5. Ethics and security

5.1 Ethics

The **SoBigData++** consortium is fully aware of the ethical implications of the proposed research and respects the ethical rules and standards of HORIZON 2020, and those reflected in the Charter of Fundamental Rights of the European Union. Generally speaking, the ethical, social and data protection considerations are crucial to this project and will be given all due attention.

The project will gather innovative and proactive responses to the structural problems we are seeing with social, and cultural data analytics. Examples include Online disinformation disorder, the Facebook data privacy breach, and algorithmic bias and discrimination, among others. This will ensure that our RI not only develops best practices and resources for practitioners of social and cultural data analytics, but also that it facilitates a better informed, wider engaged, and more equitable participation and impact for such practitioners. In this respect, **SoBigData++** will become an essential hub for European and international researchers working across numerous fields, reporting annually on the best-practices, emerging trends, and innovative approaches. The legal and ethical requirements to be elaborated for the collection, curation, storage, sharing, and use of personal data for scientific purposes support the protection of data subjects' rights under the current EU legislation. These legal and ethical requirements are also designed to ensure the proper professional behavior in compliance with accepted codes of conduct, current law, and upholding high moral standard of good scientific practice. As such, **SoBigData++** will pursue the EU views on Responsible Research and Innovation, the privilege of scientific research, and especially the values and norms of EU Data Protection law, inspired *to strengthen the protection of personal data as a fundamental right, combined with boosting the free flow of personal data as a common good*.

To drive the project along this path, **SoBigData++** will maintain and improve two *ethical and legal boards* (Operational and High level) with experts in IT law, IT ethics, and data protection, as well as experts in various disciplines such as Computer Science, Digital Humanities, Philosophy, Political Science and Sociologists along with external specialists and stakeholders.

All partners in **SoBigData++** will adhere to the Charter of Fundamental Rights of the European Union and to data protection legislation, as well as overarching ethical guidance such as the European Code of Conduct for Research Integrity. We also expect project researchers to adhere to the ethical commitments contained in their professional and institutional codes of conduct.

The project will undertake coordination and investigations that involve collection and sharing of personal data. The overriding principles will be a) active informed consent and b) appropriate ownership of data. The project will manage such data in accordance with the GDPR, local laws and institutional requirements. In addition, SoBigData++'s partner organisations will be responsible for ensuring all ethical principles relating to their country and institutional context are adhered to. In some cases, this may require additional ethical permission. Thus, where relevant and/or necessary, academic researchers in the Consortium will submit their particular research to their institution's research ethics committee for ethical approval.

In case of need, ethical approval in advance via the Board for Operational Ethics and Legality will be sought (see section 5.1.1). This consortium involves partners who are experienced academics and experts with a history of conducting research in the kinds of settings described in this research project. In addition, the creation of an

ethical and legally compliant infrastructure for data mining is the gist of SoBigData++ and all its proactive efforts are aimed at this goal.

While the General Data Protection Regulation (GDPR) intends to modernize the legal framework for the processing of personal data, **SoBigData++** aims at answering new questions on the scope, interpretation, and application of the GDPR along with the expanding role of AI, machine learning and data mining. Indeed, data processing via AI, big data mining, data analytics in the context of scientific research pose renewed concerns on legal issues, in particular, on the lawfulness, fairness, and transparency of algorithms and data (Art. 5(1)(a) and Art. 22 GDPR). Although **SoBigData++** does not intend to apply automated decision making or profiling, the consortium intends to elaborate on the social impact of algorithmic biases (T.10.6), explanatory AI (T8.09), and other pressing ethical and legal issues falling within the scope of **SoBigData++.** The legal task will deal with various ethical issues such as transparency, whether biases are pre-programmed, are unintendedly introduced by the algorithm, or are the result of disproportionate data. Ensuring adequate information to the data subjects (Art. 13, 14 GDPR), the exercise of data subjects' rights (Chapter III GDPR) and lawfulness of processing personal data in the context of big data scientific research are focal points the legal research taken by the consortium. The necessity of new regulations is also under debate (e.g., the requirement of algorithm inspecting authorities). Although there is no specific law (at least not in the EU) governing AI and big data, *soft law* and proclaimed principles are available and will be taken into account. Unfortunately, the particular assumptions, impact, and consequences of such rules and guidelines are unclear. In this sense, **SoBigData++** aims at taking part in the international discussion and knowledge transfer around these issues that pertain the relevant scientific communities. In that context (legally) non-binding state of the art policies, e. g. the Ethics Guidelines for Trustworthy AI recently published by the High-Level Expert Group on Artificial Intelligence (https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai) will be considered, as well as the development of a custom-tailored ethical and legal framework for **SoBigData++** (Figure 11).

The necessity of open access to data for scientific purposes from the perspective of IP rights, legal limitations and exceptions (e.g. to protect personal data) thereto will also have to be discussed. Undoubtedly, implementing novel technologies is essential in fostering their great potentials for scientific research. Nevertheless, implementing the appropriate data protection and IP legal frameworks is of utmost importance to protect the fundamental rights of the individuals concerned. Against this background, the legal and ethical framework proposed here is critical for correlating the freedom of research with fundamental rights of individuals.

Due to the nature of the proposed research, the achievement of **SoBigData++** tasks may require the processing of personal data of varying sensitivity ranging from user mobility data, clinical data, social networks, and other data available on the web. In this context, **SoBigData++** has a commitment to investigate the practical data analytics background and adopt rules so that the data mining, data sharing and the conduct of scientific research on the data is done in an ethical and data protection compliant way. The adoption of privacy-enhancing tools clearly illustrates this fundamental policy. Also, gender issues are considered proactively into the project work plan with a specific task in WP4.

As a consortium, we have a long experience in **Privacy-Preserving Data Mining** and **Privacy Enhancing Technology**, especially in the field of non-tabular data, such as Human mobility data collected through GPS devices and purchasing data. We firmly believe that **Privacy-by-Design** can be used for designing practical and impactful services in such a way that the quality of the results can coexist with high protection of personal data. In other words, we want to develop technological frameworks for countering privacy violations, without losing the benefits and advantages that come with implementing Big Data analytics technology.

Obviously **SoBigData**++ will apply the same approach to its own research and the ethical issues it raises.

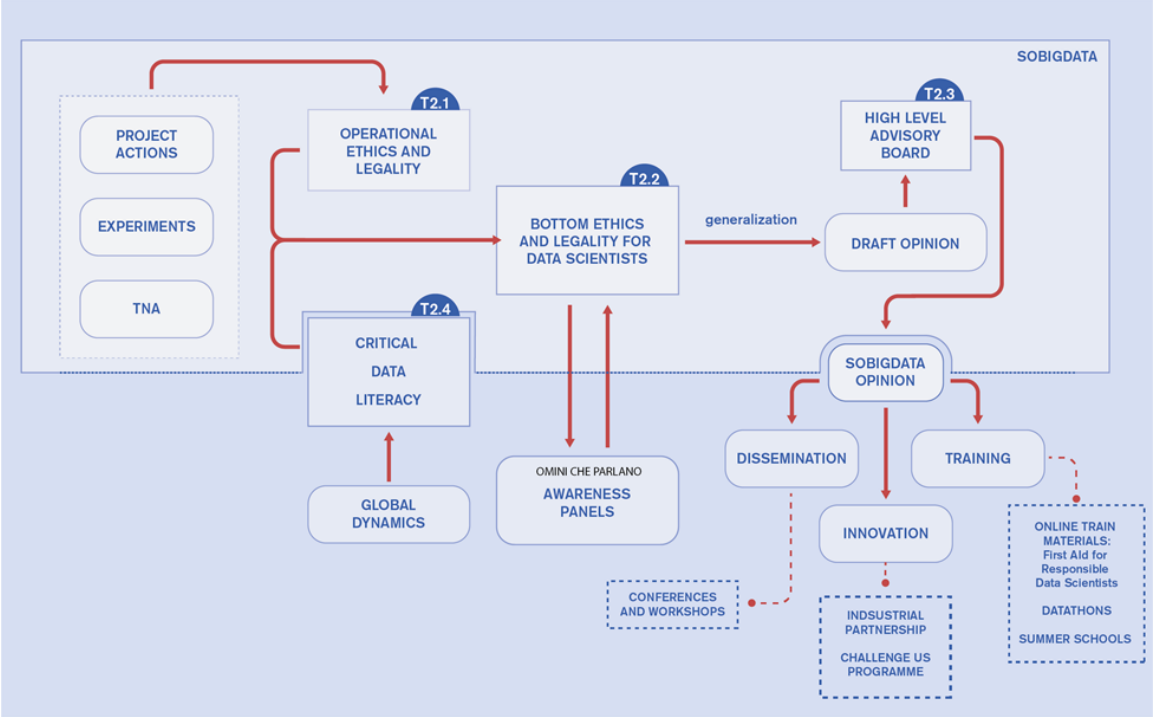

Figure 11 - SoBigData++ Ethical & privacy-enhancing tools

5.1.1 Create awareness

Privacy methodologies should be both made available in the **SoBigData**++ project, and a general explanation of the general theory behind them should be provided in the FAIR online course, as part of the dissemination activities. The FAIR course - **First AId for Responsible data scientists**, available through registration at fair.sobigdata.eu/moodle - arose as an experiment to deliver the basis of ethical data manipulation, as a general introduction of ethical concerns in data processing, the GDPR, and the Intellectual Property law. Each lesson is composed of a theoretical part, where general principles are exposed and a test part, which aims to provide a self-assessment to the comprehension of the topic and not a punitive evaluation. In less than one year of operation, more than 100 users registered at the portal and passed the core tests. Our goal for **SoBigData**++ is to spread our knowledge, expanding the course offer and increasing positively the number of persons enrolling in all the different courses. Moreover, we would like to increase the roles and duties of the Ethics Boards.

In particular, during the experience gained in SoBigData, we felt the necessity to establish the **Board for Operational Ethics and Legality** (BOEL). Indeed, even if all the various institutions composing the consortium have an official board, they are too often unknown or difficult to contact even by an internal member, or simply they meet up infrequently (possibly once or twice a year). OEB offers a fast and dynamic reference point for ethical questions. This is fundamental for researchers who want to start their experiments as soon as possible or who want to apply for the Transnational Access. *The BOEL will be active for the entire life of the project and will act as a Gateway for the data owners/managers/collectors who want to publish a*

dataset to the SoBigData Catalogue. This means that he will check the compliance with the GDPR rules. Section 5.1.5 offers a description of how it will act in the envisaged possible scenarios.

In the **SoBigData++** project, the BOEL will still offer an internal service that is able, in collaboration with all the WPs, to address ethics and legal problems in sharing data and executing social experiments on them, guaranteeing the compliance with the applicable laws and regulations. In the meanwhile, it will collect experiences in addressing problems generalizing the methodologies and making them effective in and for society. This background work sets the basis for public opinions and guidelines to be submitted to the **Board of High-Level Advisories** (HLAB, formed by members of the consortium as well as external experts selected for the specific topics addressed in each instance).

Summarizing, we believe our ethical infrastructure can qualify as a **reference point for both researchers and industries**, providing guidance and opinions, practical advice and reviewing and collecting state-of-the-art methodologies.

5.1.2 Discrimination

Fairness and related ethical issues arising in big data analytics are being recognized as an extremely relevant issue by the scientific, legal and civil-rights communities. Several recent initiatives are discussing the issue of discovering discrimination and bias, and of enforcing fairness in the design process of data-driven systems. As a consortium, we are obviously against all kind of discrimination. Moreover, we are very active in the field of discrimination discovery (both direct and indirect) as well as its mitigation. See SoBigData (D2.5 Value-Sensitive Design & Privacy-by-Design technologies for big data analytics 1) and (D2.6 Value-Sensitive Design & Privacy-by-Design technologies for big data analytics 2) respectively for a survey of the literature and a list of tools for testing/certifying fairness.

5.1.3 Intellectual property (IP)

The RI collecting and making available vast amounts of data for research brings, apart from privacy concerns, also IP issues into play. Significantly, the terms of licensing and copyrights matters are highly relevant. First, a number of datasets even though publicly available can be copyrighted, e.g. social media content, like Twitter and Facebook, Flickr images, blogs, posts, etc. Second, the use of data items, even though publicly available, is - as a rule - subject to certain license terms, such as Creative Commons (CC) licenses, etc. The reproduction, making available, modification of copyrighted content qualifies as copyright relevant actions and, in principle, require authorization by the right holder, unless exceptions apply. Against this, **SoBigData++** needs a solid IP management module to ensure the collection, use, sharing and making the data resources available for research are done in a copyright compliant way. To this end, **SoBigData++** aims to build upon the IP management module developed in SoBigData, which works on three pillars: legal, technical and educational. From the legal side, the use of data resources is governed by the RI terms of use, which refer the use of individual data items to the individual license terms. The basic rights and license terms are attached and communicated via metadata. The licensing and access restrictions are - to reasonable extent - enforced by the technical infrastructure.

Another relevant aspect concerns open access. The consortium considers that **SoBigData++** has a legal, ethical, and social responsibility with the scientific community to provide open access to highly valuable and reliable data. In this respect, **SoBigData++** embraces the FAIR principles of data management

(https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf) promoting mechanisms for finding the right data, making it accessible to interested parties, as well as advancing interoperability and reusability. **SoBigData++** is committed to upholding the standards put forward by the Open Science and Open Data movement. The project is aware of an open problem in the context of copyright management, namely how to create a fair access control in order to concede the reuse of a certain content without allowing a wide distribution to others. While this is an ongoing problem for all large open data projects, **SoBigData++** has the resources (both human and technical) to address this concern.

Finally, IP education and awareness is the third sub-set of the IP management module. It aims to ensure that the stakeholders are equipped with the required training and understanding of their responsibilities, rights, and potential infringements of IP rights. To this end, **SoBigData++** envisages a series of mechanisms for educating its members and potential researchers. Some potential ways include: (a) specific workshops designed to create awareness on the purposes of **SoBigData++** (and, indirectly, on the advantages of Open Science and Open Data, along with the FAIR principles of data); (b) online FAIR course on the platform where researchers can access and learn about the scope and limits of IP; (c) whitepapers with best practices learned in **SoBigData++** to disseminate knowledge to broad communities working in the field.

The experience gathered in SoBigData will be - to the extent possible - integrated and enhanced in **SoBigData++** with the aim to ensure IP and license compliance by knowledge and design.

5.1.4 Data management

A Data Management Plan detailing all those procedures for project data handling is already defined in SoBigData and will be updated and maintained during the project life cycle.

Informed consent procedures and information sheets will be updated with the **SoBigData++** - Data Management Plan. Furthermore, since some studies may involve minorities or vulnerable groups of people (e.g., refugees), **SoBigData++** Ethical framework will ensure not to target vulnerable individuals without taking the utmost protective measures. Among these measures are *1)* automatic triggers to initiate the ethical approval by the Board for Operational Ethics and Legality; *2)* compulsory online training for researchers on ethics and legal principles and technical measure to protect vulnerable groups from stigmatization; *3)* required confirmation by researchers to abide to the ethical principles of SoBigData++. The same protection will be taken for minor along with parents' consensus, in accordance also with CNR's Institutional Child protection policy for research. Experiments and analytical processes that involve vulnerable groups require the ethical approval in advance via the Board for Operational Ethics and Legality, that will prescribe eventual further safeguards for each Experiment or analytical process when considered appropriate. The Data Management Plan provide a set of tags in order to discover automatically the accesses to data sets that include vulnerable groups for starting the ethical approval as reported in Section 5.1.

The plan already includes the directive *"How to complete your ethics self-assessment"*. Indeed, a similar document must be filled by researchers every time a new dataset is inserted in the SoBigData catalogue. The preservation procedures shall guide the partners how to store the data, by which technology, and the period of keeping the data available. The monitoring shall ensure that the partners continue to comply with the privacy and licensing restrictions declared for their data and will take care of the costs associated for their long-term preservation. The registry of datasets will follow a defined specification.

Moreover, other two measures that help researchers in achieving awareness and transparency were developed: the *Self-Awareness* and the *Public Information Sheet.* The first corresponds to a questionnaire for the user in order to highlight possible ethical and legal issue by a set of questions, the second are use cases suggesting how to handle specific kind of data (e.g. twitter data) (see deliverables D2.3 and D2.4 in http://project.sobigdata.eu/material).

5.1.5 Protection of personal data

In **SoBigData++**, by default, all data will be at least pseudonymised **(see Section 5.1.5.4).** For any data that identifies an individual, by name or other identifying feature, that individual must give informed consent to its use in advance. Individuals will retain ownership of any data that is created, directly or indirectly, by them, and have the right to inspect that data. All participants in any **SoBigData++** activity will be made aware in advance of any **SoBigData++** activity in which they are involved, the nature of that activity, the data that will be collected and/or used and how that data will be used.

As stated above all the data inside SoBigData++ are “at least pseudonymised”, so it is improbable to violate personal information also managing "special categories of data" as the ones listed in article 9 of the General Data Protection Regulation (EU) 2016/679 ("GDPR"). In any case, the data collected and processed within the work packages and activities cannot be considered as sensitive data. **SoBigData++** will not deal with data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, health or sexual life. Moreover, as indicated in Section 5.2 Security, the e-infrastructure ensures privacy and data integrity between two communicating computer applications. Finally, all the data processing activities conducted in the **SoBigData++** project will comply with the requirements of the GDPR and other relevant national laws and regulations. By the way, in some specific (rare) situations, where could happen to manage not anonymized sensitive data a consultation with the Board for Operational Ethics and Legality will be open and a justification will be provided.

Furthermore, the Board for Operational Ethics and Legality will cover the following cases by supervising and producing the relative documents:

- in case personal data are transferred from the EU to a non-EU country or international organisation, confirmation that such transfers are in accordance with Chapter V of the GDPR, will be submitted as a deliverable.
- in case personal data are transferred from a non-EU country to the EU (or another third state), confirmation that such transfers comply with the laws of the country in which the data was collected must be submitted as a deliverable.
- detailed information on the informed consent procedures in regard to data processing must be submitted as a deliverable.
- templates of the informed consent forms and information sheets (in language and terms intelligible to the participants and eventually also including the statement for non-EU data transfer) must be kept on file.
- an explicit confirmation that the data used in the project is publicly available and can be freely used for the purposes of the project must be submitted as a deliverable.

All partners in the consortium leading a WP have appointed a DPO who will be consulted regularly. The leading partner (CNR) has appointed *Francesca Pratesi* as the interface between the DPO of ISTI-CNR and the BOEL and HLAB of **SoBigData++**.

5.1.5.1 Informed consent procedures

SoBigData++ mainly concerns the reuse of existing datasets. Researchers that collected the personal data and want to provide it to the **SoBigData++** community will have to make sure that the data had been collected legally and ethically in the first place. This will be achieved by going through the data protection law assessment form and will also be verified by the BOEL.

Although it is highly improbable that **SoBigData++** will collect personal data, informed consent procedures and information sheets will be developed with the Data Management Plan and kept on file. In these cases, participation in the project will be voluntary and all participants will be given the opportunity to ask questions and receive understandable answers before deciding whether to participate.

The consortium will outline the procedures for doing research with volunteers in order to ensure that risk and harm in research is minimised and to ensure adequate protection of the participants. Participants will be provided information in the form of a **participant information sheet** on how their **data will be processed,** who is the **data controller,** and the **possible risks or benefits of participation,** the contact details of the DPO. It will also include a description of the purpose of the research, who is organising and funding the research, and explanation about what will happen to the results of the research along with all the other information obligations as set out in art. 13 and 14 GDPR. It will be made clear that their data is stored and processed for the purposes of the project only. Each participant will be asked to sign an **informed consent** form to document and verify the information they have been given. Participants will have access –before, during and after the events –to staff who can advise them on any issues, questions, doubts, comments they may have on their data processing.

5.1.5.2 Data security

Partners will give special attention to the confidentiality of data storage and processing. They will commit to implement all appropriate technical and organisational measures necessary in order to protect potential personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorised disclosure or access, and against all other unlawful forms of processing, taking into account the particularity of the performed processing operations.

Any access will be granted only to authorised partners for data handling. Furthermore, access for information or data input (even change) will be also restricted only to authorised users to ensure their confidentiality and reserved only for these partners that collect and provide data. See Section 5.2 - Security.

5.1.5.3 Data minimization

A data minimization policy will be adopted to ensure that only data that are strictly necessary for running the activity will be processed. The principle applies to every processing including (to date an unpredicted and highly improbable case) when participants are requested to submit personal data. All processing activities will be documented, in compliance with the accountability requirements of the GDPR.

However, it is worth noticing that **SoBigData++** concerns reusing already existing data sets. In general the research exception in Art. 5 (1) lit. (b) GDPR applies, which states that the further processing of personal data

for research purposes is not considered incompatible with the original purpose for which the data had been collected as long as appropriate measures and safeguards for the rights of the data subjects are in place. The principle of data minimization and storage limitation also need to be seen with respect to the research goals. Art. 5 (1) (e) GDPR provides another exception for research. **SoBigData++** partners learn about these basic data protection principles and requirements while doing the FAIR MOOC module on data protection, but also the BOEL will verify the taken the measures and safeguards taken and compliance with these principles by the researcher.

5.1.5.4 Anonymization /pseudonymisation techniques

All research data available in **SoBigData++** will be pseudonymised. All research data publicly available, upon verification they are legally free to be used, will be anonymized by applying the opportune privacy-preserving method (such as differential privacy, k-anonymity, randomization, etc.) that guarantees adequate privacy protection, data utility and quality for analytical goals. Record of the performed anonymization technique will be maintained for each published dataset. Access to pseudonymised research data will be allowed only by on-site after (a) a positive evaluation for the MOOC, (b) signing a Non-Disclosure Agreement, and (c) a positive evaluation of the project from BOEL.

SoBigData++ will take guidance from the EU Opinion 05/2014 on Anonymization Techniques. Researchers within **SoBigData++** learn about the necessity of anonymization/ pseudonymisation of personal data while doing the MOOC and will also be made aware of this requirement while going through the data protection law assessment form. The BOEL, will double check whether data have been adequately de-identified.

SoBigData++ also provides a legal and ethical framework that makes available:

A **privacy risk assessment methodology** that aims at systematically evaluating the privacy risk level of any individual represented in the data

privacy-by-design methods for data mining and data analytics and Privacy Enhancing Technology for different types of data such as human mobility data collected through GPS devices and purchasing data. These techniques are based on well-known privacy models such as differential privacy, k-anonymity, randomization, etc.

5.1.6 Answers to the "How to complete your ethics self-assessment".

1. HUMAN EMBRYOS/FOETUSES: **ALL NO**.
2. HUMANS: **ALL NO.**
3. HUMAN CELLS / TISSUES: **ALL NO.**
4. PERSONAL DATA:

4.1 DOES YOUR RESEARCH INVOLVE FURTHER PROCESSING OF PREVIOUSLY COLLECTED PERSONAL DATA (INCLUDING USE OF PRE-EXISTING DATA SETS OR SOURCES, MERGING EXISTING DATA SETS)? **YES**

Some stakeholders might wish to donate datasets that have been collected in their activities. In such cases **SoBigData++** will keep trace of provenance with appropriate documentation

confirming lawful basis for the data processing, permission by the owner/manager of the data sets to store and share them and, if necessary, also the documentation about the Informed Consent. The data will be pseudoanonymized at the source.

Details of the database used or of the source of the data: The donated datasets will be stored into the **SoBigData++** platform and made accessible under the security policies defined in section 5.2 and made accessible (anonymized or pseudoanonimized, see 5.1.6d) under the authorization of the data owner.

Details of the data processing operations: We provide two ways to access the datasets in the catalogue: Virtual Access and Transnational Access. The first one is possible when the dataset is anonymized, it is not covered by Non-Disclosure Agreement, and the data subjects gave their explicit consent. In this case, the dataset can be directly accessed from the **SoBigData++** catalogue, and it can be downloaded or used online in combination with our integrated methods. In the case of the Transnational access, the users must provide a research project which is evaluated by both internal and external reviewers. Reviewers are experts of the research domain and the BOEL, which ensures that no ethical concerns can arise from the project, who suggest possible strategies or changes that the applicant can perform in order to be more reliable. If both the evaluations gave an affirmative answer, the applicant can physically visit the hosting site and access the dataset, under the scientific supervision of local researchers.

How will the rights of the research participants be safeguarded? Data controllers are provided with a data protection checklist and associated guide. The checklist is meant to help the researchers to evaluate whether the envisaged processing of personal data for scientific research purposes is in compliance with data protection law requirements. This refers also to the compliance with data subjects rights as the right to information, erasure etc. Additionally, researchers can turn to the Board for Operational Ethics and Legality (see section 5.1.1) and are also encouraged to seek advice from internal ethical boards or data protection officers.

How is all of the processed data relevant and limited to the purposes of the project ('data minimization' principle)? Explain. In general, the research exception in Art 5 (1) lit. (b) GDPR applies (see 5.1.6c).

Justification of why the research data will not be anonymized/pseudo anonymized (if relevant). The data will be pseudoanonymized at the source since fully anonymous data will prevent the research.

4.2 DOES YOUR RESEARCH INVOLVE PUBLICLY AVAILABLE DATA? **YES**

As stated in the previous question, the BOEL will guarantee that the publisher has the permission of the owner/manager of the data before inserting the data in the catalogue

4.3 IS IT PLANNED TO EXPORT PERSONAL DATA FROM THE EU TO NON-EU COUNTRIES? **POSSIBLE**

The SoBigData Platform is an on-line research infrastructure, and as said all the data is anonimyzed before the publishing in the on-line catalogue. In the rare case we publish data with personal information this means that it will be accessible also to non-EU countries and the BOEL will check the compliance to Chapter V of GDPR as well as the suitability of the informed consents described in section 5.1.5-6 (provided by the data collector, the data owner or the data manager).

4.4 IS IT PLANNED TO IMPORT PERSONAL DATA FROM NON-EU COUNTRIES INTO THE EU? **POSSIBLE**

As for the previous question the BOEL will act as safeguard for the data which will be published on the catalogue. It will ensure that the transfers comply with the laws of the country in which the data were collected based on the documents provided by the data collector, the data owner or the data manager.

5. ANIMALS: **ALL NO.**

6. THIRD COUNTRIES: **ALL NO.**

6.1 IN CASE NON-EU COUNTRIES ARE INVOLVED, DO THE RESEARCH RELATED ACTIVITIES UNDERTAKEN IN THESE COUNTRIES RAISE POTENTIAL ETHICS ISSUES? **YES**

Implementing the TA access, **SoBigData**++ may have researchers coming from non-EU countries (limited percentage according to the EU rules). As normal procedure the BOEL will evaluate the application of such users and for these particular categories of researchers will consider also the risk-benefit of the research proposed. The board will approve the safety of the research and the legality of its for all the other TA users applications. A particular attention will be reserved for any studies which will involve minorities or vulnerable group of people (e.g. refugees).

7. ENVIRONMENT & HEALTH and SAFETY: **ALL NO.**

8. DUAL USE: **ALL NO.**

9. EXCLUSIVE FOCUS ON CIVIL APPLICATIONS: **ALL NO.**

10. MISUSE: Does your research have the potential for misuse of research results? **YES**

10.1 Risk-assessment.
A wide range of precautions and countermeasures will be adopted to preserve privacy and ensure data protection, to comply with any relevant legislation and safeguard human rights and ethical principles. This will minimize the risk of resistance to the adoption of the **SoBigData**++ solutions, as a result of privacy and ethical concerns. The **SoBigData**++ platform has not been designed for causing any form of abuse, injury, or harm, neither intended nor unintended. However, harm might be possible by the use of the services provided (e.g., to single out minorities, increase a sense of discrimination, or hide facts about forms of bias). The BOEL (see section 5.1.1) acts as warrant that the risk-assessment is reliable and risk of misuse is kept to a minimum (see section 5.1.4).

11. OTHER ETHICS ISSUES: **ALL NO.**

5.2 Security

The **SoBigData** e-infrastructure will provide access to a set of services hosted by different sites in the EU.

The connection between the sites is ***secured with Transport Level Security*** (TLS) that provides communications security over the computer network. In particular, **SoBigData** e-infrastructure ensures privacy and data integrity between two communicating computer applications: any connections between a client (e.g., a web browser) and a **SoBigData** e-infrastructure server have the following properties:

The connection is *private* (or *secure*) thanks to the adoption of the symmetric cryptography to encrypt the data transmitted. The keys for this symmetric encryption are generated uniquely for each connection and are based on a shared secret negotiated at the start of the session. The server and client negotiate the details of which encryption algorithm and cryptographic keys to use before the first byte of data is transmitted. The negotiation of a shared secret is both secure (the negotiated secret is unavailable to eavesdroppers and cannot be obtained, even by an attacker who places themselves in the middle of the connection) and reliable (no attacker can modify the communications during the negotiation without being detected);

The identity of the communicating parties can be *authenticated* using public-key cryptography. This authentication can be made optional at client side, but is ensured at the server side;

The connection ensures *integrity* because each message transmitted includes a message integrity check using a message authentication code to prevent undetected loss or alteration of the data during transmission;

The connection ensures forward secrecy, ensuring that any future disclosure of encryption keys cannot be used to decrypt any TLS communications recorded in the past.

The **SoBigData** e-infrastructure provides access to services and data via ***Virtual Research Environments*** (VREs). Each VRE enables services and data exploitation to the users authorized to access the VRE.

The **SoBigData** e-infrastructure authorization is empowered by a ***token-based authorization system*** compliant with the Attribute-based access control (ABAC) that defines an access control paradigm whereby access rights are granted to users through the use of policies that are validated in a VRE context.

We do not identify other specific issues of security related to the current proposal, other than the above-mentioned provisions for secure data management of potentially sensitive or otherwise restricted access to data.

References for Ethics section:

[1] Ann Cavoukian. Privacy design principles for an integrated justice system. Working Paper, 2000. www.ipc.on.ca/index.asp?layid=86&fid1=318

[2] European Parliament & Council. General data protection regulation, 2016. L119, 4/5/2016.