ownership and improved privacy. "Solid (derived from 'social linked data') is a proposed set of conventions and tools for building decentralized social applications based on Linked Data principles."

Users of Solid store their data in their Solid POD, which is stored at a location decided by the user. The user decides which person or app can read or write parts of the Solid POD. The Solid POD functions like a secure USB drive for the web. All the decisions on what anyone can see are under the control of the user. To succeed applications have to be built on the Solid ecosystem.

A community around MIT professor Tim Berners-Lee is working on Solid and addresses remaining challenges. The approach is not yet fully mature.

digi.me is a European counterpart to the US-based Solid. It allows giving access to personal data stored on a server in an encrypted from. Other companies can develop services on top of this effort. There are already some applications that use digi.me. digi.me promises full control over the personal data, which is strongly encrypted and then in a personal cloud of the user's choice. The user decides what apps can have access to the data. Only the user can view or share the data. Figure 1 provides an overview of how digi.me works.

[1] https://solid.mit.edu/
[2] https://digi.me/

Grant Agreement number: 731873

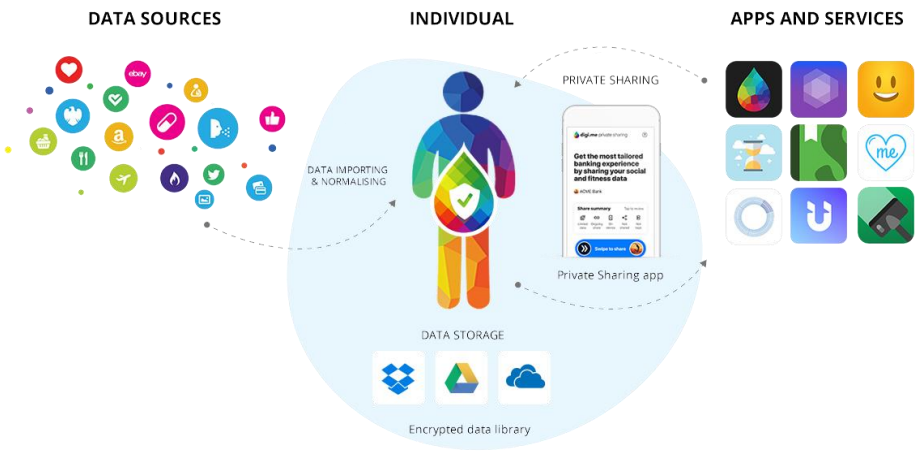

Figure 1 digi.me

digi.me is also interesting for developers and businesses. Developers can use the source code to build apps or integrate digi.me into an existing app or service. Businesses can profit from the Private Sharing platform by getting richer and more accurate data in a way that protects the privacy of the customers. digi.me is already quite mature. Companies are supporting it and APIs are available.

2.2. Exchanging personal data

Polymorphic Encryption and Pseudonymisation (PEP)[3] is a project that aims to exchange medical data for specific medical research purposes in a privacy-friendly way. To achieve this goal, PEP "combines advanced encryption with distributed pseudonymisation, and distribution of trusted data with fine-grained access management".

The motivation of the PEP project is, that especially in the medical research, big data is very important to understand diseases. However, to obtain data, the patients' privacy has to be protected. In addition, the privacy regulations of the different states or the EU have to be upheld. Therefore, the PEP project builds on the Polymorphic Encryption and Pseudonymisation technique developed by Bart Jacobs and Eric Verheul.

The key ideas of polymorphic encryption are[4]:

1. Personal data can be encrypted in a 'polymorphic' manner and stored at a central party in such a way that the central storage facility cannot get access. Crucially, there is no need to fix a priori who can decrypt the data later, so that the data can immediately be protected at the source.
2. Later on it can be decided who can de-crypt the data, via some transformation of the encrypted data (ciphtertext) which makes it locally decryptable via locally different (diversified) cryptographic keys. This decision will be made on the basis of a policy, in which the data subject should play a key role.

[3] https://pep.cs.ru.nl/
[4] https://www.cs.ru.nl/B.Jacobs/PAPERS/naw5-2017-18-3-168.pdf

Grant Agreement number: 731873

3. This transformation of encrypted data can be performed by a trusted party in a blind manner, without seeing the content; the resulting transformed ciphertext is transformed into locally decryptable ciphertext, for a specific other party.

The project is funded by public sources, the Province of Gelderland and the Radboud University. The project is developed and coordinated by professors of the Radboud Univerity in Nijmengen.

2.3. Synthesising personal data

Mostly AI[5] developed an anonymisation technique in the sense that they are synthesizing personal data. This proved to be very useful for testing algorithms and demonstrating their features without using real personal data based on machine learning and personas.

The AI-generated synthetic data retains nearly all of the valuable information while protecting personal data. This is achieved by using "deep neural networks that can automatically capture the structure and variation of an existing customer dataset"[6]. Figure 2 shows how synthetic data is generated.

In the training phase, based on machine learning, the Synthetic Data Engine automatically learns the customer behaviour. In the generation phase, synthetic customers can be generated that have the same patterns and behaviours as the actual customers. The original data is not needed anymore and therefore protected. The synthetic data is almost real data, but anonymous and can be used freely in varying areas.

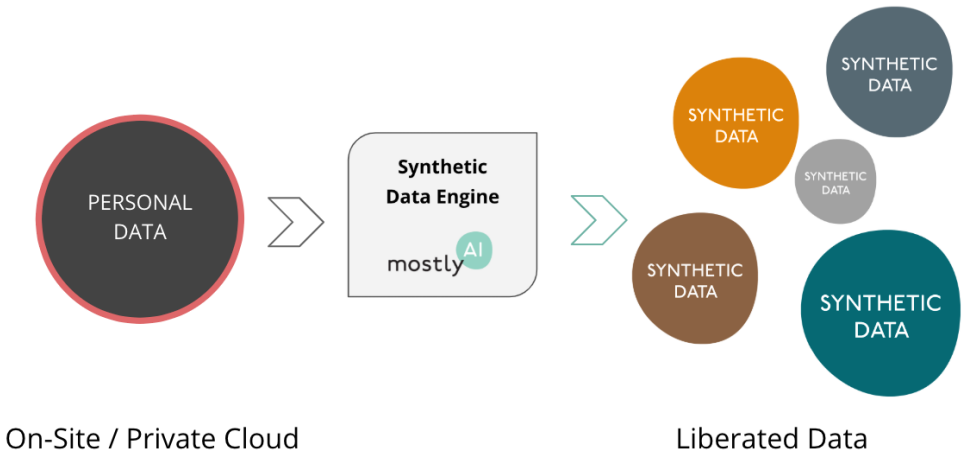

Figure 2 Generation of synthetic data

The synthetic data can be used for multiple purposes: AI training and analytics, predictive analytics and more. This can be interesting for many industries including finance, insurance or healthcare.

[5] https://mostly.ai/
[6] https://mostly.ai/finance.html

Grant Agreement number: 731873

The company publishes white papers but does not reveal how its solution really works, which impedes trust. Industrial research is making progress very quickly but both is needed industrial and academic research.

2.4. Authenticating and signing

IRMA[7] offers a privacy-friendly way of authentication that allows users to decide via app, which attributes they are willing to reveal. An example for this would be an age authentication where users only reveal their age and none of the other information. Another feature of the app is that users can sing messages where they identify themselves as the signer but also decide what attributes they want to share. IRMA furthermore offers the integration of its software into websites where user information need to be asked.

The main reason behind the development of the IRMA app is, that people often have to confirm online a special information or attribute, but often are asked for more information than needed. IRMA aims to protect the privacy of the user by only revealing the essential information in an attribute based way, which identifies people as members of certain groups (e.g., age group, student, nationality). By using the IRMA app it is not who somebody is, but what they are and therefore empowering the user. The aim is data minimisation, control and transparency for the user and the prevention of profiling.

The difference to other identity management systems is that IRMA uses a decentralised architecture and that data is only stored locally. Other identity management systems collect more information about the user and that information can be used to profile users. Figure 3 provides an overview of how IRMA works.

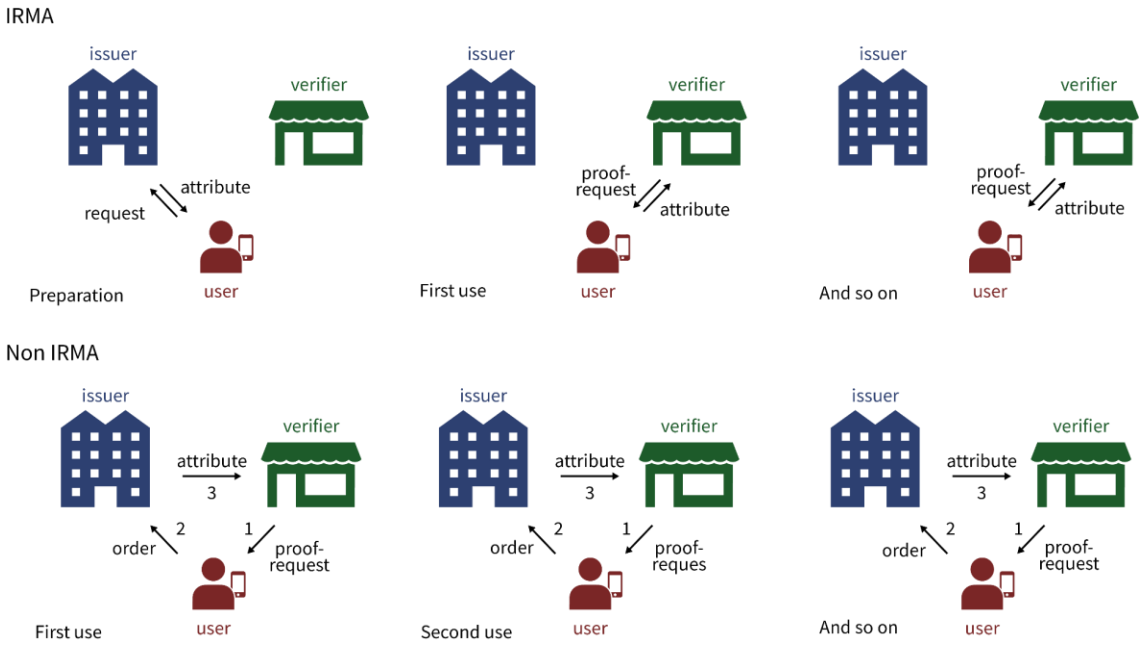

Figure 3 IRMA

The concept of the IRMA app was developed over 12 years.

[7] https://privacybydesign.foundation/irma-en/

3. Related previous work

This section provides an overview of related previous work.

3.1. RD 101: Responsible data principles

The Responsible Data (RD) community develops practical ways to deal with the unintended consequences of using data, establishes best practices and shares approaches between leaders from different sectors. RD outlines the collective duty to prioritise and respond to the ethical, legal, social and privacy-related challenges that come from using data in new and different ways. The community published a list of key elements of practicing RD[8]:

- **Power dynamics –** The least powerful actors in any situation are often the first to see unintended consequences of data collected about them. Processes like co-designing or ensuring that people from diverse backgrounds are involved in data collection or analysis processes can mitigate against this.
- **Diversity and bias –** Considering questions like, "who makes the decisions? What perspectives are missing? How can we include a diversity of thought and approach?" can highlight blind spots, and areas where adding additional voices would be valuable.
- **Unknown unknowns –** We can't see into the future, but we can build in checks and balances to alert us if something unexpected is happening.
- **Precautionary principle –** Just because we can use data in a certain way, doesn't necessarily mean we should. If we can't sufficiently evaluate the risk and understand the harms when handling data, then perhaps we should pause for a minute and re-evaluate what we're doing and why.
- **Thoughtful innovation –** For new ideas to have the best possible chance of succeeding – and for everyone to benefit from those new ideas and projects – innovation needs to be approached with care and thought, not just speed.
- **Holding ourselves to higher standards –** In many cases, legal and regulatory frameworks have not yet caught up to the real-world effects of data and technology. How can we push ourselves to have higher standards and to lead by example?
- **Building better behaviours –** There is no one-size-fits-all for RD. Existing culture, context and behaviours change the implications and ways in which data is used.

3.2. Recommendations from the Danish Expert Group on Data Ethics

The Danish Expert Group on Data Ethics[9], that was set up in March 2018, discusses how companies could handle the ethical challenges linked to the use of data and the new digital business models that are constantly evolving, and proposes recommendations that could contribute to the responsible and sustainable use of data in the business community. The Expert Group's ambition is to formulate recommendations that will help companies with a number of specific and actionable recommendations on data ethics, so that companies can start to tackle data ethics dilemmas. A starting point for that is the

[8] https://responsibledata.io/2018/01/24/rd-101-responsible-data-principles/

[9] "Data for the Benefit of the People: Recommendations from the Danish Expert Group on Data Ethics," (The Expert Group on Data Ethics, 2018), https://eng.em.dk/media/12209/dataethics-v2.pdf (accessed July 29, 2019)

so-called data ethics value compass that should help build qualified knowledge and insight. According to the Group, these values must act as the foundation for the design of data driven systems, are to be the foundation of new policy and possible legislation, and must be integrated into daily activity around data and use of data-driven systems.

The data ethics value compass includes a number of overarching values that should help implementing data ethics in practice:

- **Self-determination –** People must retain the most control possible over their own data.
- **Equality and fairness –** Technology must not discriminate.
- **Dignity –** Human dignity outweighs profit.
- **Progressiveness –** Societal progress in using data can be achieved through data ethical solutions.
- **Responsibility –** All sides must be responsible for the consequences of their technical solutions.
- **Diversity –** When developing technological solutions, involve as many trade groups of different genders, ages, ethnicities etc. as possible.

In addition to the data ethics value compass, the Expert Group on Data Ethics gives nine recommendations related to data ethics:

- **Council for data ethics –** The government needs to establish an independent Council for Data Ethics. The purpose of the council is to support an ongoing focus on data ethics.
- **The data ethics oath –** Company directors and employees actively address and take responsibility for questions and dilemmas around data ethics by taking a data ethics oath.
- **Dynamic toolbox –** The dynamic toolbox for data ethics should support the oath and provide tools and aids to help raise awareness and for specific activities in Danish companies.
- **Declaration of companies' data ethics policies –** Denmark should be the first country in the world to demand that its biggest companies incorporate an outline of their data ethics policies in their management reviews as part of their annual financial statement.
- **A data ethics seal –** A data ethics seal should be introduced as proof that a product meets data ethics requirements. A data ethics seal would make it easier for consumers to navigate digital products, and for companies to identify responsible partners.
- **National knowledge boost –** The knowledge and insight into data ethics issues of the general population and the business community need to be boosted so that we as a society gain a greater understanding of the opportunities and consequences of using data.
- **Denmark as a frontrunner –** Denmark should be visible in and impact European and global development in data ethics by being a frontrunner on the international scene.
- **Stimulating innovation and entrepreneurship –** Innovation and entrepreneurship with a focus on new data ethics business models are stimulated through co-financing, earmarking of funds and innovation contests.
- **Data ethics in public sector procurement –** It must be a requirement that digital solutions that are procured or developed by the public sector are data-ethical, so that the public sector drives demand for innovative and data-ethical solutions from companies.

Grant Agreement number: 731873

3.3. Platform for big data in agriculture: Responsible data guideline

The Platform for Big Data in Agriculture developed a Responsible Data Guideline[10] that focuses on managing privacy and personally identifiable information (PII) in the research data lifecycle. The guideline is premised on the following principles:

- research participants have the right to know and consent to the collection of information that can directly or potentially identify them, including for what purposes it will be used, who it will be shared with and why;
- any actual or potential harms associated with loss of privacy should be ethically acceptable, fully disclosed and should not be excessive in relation to the positive impacts of using PII;
- privacy protection and confidential handling of PII is paramount unless waived or reduced by a research participant to advance legitimate research objectives; and
- research ethics often involves consideration of principles that contain inherent tensions, such as individual privacy protection vs. societal benefit, leading to difficult decisions requiring professional judgement and which are the responsibility of individual researchers and their institutions.

The guideline is intended to assist agricultural researchers to:

- **Plan ahead –** Develop a data management plan to govern the use of PII in the research project and beyond. Consider how PII will be used, stored, published, shared, archived or discarded, and take into account any compliance requirements that may apply to the collection or handling of PII.
- **Anonymize PII or avoid its collection –** Only use PII if it is absolutely necessary to advance the legitimate research objectives of the project. You can maximise the participant's privacy and minimise your compliance burden by anonymising PII at the outset or not collecting PII in the first place.
- **Minimise PII as far as possible –** Minimise PII and its handling to the extent absolutely necessary for a research project.
- **Do no harm –** Research ethics call for the safety of research participants and their communities to be prioritised above all other concerns. Risks associated with privacy loss and whether they are ethically acceptable should be evaluated in the context of each project. Projects which expose research participants to significant or disproportionate risk of harm should be subject to independent ethics review and approval.
- **Obtain informed consent and be as transparent as possible –** Research participants should be empowered to give, deny or revoke their consent to share PII based on a clear understanding of why, how, by whom and for how long their data will be used. Ensure consent is fully informed by being as transparent as possible regarding the research objectives; the specific purposes for which PII will be used; how PII will be protected; benefits and risks to the research participant, their community and the public at large.

[10] https://bigdata.cgiar.org/responsible-data-guidelines/

- **Handle PII confidentially, including for transfer/access by third parties –** Maximising privacy minimises risk, therefore PII should be handled confidentially unless a research participant has explicitly consented to the contrary. Ensure appropriate security controls to protect the confidentiality of PII at rest and in transit. Transfers of PII should be undertaken on a confidential basis subject to appropriate legal and technological controls, and pro-privacy analytical tools should be used whenever feasible to do so.
- **Use PII fairly –** Check to ensure your use of the data is compatible with the purpose specification and scope consented to by the research participant, including any limitations or authorisations they may have specified or should reasonably expect regarding the use of their PII.
- **Public-use datasets containing PII are the exception –** As a general rule, public datasets should be anonymised to maximise privacy and minimise risk. PII should be included only if absolutely necessary to preserve the data's analytic potential, scientific utility or benefit to the participant, subject to prior informed consent and rigorous risk assessment.
- **Archive or delete PII –** PII should be retained only for as long as is necessary to achieve a research objectives. All copies of PII should be deleted once no longer needed. However, recognising the value of certain PII to persist with associated data (e.g. geolocation), if long-term or indefinite retention is justified, this should be clearly explained and explicitly consented to by research participants, subject to appropriate privacy and data security safeguards.
- **Review regularly –** Privacy protection and ethical research standards are fast evolving to keep pace with the rapid pace of technological change driven by big data. Periodically review institutional and other compliance requirements and don't be shy in seeking support from subject matter experts at your institution. The Big Data Platform may also be able to connect you with knowledge resources or experts to help address any challenges you are facing.

3.4. Ten simple rules for responsible big data research

Zook et al.[11] propose ten simple rules for responsible big data research that should address the increasingly complex ethical issues in big data research. While the first five rules are structures around how to reduce the chance of harm resulting from big data research practices, the second five rules focus on ways researchers can contribute to building best practices that fit their disciplinary and methodological approaches.

- **Acknowledge that data are people and can do harm –** Until proven otherwise, data should be considered people, thereby acknowledging the difficulty of disassociating data from specific individuals.
- **Recognize that privacy is more than a binary value –** Privacy depends on the nature of the data, the context in which they were created and obtained, and the expectations and norms of those who are affected.
- **Guard against the reidentification of your data –** Possible vectors of reidentification in the data should be identified and minimized.

[11] Matthew Zook et al., "Ten simple rules for responsible big data research," *PLoS computational biology* 13, no. 3 (2017)

- **Practice ethical data sharing –** Data should be shared as specified in research protocols, but concerns of potential harm from informally collected big data need to be a proactively addressed.
- **Consider the strengths and limitations of your data; big does not automatically mean better –** Provenance and evolution of your data should be documented, messiness and multiple meanings should be acknowledged.
- **Debate the tough, ethical choices –** Ethical practice for big data research should be debated with peers.
- **Develop a code of conduct for your organization, research community, or industry –** Rules for responsible big data research should be developed and established.
- **Design your data and systems for auditability –** Responsible internal auditing processes flow easily into audit systems and also keep track of factors that might contribute to problematic outcomes.
- **Engage with the broader consequences of data and analysis practices –** Big data research has societal-wide effects.
- **Know when to break these rules –** Responsible big data research depends on more than meeting checklists.

3.5. Big data ethics: 4 Guidelines to follow by organisations

The founder of Datafloq, Mark van Rijmenam[12], formulated four big data ethics guidelines to be followed by organisations. These guidelines should ensure a proper usage of big data strategies. Combining and analysing the correct datasets and using it in decision-making will help to grow the organisation. Doing it the correct way will help to sustain that growth for the long term. The proposed guidelines are the following:

- **Radical transparency –** Organisations should tell their customers in real-time what sort of data they are collecting and for what they will use it. In case they want to offer a service for free, organisations should be honest and transparent about it so that users know what they are up to when using the 'free' service. If possible, they should create also a paid version of the service that does not collect any data but still allows the user to use the service provided.
- **Simplicity by design –** Users should be able to simply adjust any privacy setting and they should be able to determine what they want to share or not. This process should be simple and understandable, also for the digital immigrants.
- **Preparation and security are key –** Organisations should define what information and data they really need to do business. Moreover, they should develop a crisis strategy in case the company gets hacked and any data gets stolen.
- **Make privacy part of the DNA –** When an organisation embraces transparency, simplicity and security, its customers will embrace it. Organisations should hire a Chief Privacy Officer or a Chief Data Officer that is also responsible for data privacy and ethics. This C-level position should be accountable for whatever data is collected, stored, shared, sold or analysed.

[12] https://datafloq.com/read/big-data-ethics-4-principles-follow-organisations/221

3.6. Principles for data handling

The Internet Society[13] sets out ethical principles and recommendations to provide help concerning responsible data handling. The following guiding principles should help all data-handlers – in public and private sectors and civil society – to fulfil not just the letter of the law, but its spirit and broader intent:

1. Transparency
2. Fairness
3. Respect

Recommendations for policy makers:

- Strengthen the incentives for better practice: e.g. include responsible data-handling criteria in government procurements, provide a framework for certification schemes.
- Use the full range of applicable policy, legal and regulatory options: this includes current and strengthened privacy and data protection laws; consumer protection and competition laws; increasing accountability through a 'polluter pays' principle.

Recommendations for data handlers:

- Be custodians of data, on the individual's behalf and in their interests.
- Adopt a principle of "no surprises": minimal collection and total transparency.
- Do not use personal data out of context, or for purposes the individual would not expect or to which they have not consented: consent is no excuse for bad practice.
- Make ethical considerations explicit in your development process, so that you can show why you made the design and implementation decisions you did.
- Respect the individual's interests, time and attention.
- Build an operational culture of transparency, fairness and respect.

3.7. Universal principles for data ethics

A report by Accenture[14] discusses the dynamics involved in generating a code of ethics that could guide the profession of data science to immediately help organisations shape their own internal guidelines related to data. A broad set of principles is proposed and intended to inform the development of domain-specific codes of ethics for specific organisations or industries. These "principles for data ethics" are aimed at both data science professionals and practitioners:

- **The highest priority is to respect the persons behind the data –** potential harm should be considered, as big data can produce compelling insights about populations, but those same insights can be used to unfairly limit an individual's possibilities.
- **Attend to the downstream uses of datasets –** Correlative uses of repurposed data in research and industry represents both the greatest promise and the greatest risk posed by data analytics.

[13] https://www.internetsociety.org/wp-content/uploads/2019/06/Responsible-Data-Handling-Policy-Brief-EN.pdf
[14] https://www.accenture.com/t20160629t012639z__w__/us-en/_acnmedia/pdf-24/accenture-universal-principles-data-ethics.pdf

Grant Agreement number: 731873

- **Provenance of the data and analytical tools shapes the consequences of their use –** mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data.
- **Strive to match privacy and security safeguards with privacy and security expectations –** Designers and data professionals should give due consideration to expectations regarding the privacy and security of the data subject's data and align safeguards and expectations as much as possible.
- **Always follow the law, but understand that the law is often a minimum bar –** leaders must define their own compliance frameworks that outperform legislated requirements.
- **Be wary of collecting data just for the sake of more data –** less data may result in both better analysis and less risk.
- **Data can be a tool of inclusion and exclusion –** Data professionals should strive to mitigate the disparate impacts of their products and listen to the concerns of affected communities.
- **As much as possible, explain methods for analysis and marketing to data disclosers –** Maximizing transparency at the point of data collection can minimize risks.
- **Data scientists and practitioners should accurately represent their qualifications, limits to their expertise, adhere to professional standards, and strive for peer accountability –** Data professionals should develop practices for holding themselves and peers accountable to shared standards to enhance public and client trust.
- **Aspire to design practices that incorporate transparency, configurability, accountability, and auditability –** being aware of design practices can break down many of the practical barriers that stand in the way of shared, robust ethical standards.
- **Products and research practices should be subject to internal, and potentially external ethical review –** Internal peer-review practices can mitigate risk, and an external review board can contribute significantly to public trust.
- **Governance practices should be robust, known to all team members and reviewed regularly –** organizations engaged in data analytics require collaborative, routine and transparent practices for ethical governance.

3.8. Responsible data frameworks

The Center for Democracy and Technology[15] reviewed 18 data use frameworks and organises their principles into six common themes that exist across the frameworks:

- **Respect for individual rights and autonomy –** including concepts such as consent and access to one's personal information.
- **Fairness or justice –** as in distribution of resources, including non-discrimination.
- **Beneficence and risk-benefit assessment –** the necessity of assessing the risks and benefits of collecting or using data, including data-sensitivity.

[15] https://cdt.org/files/2018/06/2018-06-25-Responsible-Data-Frameworks-In-Their-Own-Words-FULL.pdf

- **Fair-information-practices-based privacy and data protection principles –** including data minimization, collection limitation, retention limits and disposal, de-identification, use limitation and purpose specification, data quality, privacy by design.
- **Transparency and accountability for information practices –** including compliance, oversight, and also considers third parties and collaboration partners.
- **Data/information security –** an essential component of responsible data use.

4. Recommendations

This section provides a set of recommendations.

In general, spreading privacy-preserving technologies requires the simultaneous involvement of a broad range of stakeholders. When viewing the entire data value chain, privacy preservation should be considered a shared responsibility. However, the strongest party should have the largest responsibilities.

Court cases are very effective, but it can be time-consuming to wait for judgements. Yet, their effects can be far reaching in terms of privacy.

Beyond top-down approaches, bottom-up approaches should also be embraced. For instance, citizens can demand for implementing privacy-preserving technologies in a given big data context, if the lack of these technologies negatively impacts their lives.[16]

In order to effectively stimulate the implementation of privacy-preserving technologies, business models should incorporate privacy by design as a competitive advantage.

4.1. Developers and operators of big data solutions

Recommendations for developers and operators of big data solutions:

A. Big data solutions need to **comply** with laws and corporate policies and must be **flexible** enough to adapt to changing demands. Legal compliance is a precondition for every further optimisation step regarding big data solutions. The same applies to corporate policies. Accordingly, compliance with laws and corporate policies was defined as a key requirement for the design and use of privacy-preserving big data technologies in D4.2. Without doubt, laws need to be updated regularly to meet the changing needs of the data-driven world we live in. Developers and operators of big data solutions can support policy makers in this regard by clearly stating the legal challenges they are facing. It is also important to keep in mind that in many respects, legal compliance can only be seen as a baseline. It is up to developers and operators of big data solutions to go beyond legal compliance and take aspects such as fairness and responsible innovation into account wherever useful and feasible. This has also been highlighted within the scope of the e-SIDES Community Position Paper (D5.3).
B. Operators of big data solutions should take their responsibility seriously. Therefore, the function of a **Chief Privacy Officer** or a **Chief Data Officer** should be defined and filled. A C-level position focusing on responsible use of data does not only increase the chance that relevant things are

[16] For example, see https://www.hrw.org/news/2019/10/04/governments-facebook-stop-making-encryption-easy

getting implemented but is also a strong sign to the outside world. The Community Position Paper stresses that it is essential to point out clearly who is accountable for what, particularly, when decisions are taken with the help of a big data solution.

C. Operators of big data solutions should make available a **declaration** of their data ethics policies. Lack of transparency undermines trust in big data solutions as well as in those using big data solutions. Such a declaration should detail the issues faced as well as the measures. When doing that, it seems reasonable to take sector-specific standards into account. The relevance of doing this has been emphasised by many of those who contributed to the production of the Community Position Paper. Declaring data ethics policies increases the chance that preventive measures are taken. Taking preventive measures rather than reactive ones was defined as a key requirement for the design and use of privacy-preserving big data technologies in D4.2.

D. Developers and operators of big data solutions should increase the **dialogue** and interact with other stakeholders – such as developers of privacy-preserving technologies, policymakers, and civil society organisations – in order to address concerns, improve legal compliance and to ultimately improve privacy and security measures.

E. Developers and operators should make use of legal, ethical and social **impact assessment** tools when introducing new technologies into big data environments. The Community Position Paper reveals that assessing the impact of data use is still perceived as a challenge. People often do not even know what questions to ask. Different approaches towards impact assessments and their application in different domains have been discussed in detail in D5.1. It was found that the assessment of impact often requires a combination of different assessment tools. The detailed investigation demonstrates that data-driven innovations bring along fast change in professional, public, private and commercial relationships of citizens. This requires the continuous assessment of impacts not only with respect to the privacy of individuals but also to the accountability of data sharing processes.

F. Data protection and privacy **by design** policies should be incorporated into any existing big data solution and should be a mainstay of any big data solution in development from the very beginning. Privacy by design is a core principle that most guidelines, including the ones introduced on section 3, require to be taken into account. The concept has received considerable attention in D4.1 where its relevance in the context of big data solutions is described in detail. Privacy by design is related to two of the key requirements for the design and use of privacy-preserving big data technologies specified in D4.2. It is related to the requirement to embed security and privacy features in solutions as well as the requirement to take preventive measured.

G. **By default**, a high level of privacy should be set for big data solutions. Privacy by default is often discussed together with privacy by design. Several reasons for its relevance are outlined in the Community Position Paper. One key reason is that privacy preservation is not yet the normal setting for operations. It is increasingly possible to configure big data solutions in a way that they are safe, but this requires additional effort. Another one, that is somehow related, is that individuals tend to face a cognitive overload quite quickly when they can, but also have to, control what happens with their data themselves.

H. Developers and operators of big data solutions should not only implement appropriate **security controls** but also have a contingency plan at hand. This recommendations build upon three of the four requirements for the design and use of privacy-preserving big data solutions stated in D4.2.

Grant Agreement number: 731873

Among them are quite obviously the requirements to embed security and privacy features in solutions and to take preventive measures. However, also the requirement to connect people, processes and technology is closely related. As has been discussed in detail in D3.2, neither privacy-preservation nor responsible innovation in general can be achieved by applying technology alone. Technology has to be embedded in organisational processes and structures in an ethically informed and legally compliant manner. The need for this has been clearly voiced by those contributing to the Community Position Paper.

I. Big data solutions should be subject to regular ethics **reviews**. Such reviews, ideally conducted by peers, may be internal and/or external and may result in some kind of seal or certification. The relevance of seals and certifications has been emphasised in the Community Position Paper. Many of the contributors agreed that seals and certifications should be developed with the involvement of the industry and a broad representation of the society, including independent experts but also citizens' assemblies. Certifications and seals may focus on processes, outcomes, governance models or other parts of the big data value chain that should be made transparent. Ethics or privacy seals and certifications are perceived as beneficial. However, they are complex to arrange because they need to gain general trustworthiness among the broad public including end users.

4.2. Developers of privacy-preserving technologies

Recommendations for developers of privacy-preserving technologies that may be integrated into big data solutions:

A. Just as developers and operators of big data solutions, developers of privacy-preserving technologies should engage in a **dialogue** and foster partnerships. In particular partnerships with developers and operators of big data solutions promise to hone their technologies and to make sure they are put to use. This is underlined by one of the key conclusions of the Community Position Paper, which stated that all relevant stakeholders need to be involved in decision-making processes as well as in the development and application of big data solutions. Dialogue is also promising to allow resolving or dealing with existing trade-offs between privacy, confidentiality interests and business interests. Such trade-offs have been discussed in D5.1.

B. **Provenance** is a key issue with respect to both data and big data solutions. Mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data are needed. Assessing data quality as well as the fit of a big data solution for a specific context is difficult. Yet, limiting potential biases is crucial. Therefore, and this has also been confirmed by many of the contributors to the Community Position Paper, it is important to know what kind of data is used and reused for different purposes. There is no doubt that traceability of data for users is essential.

C. Developers should strive for the highest levels of privacy and security, while also taking into account the realities of big data environments and **business models**. Putting privacy principles such as purpose limitation or data minimisation into practice may be in conflict with current or desired business models. As stated in the Community Position Paper, the majority of currently existing business models rely upon personal data as a currency in exchange for free services. These business models are doomed with respect to privacy violations. Taking privacy and transparency seriously, and making this public, however, has the potential to allow for

competitive differentiation. As documented in D4.1, business model conflicts represent a key challenge that is partly responsible for the slow deployment of privacy-preserving technologies.

D. Developers of privacy-preserving technologies should pay particular attention to the **interface** between their technologies and the user. If settings need to be adjusted manually, everything should be clear and understandable. There should be no surprises for users. User self-determination is essential. Users should retain the most control possible over their data. However, as outlined in the Community Position Paper, users quickly face a cognitive overload. Privacy by default as well as appropriate interface design can help here. In any case, users do not only have to be informed but also empowered to exercise their rights.

4.3. Policy makers dealing with relevant issues

Recommendations for policy makers dealing with relevant issues:

A. Knowledge about data and in particular knowledge related to opportunities and threats should be boosted. Data literacy is relevant for everybody in an increasingly data-driven society. Consequently, it needs to be part of **general education**. Education as a holistic approach is also considered as a beneficial method to raise awareness. This has been stressed by contributors to the Community Position Paper. As stated in D4.1, inadequate skill level as well as lacking cultural fit are key challenges slowing down the deployment of privacy-preserving technologies.

B. Apart from taking care of education, policy makers need to make sure that an appropriate regulatory framework is in place. Regulations should **promote**, and in some cases **require**, the use of privacy-preserving technologies and good security practices, while being aware of business models and the fact that they many hamper scientific progress. A critical task in the context of regulations is to make sure the rules are updated continuously. The Community Position Paper states that regulations need to be reviewed and make fit for the big data and analytics requirements of the 21st century. An interesting development is that companies increasingly differentiate by being privacy preserving. This development has received attention in D4.1.

C. Regulations should be better **enforced**. Regulations such as the GDPR already promote the use of privacy-preserving technologies and good security practices. However, there still exist gaps in the implementation of privacy-preserving technologies. Policy makers need to make sure that these gaps are further examined. Subsequently, they should remove barriers and bolster compliance with the regulations. One aspect that is very relevant in the context of the enforcement of regulations is the availability of strong bodies of oversight. Quite some contributors to the Community Position Paper emphasised that such bodies are needed in order to evaluate compliance.

D. Robust **bodies of oversight** should be established. Such bodies are needed both within organisations as well as on a national or international level. Oversight bodies need to know how to inspect big data related products, services and organisations in a manner that is accountable, particularly to the public. Moreover, they have to be able to ensure the continuity of monitoring. They could be supported by company audits and the publication of independent reviews of companies. Still, the practices of investigative journalists and members of civil society remain crucial to make sure regulations are enforced.

E. Public **co-financing** should be linked to ethical behaviour. Consequently it should be dependent on the availability of independent reviews and supported by data management plans and company declarations on ethics policies in place.
F. Data ethics should play a key role in **public sector procurement**. Creating data management plans should be mandatory. By setting an example, the public sector can encourage private actors to adopt high ethical standards. Moreover, the emphasis on ethical standards in the context of public projects can be expected to increase the general awareness of the topic. As outlined in the Community Position Paper, when conducting privacy impact assessments (PIAs) in Canada, PIAs must be attached to all data exchanged. If the assessment leads to negative results, tis will be taken into account. This mechanism is in place since the mid-1990s in Canada and the EU could requests assessments in a similar fashion with respect to public procurement.

4.4. Civil society (organisations)

Recommendations for civil society organisations and private individuals:

A. Civil society organisations should help **inform individuals** about risks and provide training to end users and data subjects so that they can better protect themselves. Raising awareness should also include awareness about biases. A system itself may not be discriminatory but it may learn to discriminate from members of society, which feed information into. The fact that this can lead to severe discriminatory is explained in the Community Position Paper. As stated in D4.1, it is essential that people are able to assess the value of their data. Still, many people attach way to little value to their data and are not aware of the risks that they face if their data is misused.
B. Civil society (organisations) should look for opportunities to **provide input to developers and operators of big data solutions** as well as policy makers to raise awareness of risks and to promote the use of privacy-preserving technologies. As pointed out by contributors to the Community Position Paper, collective platforms could facilitate the dialogue between different stakeholders. The good thing about such platforms is that they can be initiated by users themselves.
C. Civil society organisations should put emphasis on the adherence to **professional standards** and **codes of conduct**. For instance, as an alternative to free services, paid but privacy-friendly services with the same functionality should be offered. Moreover, a high level of transparency should be demanded. Responsibility and accountability should be taken seriously by actors collecting and using data. The “polluter pays” principle should be promoted.
D. Finally, the idea of a **data ethics oath** could be promoted by civil society organisations. As such an oath would allow abstaining from certain regulations, it could also be an advantage for companies.

Grant Agreement number: 731873

Bibliography

"Data for the Benefit of the People: Recommendations from the Danish Expert Group on Data Ethics.". https://eng.em.dk/media/12209/dataethics-v2.pdf (accessed July 29, 2019).

Zook, Matthew, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta P. Gangadharan, and Alyssa Goodman et al. "Ten simple rules for responsible big data research." *PLoS computational biology* 13, no. 3 (2017): e1005399.