Social Mining & Big Data Analytics

SoBigData

RESEARCH INFRASTRUCTURE ++

Deliverable D2.2

Ethics and Legality Framework activities 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

DOCUMENT INFORMATION

PROJECT	
PROJECT ACRONYM	SoBigData-PlusPlus
PROJECT TITLE	SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics
STARTING DATE	01/01/2020 (60 months)
ENDING DATE	31/12/2024
PROJECT WEBSITE	http://www.sobigdata.eu
TOPIC	INFRAIA-01-2018-2019 Integrating Activities for Advanced Communities
GRANT AGREEMENT N.	871042

DELIVERABLE INFORMATION	
WORK PACKAGE	WP2 NA1 - Responsible Data Science
WORK PACKAGE LEADER	TUDelft
WORK PACKAGE PARTICIPANTS	CNR, UNIPI, SSSA, KCL, LUH, CNRS, URV
DELIVERABLE NUMBER	D2.2
DELIVERABLE TITLE	Ethics and Legality Framework Activities 1
AUTHOR(S)	Juan M. Durán (TUDelft)
CONTRIBUTOR(S)	Giorgia Pozzi (TUDelft), Francesca Pratesi (UNIPI), Denise Amram (SSSA), Giulia Schnider (SSSA), Mark Coté (KCL), Josep Domingo-Ferrer (URV), Iryna Lishchuk (LUH), Francesca Donati (SSSA), Giovanni Comandé (SSSA)
EDITOR(S)	Valerio Grossi (CNR), Beatrice Rapisarda (CNR)
REVIEWER(S)	Roberto Pellungrini (UNIPI), Giovanni Comandé (SSSA)
CONTRACTUAL DELIVERY DATE	30/06/2021
ACTUAL DELIVERY DATE	02/07/2021
VERSION	V1.1
TYPE	Report
DISSEMINATION LEVEL	Public
TOTAL N. PAGES	35
KEYWORDS	Privacy, ethics, data science

EXECUTIVE SUMMARY

This deliverable provides a full report of the relevant activities carried out, ongoing, and planned during the period 2nd March 2020 - 30th June 2021 by Work Package 2: NA1 - Responsible Data Science (hereby WP2) and its Tasks. This report builds on the previous deliverable D2.1 which provided a full description of the Board of Operational Ethics and Legality submitted on the 1st of March 2020.

The document is structured as follows. Section 1 reports on the relevance of WP2 for the other work packages within the consortium SoBigData++. Section 2 reports on the general activities carried out by all the members of WP2. Subsections 2.1 to 2.4 report on each individual task's activities as well as collaborations. Section 2.5 reports on the publications by the members of WP2. Finally, the appendixes include extra details about activities mentioned in subsections 2.1 to 2.4. Concretely, appendix A reports on the statistics of TransNational Access; appendix B is a graphical interpretation and report of the survey conducted by WP2; finally, appendix C contains the white paper to be published by the High-Level Advisory Board

DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042.

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society. SoBigData++ is set to advance on such ambitious tasks thanks to SoBigData, the predecessor project that started this construction in 2015. Becoming an advanced community, SoBigData++ will strengthen its tools and services to empower researchers and innovators through a platform for the design and execution of large-scale social mining experiments.

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 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.

GLOSSARY

EU	European Union
EC	European Commission
H2020	Horizon 2020 EU Framework Programme for Research and Innovation
RI	Research Infrastructure
VA	Virtual Access
TA	Transnational Access
BOEL	Board for Operational Ethics and Legality

TABLE OF CONTENTS

1 Relevance to SoBigData++

This document provides a full report of relevant activities carried out, ongoing, and planned during the period 2nd March 2020 - 30th June 2021 by Work Package 2: NA1 - Responsible Data Science (hereby WP2) and its Tasks. This report builds on the previous deliverable D2.1 which provided a full description of the Board of Operational Ethics and Legality submitted on the 1st of March 2020.

1.1 Purpose of this document

The document corresponds to the deliverable D2.2: Ethics and Legality Framework activities 1 according to WP2. The deliverable must be a report describing activities performed in the WP2 as a whole and specifically activities carried out in the various boards and tasks.

1.2 Relevance to project objectives

This document complies with the objectives established for WP2 consisting in gathering innovative and proactive responses to structural problems currently emerging in social and cultural data analytics, such as online information disorder, the Facebook data privacy breach and algorithmic bias and discrimination. This is ensuring that the project not only develops best practices and resources for social and cultural data analytics practitioners, and it is granting a wider, informed, engaged and equitable participation and impact. To this end, the tasks related to WP2 have carried out a series of activities as described in section 2. This report follows the previous deliverable D2.1 where the BOEL and the High-Level Advisory Board have been established.

1.3 Relation to other work packages

WP2 and its tasks carried out and planned different activities independently as well as in close collaboration with several other WPs. These activities focus on emerging ethical and social concerns that transpire from the analysis and validation of usage of data mining resources. Of particular relevance are the following WPs:

- WP3: Dissemination, Impact, and Sustainability
- WP4: Training
- WP5: Accelerating Innovation
- WP6: Transnational Access
- WP7: Virtual Access
- WP8: Social Mining and Big Data Resource Integration
- WP10: Exploratories

Naturally, the BOEL as well as the High-Level Advisory Board is attending all ethical and legal consultations required within SoBigData++ by any WP and/or its members.

1.4 Structure of the document

Section 2 reports the relevant activities carried out, ongoing, and planned during the period 2nd March 2020 - 30th June 2021 by Work Package 2: NA1 - Responsible Data Science and its Tasks, while Section 3 reports some conclusions.

2 Report on WP2 activities

As Work Package 2, we carried out activities together and as individual task. In what follows, we describe the joint activities. The individual activities are described in the subsections.

As WP2, we conducted a survey among the SoBigData++ community with the purpose of identifying which ethical, legal, and societal issues were deemed to be of most importance. The purpose of the survey was also to collect enough information to base the first white paper (see activities performed by Task 2.3). The results are shown in Appendix A, as well as an interpretation of the responses. The complete survey can be found here: https://docs.google.com/forms/d/1xysI2iUkztQSXgSRaMjm4NSTy0Q7pRkeV8d_l-gy12A/prefill.

Together with Task 2.4 Critical Data Literacy we contributed to the creation of the working environment "SoBigData literacy" (https://sobigdata.d4science.org/group/sobigdataliteracy/literature), which has been developed as a part of the SoBigData catalogue. The Literacy is conceived to support the collaborative development of a curated collection of literature of interest for the SoBigData community as a whole. The collection consists of a catalogue service enacting authorized members to publish and organize the selected contents to facilitate discovery and access, promoting the exchange of relevant information among the members of the project. The topics of the literature gathered in the SoBigData Literacy are particularly broad and comprehensive since the catalogue aims at mirroring the thematic and conceptual variety present within the project, which touches upon different aspects of highly relevant issues related to, among others, big data and artificial intelligence-based developments.

On April 20th 2021 we submitted a contribution for the SoBigData++ Magazine entitled "Introducing the novel SoBigData Literacy database. Paving the way for consolidating a shared knowledge base for the SoBigData community as a whole", in which we described the contents and functions of the SBD Literacy and its relevance for the project. The contribution will be published in the upcoming Summer Edition in the next issue of the SoBigData Magazine.

Finally, we would like to make it official that Prof. Dr. Juan M. Durán, the representative for TUDelft and WP2 and Task leader, will enter paternity leave from the 1st of September to the 30th of September 2021. In his place, Ms. Giorgia Pozzi will temporarily lead the WP2 meeting in September. Ms. Pozzi is Prof. Durán's PhD student for TUDelft, and she is also a member of the SoBigData++ consortium. Despite Prof. Durán's paternity leave, he and Ms. Pozzi will be discussing the agenda for said meeting. Any pending issues from the meeting in September will be discussed in the next meeting.

In what follows, each task reports on its activities carried out during the period 2nd of March 2020 to the 30th of June 2021.

2.1 Task 2.1. Board of Operational Ethics and Legality

Task leader: TUDelft; **Participants:** LUH, CNR, SSSA

The Board of Operational Ethics and Legality was established for Deliverable 2.1, submitted on the 1st of March 2020. Since then, there are no changes to be reported.

TUDelft, as Work Package leader and as Task 2.1. leader, have the following activities to report: It promoted the Micro-Project “BOEL Works”, which is providing a recommendation service to all the members of SoBigData++. This service is exclusively focused on assisting the members' concerns regarding legal, ethical, and societal questions brought about by their research. This recommendation service could be used to seek advice on the methodological approach (references, tools, standards, and policies etc.) to be applied in several contexts of their research (e.g., for grant applications, academic essays, databases, and research outreach, workshops etc) in order to successfully deal with the ethical-legal and societal-related aspects. In this context, the BOEL provides first assistance to improve individual as well as groups awareness on the ethical, regulatory, and societal implications in data science and to facilitate an accountable problem-solving approach towards the research.

In the first year and half of the project, the BOEL received only one TransNational Access (TA) request on 18th February 2020, i.e., before the COVID-19 pandemic started. Then, we did not process any other TA applications, and this certainly affected the number of requests we received in the period. Indeed, during the SoBigData project, we established a board similar to the BOEL, and TA requests represented most parts of the total applications we received.

Nevertheless, the BOEL received a total of 13 requests from the consortium members.

As already said, one was related to TA application, while the classification of the other is as follow:

- requests are related to industrial projects, i.e., from consortium members who are also part of SMEs, and with the activity described in the request, they want to contribute to both industrial and research activities;
- 3 requests are related to research projects, i.e., activities totally carried out within the SBD++ umbrella;
- 4 requests are related to the creation or the integration in the SBD++ catalog of a new dataset;
- 2 requests are integrations of the previous requests, as, after first processing, the BOEL highlighted the lack of some information necessary to process the request correctly.

The average response time for these requests is 30 days. All the above requests (eventually after the required integration) received positive answers, except four of them. In all these cases, the problem was not referable to the lack of ethical commitment from the writer, while the writer would need a formal opinion given by an official ethical committee due to the particular nature of the involved data/participants. However, in these cases, the BOEL gave suggestions and recommendations to both add value to the work or speed up the process.

The activities reported here are aligned with, and fulfil the specification given in the description of Work Package 2 in the SBD++ consortium, as per agreement. Future activities for the BOEL include:

- increasing the number of TNA by means of promoting access to SBD++,
- organization and management of micro-projects where members of WP2 are involved, whether as leaders or as participants,
- fostering inter-WP2 collaboration (e.g., contributing to the maintenance of the literature repository managed by Task 2.4; facilitate the organization and writing for the next white paper by Task 2.3)

2.2 Task 2.2 Bottom-up Ethics and Legality for Data Science

Task leader: SSSA; **Participants**: TUDelft, CNR, URV, KCL, SSSA, LUH

SSSA's participation in WP2 activities has first of all involved participation in the SoBigData++ Kick-off event held on 19-21 February 2020. Starting from this date, the SSSA team (Giovanni Comandè, Denise Amram, Giulia Schneider) has actively taken part to monthly meetings that have been scheduled until now on 12th May 2020 14-15.30; 4th June 2020 14-15.30; 6th July 2020 14- 15.30; 3rd August 2020 15-16.30; 4rd September 2020 17-18.30; 6th October 2020 16-17.30; 29th October 2020 15.30-17.00; 23rd November 2020 14-15.30; 28th December 2020 10-11.30; 1st February 2021 14-15.30; 31 March 2021 11-12; 29 April 2021 15.30-16.30.

SSSA has also identified key challenges and topics to be addressed to the WP2 Survey on Responsible Data Science that has been circulated among the project's participants. Moreover, from July 2020 on, the SSSA team has organised four awareness panels, respectively concerning "Data Protection For Research and Statistical Purposes: Towards Legally Attentive Datathons" (22 July 2020); "Research Infrastructure Platform: Data Protection & IP Issues" (10 November 2020); "Medical Device Regulation and Digital Health: Problems and Perspectives" (15 February 2021); "Mobility Data and Ethics" (10 June 2021), in collaboration with URV . An additional Awareness Panel organised by SSSA is scheduled for the 6th July 2021 on the topic "Legal materials as Big Data: algo(Rithms) to support legal interpretation. A dialogue with data scientists".

In addition to this, on the 12th November 2020, Prof. Dr. Giovanni Comandè was speaker at the International Forum on Digital and Democracy 2020 (Satellite Session) with the Keynote "Matching Ethics and Law in AI: policy and practical implications of the Legal Legs of a 'Trustworthy' SoBigData++".

From the 7th December 2020 to the 1st March 2021 SSSA has performed, under the coordination of Dr. Amram, an internal audit activity of the SoBigData++ research infrastructure as a micro project approved by the WP2 leader and the management: it has developed an Internal Audit aimed at analysing the Online Datasets of the SoBigData++ Infrastructure and providing methodological recommendations to boost the research infrastructure potentialities within the research community. These activities are also functional to the ongoing platform review process provided under WP7.

Moreover, SSSA and CNR created synergies with other European Project sas with LEADS (Legality Attentive Data Scientists; Grant Agreement No. 956562).

In order to liaison with other EU initiatives, Prof. Comandé (SSSA), as the Italian representative on behalf of the Italian Ministry of Health for the 1MG project financed by the EC and participated by most of the Member States https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative (legal issues), is linking the SoBigData++ Ethical and legal analysis with this initiative.

Finally, CNR and UNIPI are organizing a workshop on ”Ethics and privacy of big data use for migration research”, a joint online workshop organised with the HumMingBird consortium (Enhanced Migration Measures from a Multidimensional Perspective, H2020 project - GA 870661) and the IMISCOE Meth@Mig (Methodological Approaches and Tools in Migration Research) standing committee: https://hummingbird-h2020.eu/news/news-items/call_for_papers .

All these activities have been functional to achieve the following purposes:

- to establish a fruitful interdisciplinary dialogue between data scientists and ethical-legal scholars
- to develop cross skills and competence aligned to the European Strategy of Data
- to define a more compliant environment for the SoBigData++ infrastructure
- to contribute to the debate on European data strategy and open science and disseminate its results

2.3 Task 2.3 High-Level Advisory Board

Task leader: LUH; **Participants:** TUDelft, CNR, UNIPI, URV, CNRS, SSSA

2.3.1 Kind of activities performed

Collection, advice and reporting on best practices, emerging trends, innovative approaches resulting from WP2 activities in the Project. The High-Level Advisories Board (HLA B) has been assembled, as reported in D.2.1, and carried out its activity based on awareness conferences (T.2.2) and collection of experiences accumulated by the Board of Operational Ethics and Legality (BOEL) (T.2.1). The topic for the first annual white paper 2021 has been framed by the following sources:

1. experiences of the BOEL (T.2.1);
2. activities of the WP2 in general (M3 - 18) and outcome of the survey on ethics and legality, in particular (See: Appendix B);
3. awareness panels conducted within T.2.2;
4. discussion held by the Expert Panel of the Project, incl. HLAB Members, at the "Expert Roundtable on Data and Ethics in a Post-COVID World", a parallel session by Re-Imagine Europe and SoBigData++ to the International Forum of Digital and Democracy, 10 December, 2020.

In particular, the discussion centered around the data sciences potential for innovation, policy making and society and framing the European Digital Governance Strategies around the fundamental principles of ethics. Fairness, accountability, confidentiality, transparency, respect to privacy are the aspects to consider when building the European “Trustworthy” AI model. The maximising benefits of data science should be reached via balance with the individual and collective rights, data sustainability, global connectivity and legitimate expectations of trust.

The core for the white paper 2020/2021, as agreed by the Board, is to discuss Europe's plans to profile itself as promoter of ethics and values with the aim to create a digital ecosystem of trust in which it is easy and safe for people to share their data (with focus on data reuse). The draft version of the paper is available at: https://drive.google.com/file/d/1hjRMKFpTbU3xR0TEKzelCQ2o7lLAuDBG/view?usp=sharing

The timeframe for the white paper 2021, as agreed with the Project Coordinator, is as follows:

- First Internal Version: 30 June 2021
- Publication Ready Version: 31 August 2021

2.3.2 Participants involved

Task leader: LUH; **Participants:** TUDelft, CNR, UNIPI, URV, CNRS, SSSA

The High-Level Advisory Board consists of the following members:

1) TUDelft: Jeroen van den Hoven (Chair)
2) LUH: Marc Stauch (Vice Chair)
3) SSSA: Giovanni Comandé (Vice Chair)
4) CNR: Fosca Giannotti (regular member)
5) UNIPI: Salvatore Ruggieri (regular member)
6) URV: Josep Domingo-Ferrer (regular member)
7) CNRS: Francesca Musiani (regular member)
8) Giovanni Sartor – UEI (external expert)

2.3.3 Goals of the activities

Production of annual white papers to be published and disseminated via different channels (T.3.2, T.3.4, and T.5.1) to impact European society at different levels

2.3.4 Possible sponsorships other than SoBigData++ (if applicable)

N/A

2.3.5 Outcome(s) produced

White paper 2021 on the core issue: digital ecosystem of trust where data research (primary and/ or secondary) is conducted with due respect to subjects' rights (reinforcing subjects' trust to feel free to share data against the background of secondary data use for research). (Version 30th of June 2021, See Appendix C)

2.3.6 Possible follow up activities

Annual white papers:

a. white paper 2022;
b. white paper 2022/2023;
c. white paper 2023/2024

2.3.7 Contribution to the task and WP2 in general

Annual reporting on best-practice, emerging trends, innovative approaches resulting from the Project. Advice, formation of high-level opinions on the legal and ethical matters emerging within the project.

SSSA Prof. Caterina Sganga has actively engaged in the Activities of the BOEL while Giovanni Comandé has been actively engaged in the activities of the HLAB especially discussing the first Whitepaper emerging from the SoBigData++ activities.

2.4 Task 2.4 Critical Data Literacy

Task leader: KLC; **Participants:** LUH, TUDelft, CNR, SSSA

As per Grant Agreement, Task T2.4 is focused on the monitoring of 'the evolving global developments regarding data literacy and their long-term impact on the research infrastructure'. In order to do so, consortium partners KCL, TUDelft, CNR, LUH and SSSA collaborated in the inception and creation of a Critical Data Literacy environment which has been embedded into the SoBigData Research Infrastructure.

The Data Literacy (https://sobigdata.d4science.org/group/sobigdataliteracy/literature) environment was designed as part of the SoBigData Research Infrastructure Catalogue, in order to promote its ease of use and access, as well as integration both by word search, tag search and group search. In this manner, all the content that has been selected to become part of the Critical Data Literacy environment is easily available to users. The data literacy working environment includes works on Privacy, Fairness and Justice, Legal and Ethical Concerns, Transparency and Explainability, Accountability and Responsibility and moreover includes all the Training Materials which were already part of the SoBigData catalogue as part of the e-Learning Area.

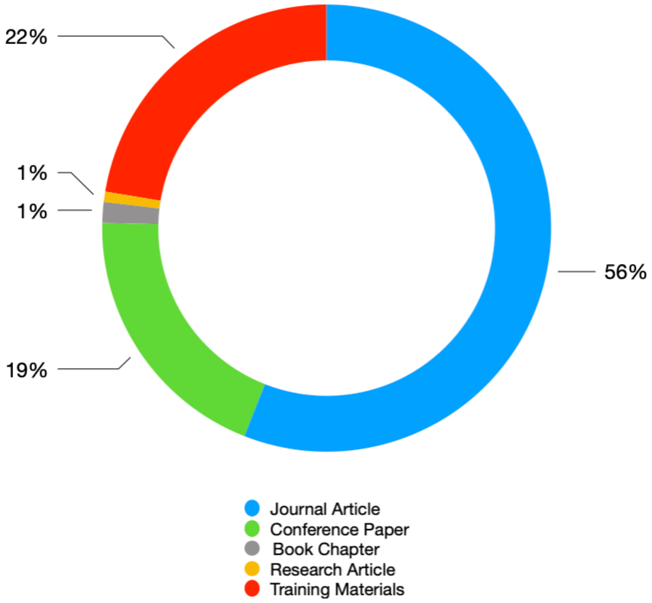


Figure 2.4.1 The composition of the Data Literacy Environment as of June 2021

The total number of data literacy sources which are currently uploaded (as of June 2021) is 134, including 30 online training materials. Figure 2.4.1 displays the variety of sources, between journal articles, conference papers, book chapters and research articles which have been selected to join the online training materials in the new Data Literacy section. Each item displays title, abstract, author and publication information, in order to aid the user in assessing if the resource might be helpful to her\him. Moreover, through a thematic clusters and tags, further connection between resources can be explored. Moreover, a contribution form has been created in order to let anyone from the SoBigData++ community suggest resources. A review process has been put in place in order to then select the relevant entries and make them become part of the Data Literacy environment, allowing it to grow in following the suggestions of the whole SoBigData++ community.

2.5 Publications by members of WP2

1. D. Amram – G. Comandé, Feedback for the EU Commission Inception Impact Assessment towards a "Proposal for a Regulation of the European Parliament and the Council laying down requirements for Artificial Intelligence" https://ec.europa.eu/info/law/better- regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and%20legal- requirements/F551050;
2. D. Amram. The Role of the GDPR in Designing the European Strategy on Artificial Intelligence: Law-Making Potentialities of a Recurrent Synecdoche. Opinio Juris in Comparatione, [S.l.], jul. 2020. ISSN 2281-5147. Available at: http://www.opiniojurisincomparatione.org/opinio/article/view/145/153;
3. G. Comandé. Unfolding the legal component of trustworthy AI: a must to avoid ethics washing. In Annuario di diritto comparato, ESI, 2020, 39-62*
(available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690633);

4. G. Schneider, Health Data Pools under European Policy and Data Protection Law: Research as a New Efficiency Defence?, in JIPITEC, 2020,1,11,1 ff.;
5. G. Schneider, Data Sharing for Collaborative Research under Art. 101 TFEU: Lessons from the Proposed Regulations for Data Markets, in European Competition Journal, 2021, DOI: 10.1080/17441056.2021.1921515;
6. G. Schneider-G. Comandè, Can the GDPR Make Data Flow for Research Easier? Yes it Can! By Differentiating!, in Computer Law & Security Review, 2021, 41 (2021) 105539.
7. G. Schneider-G. Comandè, Differential Data Protection Regimes in Data-driven Research: Why the GDPR is More Research-friendly Than You Think, in German Law Journal, 2021, forthcoming.
8. JM Durán (2021) Dissecting scientific explanation in AI (sXAI): a case for medicine and healthcare. *Artificial Intelligence*, 297 https://doi.org/10.1016/j.artint.2021.103498

3 Conclusions

This document reports the scope, activities, and members involved in the most recent activities carried out, ongoing, and planned by Work Package 2: NA1 - Responsible Data Science during the period 2nd March 2020 - 30th June 2021.

Appendix A. TransNational Access – Statistics

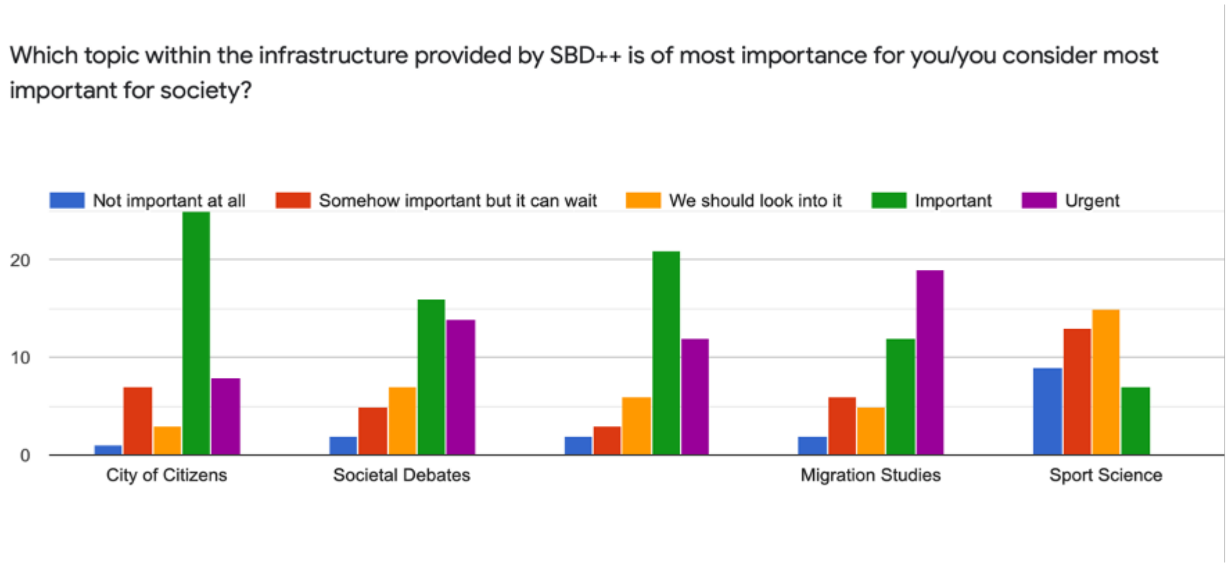
Date	Type	Answer Date	ReferenceID
2020-02-18	TA project	2020-02-25	
2020-07-27	Industrial project	2020-08-26	BOEL20200727IP
2020-08-21	Industrial project	2020-08-30	BOEL20200821IP
2020-09-01	Industrial (Integration)	2020-09-28	BOEL20200821IP
2020-10-15	Industrial project	2020-10-22	BOEL20201015IP
2020-10-23	Industrial (Integration)	2020-12-18	BOEL20201015IP
2020-11-01	Research project	2020-11-24	
2020-11-20	Research project	2020-12-18	BOEL20201120RP
2020-11-25	Dataset in SBD++ Catalog	2020-12-11	BOEL20201125D
2020-12-07	Dataset in SBD++ Catalog	2020-12-11	
2021-03-09	Research project	2021-03-15	BOEL20210309RP
2021-03-15	Research (Integration)	2021-03-22	BOEL20210309RP
2021-03-17	Dataset in SBD++ Catalog	2021-06-28	BOEL20210317D
2021-03-24	Dataset in SBD++ Catalog	2021-04-16	BOEL20210324RP

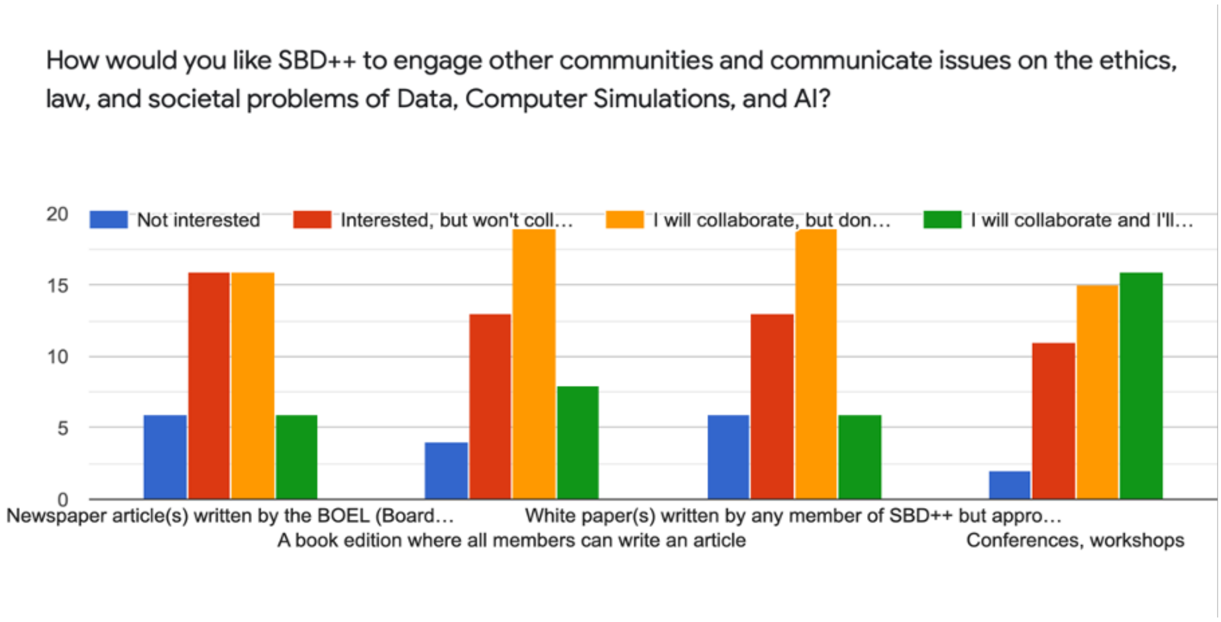
Appendix B. Graphic interpretation of the responses to the survey

Number of results: 44/250

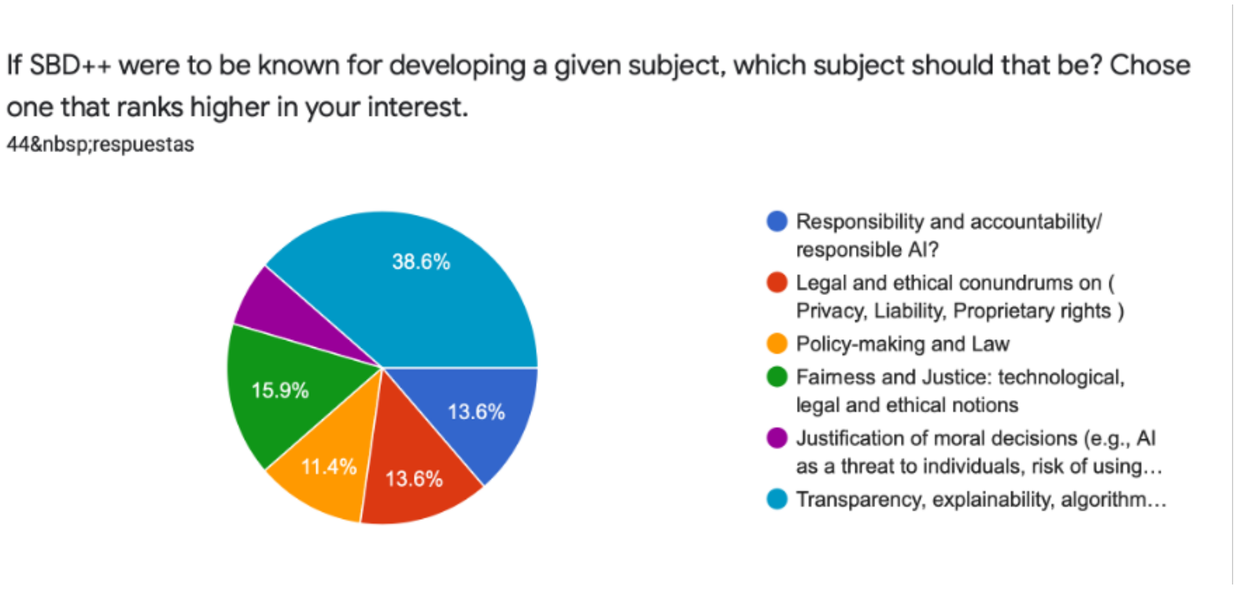
Summary of the results: The survey conducted by WP2 was successful in showing the preferences and concerns of members of SBD++ consortium, as well as provide a good sense of the priorities of research. Questions were divided into identifying the area to which the individual completing the survey belongs, and the societal, ethical, and legal concerns that this individual see as more urgent. To quantify the responses, each individual question ranked from “not important at all” to “urgent”. The topics survey include: “Transparency, explainability, algorithmic black box”, “Justification of moral decisions”, “Fairness and Justice”, “Policy-making and Law, Legal and ethical conundrums (e.g., privacy, liability, proprietary right)”, and “Responsibility”. Whereas each topic was overwhelmingly valued either as “Important” or “urgent”, the overall results show a majority of interest in ethical, societal, and legal issues related to transparency, explainability, and black-boxes. The survey also included the possibility to include a final comment, which is one of the following possibilities: “I want to add some general remarks”, “I want to tell you about past experiences”, “I want to be clear what it is in for me with the WP2 - Responsible Data Science”, “A somehow different comment”.

As mentioned, the results showed where the ethical, social, and legal concerns and interest lies among the SBD++ community. Besides providing this valuable information, the results are used as the basis for the white paper the High-Level Advisory Board will be publishing (see Appendix C).





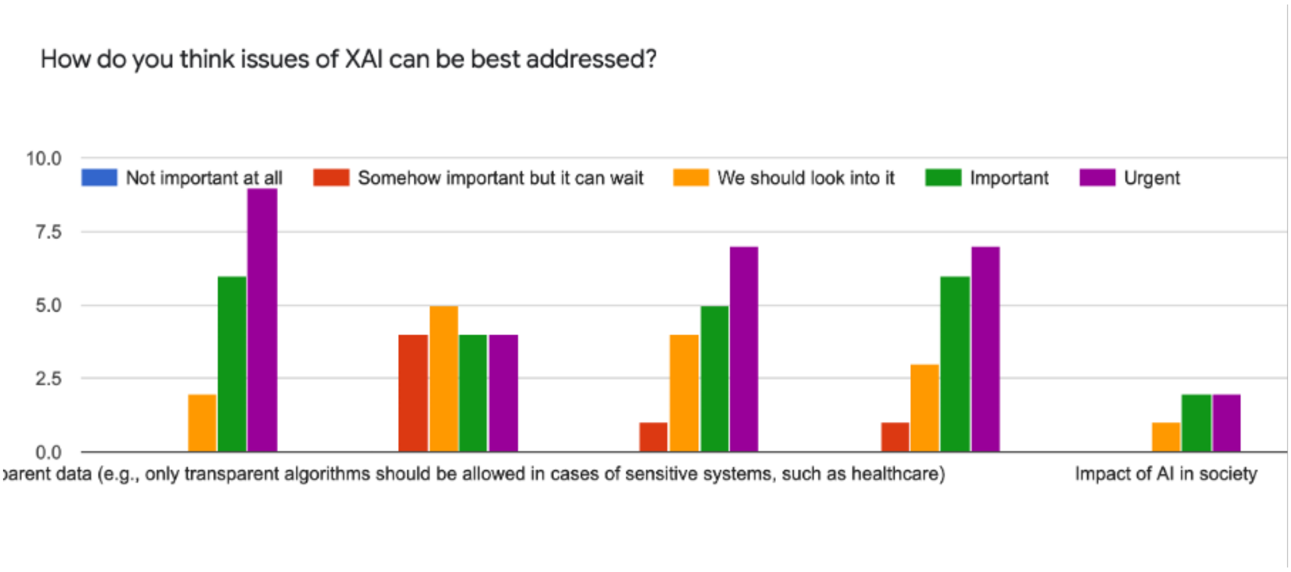
How would you like SBD++ to engage other communities and communicate issues on the ethics, law, and societal problems of Data, Computer Simulations, and AI?



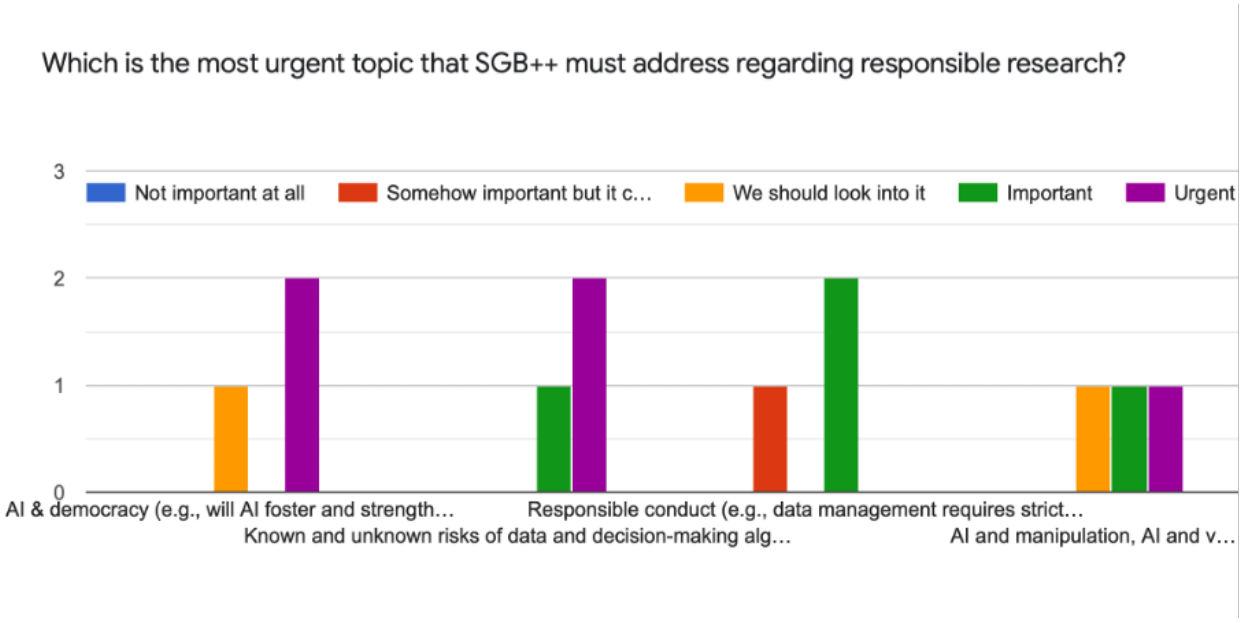
If SBD++ were to be known for developing a given subject, which subject should that be? Chose one that ranks higher in your interest.

44 respuestas

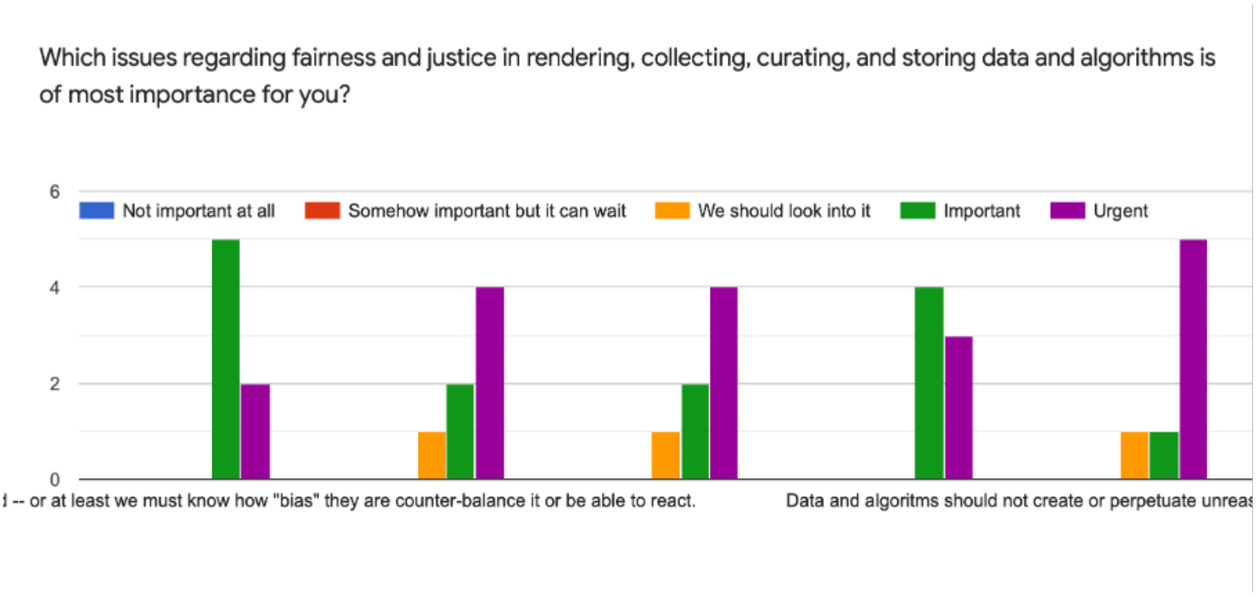
Transparency, explainability, algorithmic black box



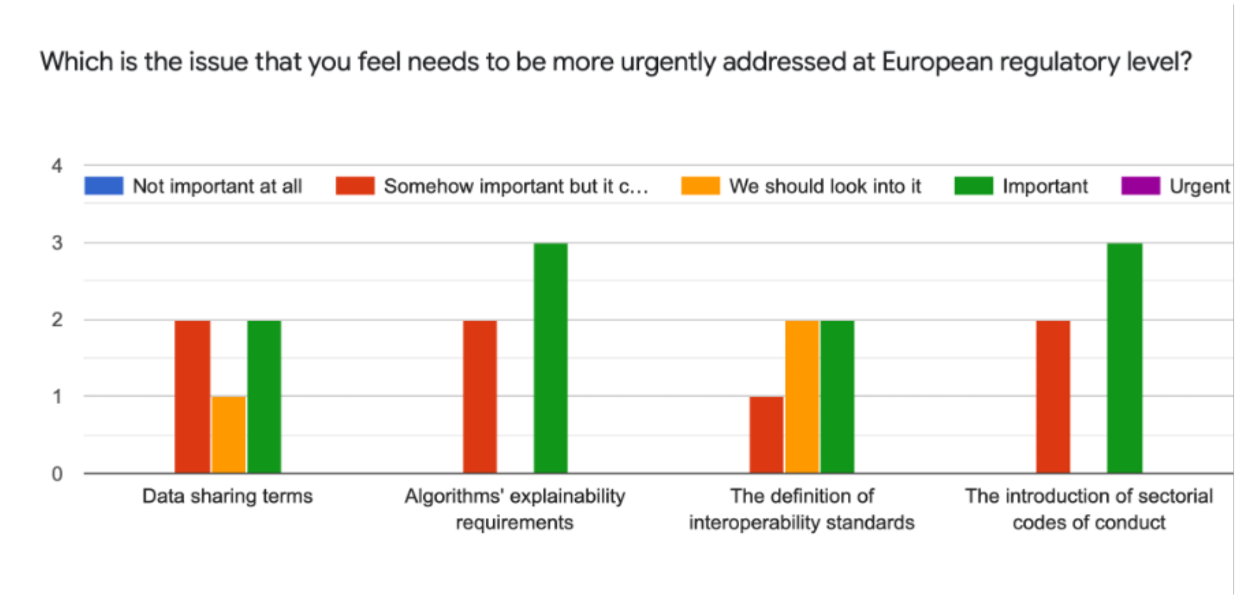
Justification of moral decisions



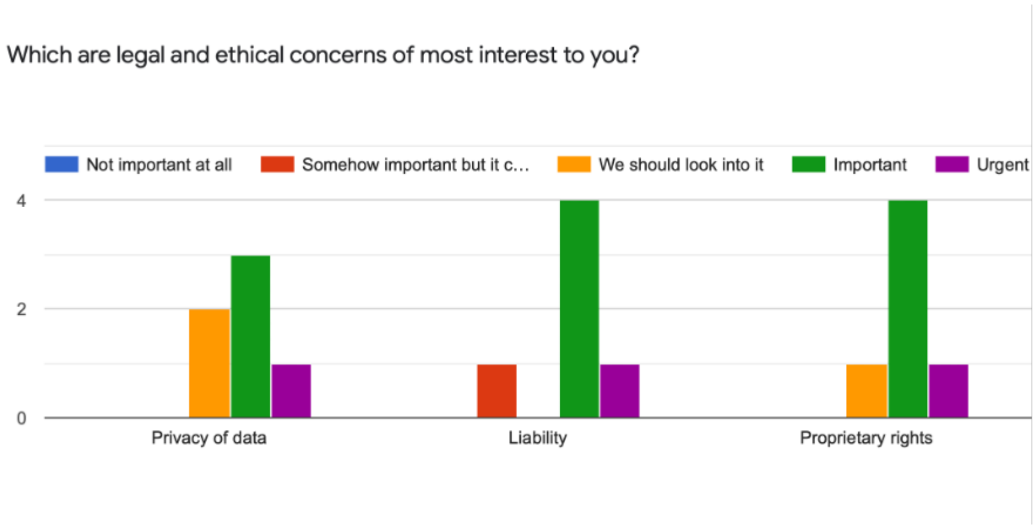
Fairness and Justice


Which issues regarding fairness and justice in rendering, collecting, curating, and storing data and algorithms is of most importance for you?

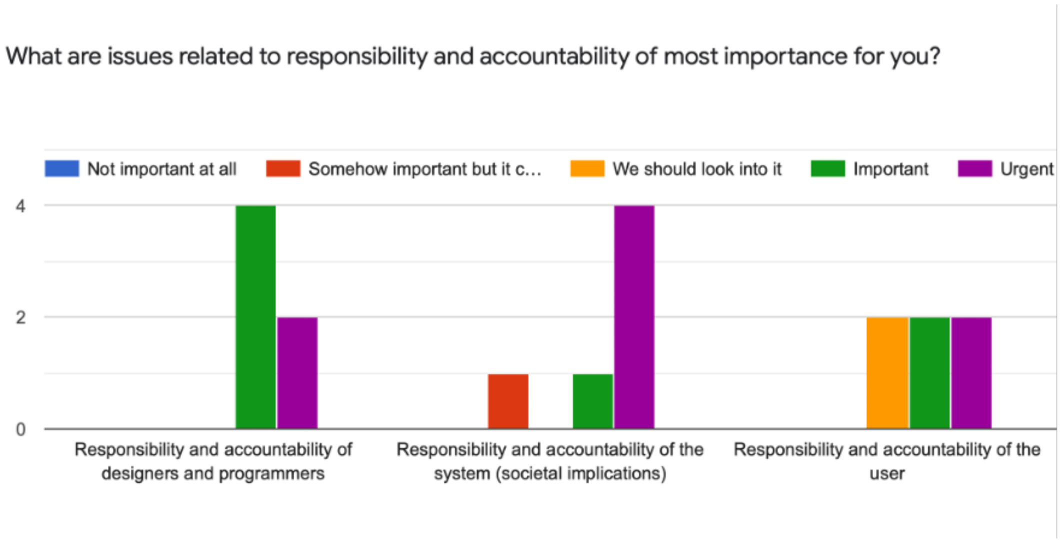
Policy Making and Law


Which is the issue that you feel needs to be more urgently addressed at European regulatory level?

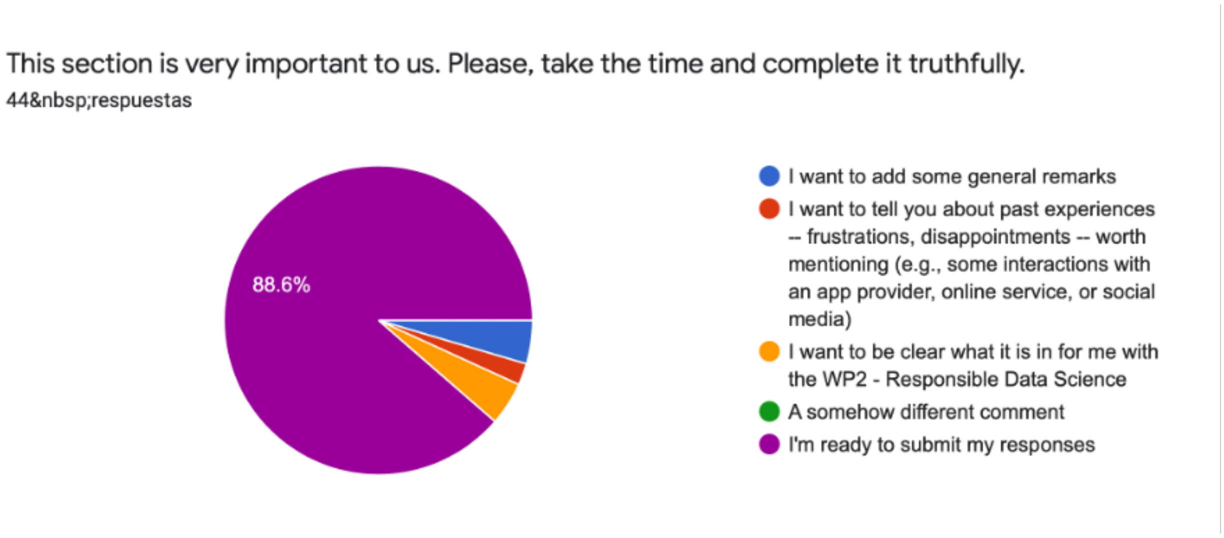
Legal and ethical conundrums



Responsibility



Final Comment



This section is very important to us. Please, take the time and complete it truthfully.

44 respuestas

Appendix C. White paper 2021

Working Title: Towards a Digital Ecosystem of Trust: Ethical, Legal and Societal Implications to Consider

Jeroen van den Hoven, Giovanni Comandé, Salvatore Ruggieri, Josep Domingo-Ferrer, Francesca Musiani, Fosca Gioannotti, Marc Stauch and Iryna Lishchuk

Abstract: The European vision of a digital ecosystem of trust rests on innovation, powerful technological solutions, comprehensive regulatory framework and respect to the core values and principles of ethics. The data science and digitalization strongly rely on data, what became obvious in view of the recent pandemic. Successful data science, especially when health data is concerned, necessitates establishing a framework where data subjects can feel safe to share their data. Use cases from research projects, methods for facilitating data-sharing, privacy-preserving technologies, de-centralization, and data altruism, interplay between the Data Governance Act and the GDPR are the main aspects this paper elaborates on.

Keywords: Digital ecosystem of trust, responsible data science, data altruism, de-centralization

1. Introduction

Europe has developed ambitious plans for its digital leadership in the remainder of the 21th century. On the basis of 2016 EU 'General Data Protection Regulation' (GDPR) and new plans for Data Governance Act and New Regulation for AI it hopes to set global standards for the Digital Age on the basis of EU law. It has rightfully foregrounded ethical principles and fundamental rights, since they are enshrined in the constitutive and binding treaties of the European Union. On this basis it aims at building a European digital ecosystem of trust and excellence that will allow the EU to make the best possible use of the potential of Digital Innovations to help solve grand societal challenges. There is however a recurrent concern in Europe itself and a point of astonishment and disbelief outside of Europe: how can one prosper in a digital economy, how can one lead in digital innovation and spearhead data driven research and AI development while being firmly committed to the highest ethical standards, especially when others don't.

This paper seeks solutions to this challenge. In doing so it draws upon the findings, results and experience in the SoBigData++ research environment, comprising over thirty research institutions, spreading through thirteen countries, united by the goal to establish a pan-European research infrastructure (RI) for social science big data. An important part of the response to the central concern resides in the fact that this approach fosters trust and augments the quality of relevant institutions. Trust and the quality of institutions is a key determinant in the success of Nations (see *Why Nations Fail,* Acemoglu & Robinson*)* is therefore a key to successful digital societies. Trust is at the same time an elusive moral concept. Trust implies the belief that the trusted are well-intentioned and are taking the moral view. Like friendship it cannot be produced at will and those who set out to 'manage' our trust in relations may find their attempts to be counterproductive. Trust usually does not appear in one's Excel sheets, but when there is no trust, the costs associated with (re-)establishing it become evident. Trust in the digital economy requires that infrastructures, institutions, mechanisms and habits are in place that allow people to receive reliable signals of the moral quality of intentions and plans of others, so that they can distinguish when trust and when distrust is appropriate in their interactions.

In a digital ecosystem of trust, appropriate norms are clear to parties, responsibilities are well defined and adequately and fairly allocated to actors and agents. Here trust, both horizontally between citizens and parties, but also trust between citizens and governments is. The SoBigData++ project provides examples of designing for trust in big data ecosystems by furthering (i) data altruism and generosity, (ii) practices of responsible data science, (iii) responsible innovations for privacy respecting technologies, (iv) research integrity review boards in AI and data driven research, (v) adequate governance schemes. In this way both primary and secondary use of data can be responsibly geared towards Big Data and Data Analytics for good and for all.

The core of the paper is to discuss Europe's plans to profile itself as promoter of ethics and values with the aim to create a digital ecosystem of trust where people feel safe to share their data and data science and analytics (either primary or secondary) is conducted with due respect to the subjects' rights. This is connected to the central questions: *What conditions need to be fulfilled for people to trust an ecosystem in which they would feel safe to share their data? How to build a research infrastructure for data science - a prototype of an ecosystem of trust?*

In order to address this central question in an innovative way, we analyse use cases from the SoBigData++ project (Section 2), privacy-preserving technologies (Section 3), the novelties of data altruism, de-centralization and data intermediaries introduced by the Data Governance Act (DGA) (Section 4), complemented by an interplay between the DGA and the GDPR, and the GDPR intricacies for research (Section 5). Conclusions marking steps forward finalize the paper.

2. Ethics integrating approaches proposed by the SoBigData++ project

Drawing upon concrete cases and examples within the SoBigData++ project - namely mechanisms in play to address data-related issues - shows what the above-mentioned aim practically means.

The SoBigData and SoBigData++ projects have developed a few vertical, thematic environments, called *exploratories*, focused on specific contexts and research questions. They are intended to test the effectiveness of the cross-disciplinary social mining research conducted on top of the SoBigData research infrastructure. The core exploratories are as follows:

- *Sustainable Cities for Citizens*: models and patterns extracted from data about cities and people living in them, allows citizens and local administrator to better understand cities and to improve services offered and overall quality of living.
- *Societal Debates and Misinformation Analysis*: the analysis of discussions on social media allows for understanding public debates and opinion, for tracking them through time and space, for investigating the spread mechanisms of misinformation and bias.
- *Demography, Economy & Finance 2.0*: data of supermarket purchases, of people's mobility, and of financial transactions, allows for investigating the changes in the well-being of people and in the network structure of companies because of the economic crisis.
- *Migration Studies*: the phenomenon of international migration is studied with models extracted from big data (mobile phone data, social media, surveys, official statistics, etc.), including economic models of migration, nowcasting migration flows and stocks, identifying perception of migration, understanding cultural diversity and integration.
- *Sports Data Science*: starting from massive data describing several sports (especially soccer, cycling and rugby) interpretable and easy-to-use models of player performances are offered to practitioners, fans, coaches, and managers.

Let us focus on *Sustainable Cities for Citizens* as a prototypical example challenging the digital ecosystem of trust offered by the research infrastructure. Data about people mobility (Andrienko et al. 2021) can be collected from mobile phones, vehicle trajectories, geolocated content uploaded to social media, travel tickets and cards, vehicle sharing services (bikes, scooters, cars, etc.), traffic volumes road sensors, video and photograph streams of security cameras, satellite images, credit card transaction data, shopping records, wi-fi connection, etc. Several useful services for the citizens and the public-policy decision makers can be designed using models built from such big data: optimizing mobility and location-based services (car sharing, tour recommendation, public transportation scheduling); supporting urban sustainability (through understanding of urban social activities); planning for different profiles of city users (residents, commuters, visitors, disabled, poor); optimizing resource distribution (residential energy management, load balancing of shared bikes).

The downside is that data collection and models/services may put the privacy of people at risk, e.g., the risk of disclosing the sensitive position of an individual. The trade-off here is to balance the utility of the discovered mobility patterns with the necessary privacy safeguards (Asikis and Pournaras 2020; Pratesi et al. 2018). Methods offered by the SoBigData platform are applicable at different stages of the data analysis process. Data can be perturbed or aggregated to obfuscate (Fiore et al. 2020) individual information. Private-by-design methods (Andrienko 2016) are offered to account for privacy risks when disclosing discovered patterns and models. Finally, privacy risk estimators support the data analyst to quantify and monitor the risk of re-identification from individual mobility patterns (Pellungrini et al. 2018) and from mobility profiles (Pratesi et al. 2020). The platform also offers general mechanisms to tag data with meta-information for ease of search, to control for accesses to data and methods, and to run methods on the cloud.

In summary, maturity of tools from the privacy-preserving literature is a prerequisite for acceptance and trust of a technology. The *Cities for Citizens* exploratory is a significant example showing how privacy of data subjects and utility of models extracted from those data can be dealt with at the same level of importance in the design of individual and society-wide data-driven services. The emergence of privacy-respecting new technologies is another proof that progress and innovation can be furthered by the consideration of ethical challenges and constraints, as explored next.

3. Privacy-respecting new technologies

It is sometimes argued that Europe's strong regulations on privacy hamper the scientific and economic progress that could ensue from massive data collection and processing (Eiss 2020; but see also Schneider and Comandè 2021b). While eliminating all barriers would no doubt facilitate progress in certain directions, the issue deserves more careful consideration.

On the one hand, unlimited collection and processing of personal data conflicts with fundamental rights and ethical values such as privacy, autonomy, fairness and security (Domingo-Ferrer and Blanco-Justicia 2020). On the other hand, the need to reconcile progress with the aforementioned rights and values spurs technology research, innovation and development (see Schneider and Comandè 2021a).

Privacy-preserving technologies are the workhorse that enforces the protection of digital assets, whether they are personal or corporate (Danezis et al. 2015). In particular, such technologies are instrumental to implement privacy and data protection by design. Due to its legal framework and its expertise in information technologies, Europe is very well placed to take the lead in innovation on privacy-preserving technologies. We next sketch directions that hold promise.

Right now, a very common setting in privacy preservation is to rely on trusted third parties (certification authorities, data controllers that take care of anonymizing or encrypting data, etc.). The trend of future information technologies is to follow the ethics-by-design approach, which inherently reduces the need for trust by empowering individual users. This is substantiated by the following design principles:

- *Decentralization*. Most individuals currently have powerful personal computing devices (smartphones, tablets, etc.). Hence, it is possible for them to carry out a fair amount of computation. This has resulted in new paradigms for decentralized machine learning (federated learning (McMahan et al. 2017), fully decentralized learning (Koloskova et al 2019), etc.), for decentralized anonymization (local anonymization (Domingo-Ferrer and Soria-Comas 2021)), for decentralized COVID-19 contact tracing, etc.

- *Incentivization*. Decentralized computing relies on the willingness of individual participants to play their respective roles as specified in the computation protocols. But this cannot be taken for granted. Without proper incentives, a rational participant might be better off by not joining, deviating, free-riding or dropping the protocol. The poor uptake of COVID-19 contact tracing apps in spite of most of them being privacy-preserving is a recent and painful example of what can happen when incentives are lacking (Toussaert 2021): people do not feel very motivated to install and run an app that can only give them negative (and maybe false) news. Offering additional services might be a better way to follow (Nanni et al. 2020).

In behavioral economics, it has long been known that moral behavior can be incentivized (Frey and Oberholzer-Gee 1997). In decentralized computing, the co-utility approach (Domingo-Ferrer et al. 2020) follows this idea by designing protocols in such a way that adhering to them is the best option for all participants: in game-theoretic terms, following a co-utile protocol as specified is an equilibrium for all participants.

Crafting decentralized, co-utile protocols allows embedding not only privacy preservation, but virtually any ethical values by design. This is open ground for the European academia and industry to conquer and cultivate. If this opportunity is properly seized, an "*IT made in Europe*" seal might become synonymous of ethically-compliant technology. Beyond a pay-off in moral and legal terms, this could also give a new purpose and competitive strength to the European IT industry.

A significant step towards decentralization of the web and de-monopolization of the data is expected to be achieved under the DGA. The DGA aimes to create regulatory framework to facilitate data sharing, *inter alia* in support of data science and open innovation, and to foster altruistic uses of the data.

4. The DGA instruments to foster data sharing and data altruism

The DGA can set a centerpiece in the EU strategy for unleashing data sharing and fostering altruistic use of personal and non-personal data. The proposed Data Governance Act introduces information intermediaries to replace big tech players, encourages 'data altruism' with citizens to facilitate data sharing, opens avenues for self-sovereign identities.

4.1 From data monopolies to data commons

In times such as these, as the world struggles with the Covid-19 pandemic and related economic and social upheavals, there seems to be momentum for action to take place in order to "reclaim" digital services and data from centralized monopolies, and for practices of "data altruism" to take place.

- In the context of smart cities and algorithmic governance, citizen data could either be managed as a commons, or handed over to private companies developing applications and controlled by centralized "control points" (DeNardis 2014). Such a dynamic raises similar questions of control and opportunities for citizen re-appropriation and governance as data commons, without exclusive intellectual property.
- Data citizens produce when using municipal digital services can be governed democratically, as urban or data commons. Such policies are needed to avoid smart cities to turn into dystopian 'safe cities' based on surveillance capitalism.
- Open data on public transportation designed as commons (Teli et al. 2015), P2P energy production within decentralized networks (Giotitsas, Pazaitis and Kostakis 2015), data generated by applications such as participatory-science Internet of Things captors to measure street pollution rates, or Internet

of Bodies-based smart devices to monitor our health signs (Matwyshyn 2019): all of these projects generate big data. If kept open, massively clustered and mined, such data can be useful for policy decision and scientific research; however, the data includes sensitive personal information, such as individual location or health status, which should be protected as such.

- Privacy and commons may intuitively seem at odds, but proposals to apply the analytical framework of knowledge commons to private data, seen as contextualized personal information flow, have been made (Sanfilippo, Frischmann and Standburg 2018).
- Aufrère and Maurel (2018) propose the linking of privacy to labor law negotiation mechanisms and social protection, in order to develop a legal framework to protect social rights on data we generate as digital labor collective rights and to exploit them as a commons.

The conceptualization of theoretical alternatives to govern algorithmic decision-making systems and the data these systems currently collect and process within closed boxes is urgently needed. Springing from such collective, inclusive, participatory models, legal and licensing framework could be developed to accompany the urban and AI data flow, governed as a privacy-friendly commons flowing through P2P infrastructure based on post-capitalist, non-proprietary values of sharing, rather than controlled by centralized organizations.

4.2 Decentralised architectures and self-sovereign identities

In Europe there have been experiments with self-sovereign identity and application of technology to re-decentralize the web. One example is Tim Berners Lee’s SOLID. Another example is Ernst Hafen’s data cooperation MIDATA. MIDATA contributed to the creation of an ecosystem of trust by way of giving patients the control of their own data. Drawing upon functional equivalent initiatives goes in the same direction.

These data practices are likely to be best supported by experiments with decentralized network architectures: a network of peers, or equals, that allows several individuals to collaborate spontaneously (and, in most cases, without the need for a central coordinating entity: Schollmeier 2001). On the basis of some technical principles (including that each node of the network can act as both a supplier and a consumer of resources; there is no central coordinating authority; and there is no entity that has a global vision of the network), philosophers and social scientists have explored these decentralized forms of organizing networks as alternative ways not only to distribute software, files and cultural works among peers, but also to manage the Internet, and to develop alternative applications, platforms, knowledge, or creations. This ensemble of models can potentially form the basis of efficiency, security and “sustainable digital development” (Linkov et al. 2018) for P2P, and of agile sharing, open access and collaborative work within the digital commons. Practical examples of these experimentations with decentralized architectures, which have originated in Europe, include Sir Tim Berners-Lee’s SOLID Web decentralization project, or the PeerTube video platform.

Individual citizen´s data stores, as has been proposed by Nanni et al. (2021) for tracking the dynamics of COVID-19, also relies on a decentralised approach. The individual citizen´s data stores have been developed to collect contact and location data of persons tested positive for COVID-19. The idea behind is to enable tracking of virus transmission chains and early detection of outbreaks in a privacy preserving manner. The conceptual advantage of decentralised approach lies in enabling sensitive categories of data to be shared separately and selectively - either with a backend system or to the other citizens - voluntarily and with a privacy preserving level of granularity. It allows for detailed information gathering for infected people, enables contact tracing, and is also scalable to large populations (See Nanni et al. 2021).

Decentralised data governance schemes strongly interrelate with self-sovereignty of the networks. The vision of self-sovereignty is beautiful. It is attractive not only for decentralised schemes, but also for big tech companies in control of data. However, in the contexts of global (e.g. COVID-19 pandemic) or pan-European actions (e.g. UEFA EURO) individual self-sovereignty can be counter-productive, unless supported by sovereignty on a geo-political level. Europe ideally positioned to push innovation forward because of data quality and diversity. A prominent example is healthcare and life sciences.

4.3 Sovereignty on a geo-political level

The COVID-19 pandemic has fueled a crisis of sovereignty. It showed "*the limits of national policy, politics and borders*," observed anthropologist Arjun Appadurai (2020). Revealing that "*all national sovereigns are weak*" it "*knocks on the door of the Westphalian model of sovereignty in a way that Ebola, SARS, and even HIV did not*." Recently, the sovereignty discourse has been mobilized in reference to the digital, acknowledging that digital infrastructure puts (national and individual) sovereignty under strain (e.g., Irion, 2012; Amoore & Raley, 2017; Couture and Toupin, 2019; Hummel et al., 2021). Ongoing EU efforts to reclaim digital sovereignty are a case in point—think of the plans for a Digital Services Act and a Digital Markets Act or the GAIA-X project, tasked with developing EU data infrastructures to counter the dominance of global tech giants. But digital infrastructure provides also a fresh terrain to exercise sovereignty—see Russia's "*sovereign Internet*" (Daucé and Musiani, 2021) or the comeback of the state in the governance of the internet (Haggart et al., 2021).

More work is needed to establish a systemic view on the distinct levels at which sovereignty is exercised, namely how the citizenry, government institutions and the private sector, evolve and interact with each other. However, it seems safe to observe as a starting point that current data infrastructure, especially all the regulatory devices based on the treatment of data that have been deployed during the pandemic, alter "the social conditions under which information on the social world is produced" (Desrosières, 1998), managed and acted upon—colliding with the state's traditional role in "making up people" (Hacking, 2004). Further, this data infrastructure contributes to enact what Isin and Ruppert (2020) called "sensory power"—a type of power based on the "the accumulation of subject peoples" by means of sensors, meaning "technologies of detecting, identifying and making people sense-able through various forms of digitized data (…) about their conduct" (2020: 2). While Europe is ideally positioned to push innovation forward in this regard, because of data quality and diversity (e.g., in healthcare and life science), it also faces unique challenges due to the particular configurations of sovereignty and data sovereignty it supports. In particular, when personal data, and especially health data - a special data category under Article 9 (1) GDPR) - are at stake, harmonization with the data protection framework is required. The task becomes even more complicated when the data critically required for pan-European actions *a priori* rests in hands of individual entities. The matter merits attention in view of highly fragmented and regulated data landscape in healthcare (bound by regulatory constraints, the obligations of professional secrecy, highly divergent data formats and encoding systems, languages, strict legitimation requirements, et cetera). A certain degree of synchronisation established between the DGA and the GDPR lay foundation for pan- European data research initiatives, as we consider next.

5. Interplay between the DGA and the GDPR

The DGA can set a centerpiece in the EU strategy for unleashing data sharing and fostering altruistic use of personal and non-personal data. In itself, it builds upon the established frameworks for research in general and avenues opened for research in support of public good (such as medical research) by the GDPR, in particular (see Schneider and Comandè 2021a).

5.1 General interconnection points

By adding a clear missing infrastructural and normative link, the trustworthy setting for intermediaries should be created allowing personal data to be used with the help of a "*personal data-sharing intermediary*". It is centered on allowing data use on altruistic grounds. From a technical point of view, its nature of a proposed EU Regulation (contrasted to a Directive) permits uniform and direct application of the many elements requiring a clear common framework. Chiefly, the DGA introduces a uniform system and interpretation of the notification for data sharing service providers, the mechanisms for data altruism, the basic principles that apply to the re-use of public sector data that cannot be available as open data or are not subject to sector-specific EU legislation, and the set-up of coordination structures at European level.

The DGA exemplifies and provides content to the so-called FAIR principles limiting the conditions for reuse "*to what is necessary to preserve the rights and interests of others in the data and the integrity of the information technology and communication systems of the public sector bodies"* (ref. 11; see also art. 5 and art. 11.4 DGA). Such a FAIR approach is made possible exactly by the GDPR regulatory background. Indeed, the very same referral 11 echoes art. 89 GDPR in its call for transmission (thus reuse) of anonymous data as a default approach, tempering the limit with the understanding that "provision of *anonymised or modified data"* might *"not respond to the needs of the re-user"* and suggesting alternative safeguards such as *"on-premise or remote re-use of the data within a secure processing environment*". A strikingly similar approach has already been experimented within the SoBigData++ project (practised as transnational access and/or virtual access).

In the same line of deference to the GDPR, the DGA leverages the principle of lawfulness of the processing to establish trust, reasserting that "*personal data should only be transmitted for re-use to a third party where a legal basis allows such transmission."* The evident preference for general interest research and data sharing of the DGA emerges in various instances. Among them is worth mentioning the possibility for public sector bodies "*to make the data available at lower or no cost, for example for certain categories of re-uses such as non-commercial re-use, or re-use by small and medium-sized enterprises".*

As a possible response to the criticisms that the GDPR might be excessively limiting reuse and personal data sharing the notion of "*Data cooperatives",* a specific category of data intermediaries includes providers of data sharing services that offer their services to data subjects in the sense of Regulation (EU) 2016/679 "*to enhance individual agency and the individuals' control over the data pertaining to them.*" By way of the intermediaries regulated in the DGA, data subjects would, for instance, be enabled to use their autonomy not only by designing wider consent (articles 6.1.a and 9.2.a GDPR) but also to make their personal data "manifestly public" (article 9.2.e GDPR) for specific purposes of general interest (see Schneider and Comandè 2021a and 2021b).

5.2 Truly enabling personal data altruism

Indeed, data altruism is a landmark for re-use of data that needed to be encouraged and leveraged within the framework of the GDPR. It is recital 35 that states "*There is a strong potential in the use of data made available voluntarily by data subjects based on their consent or, where it concerns non-personal data, made available by legal persons, for purposes of general interest.". While it stresses at the same time that "Support to scientific research, including for example technological development and demonstration, fundamental research, applied research and privately funded research, should be considered as well purposes of general interest"*.

The interplay between data altruism and the GDPR in the prism of fostering research is clearly highlighted in the DGA stressing how the intermediary tools it institutes and regulates: "*In accordance with Regulation (EU) 2016/679, scientific research purposes can be supported by consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research or only to certain* *areas of research or parts of research projects*" (ref. 38). A highly important overlap between big data science and the data protection legal framework has been achieved with regard to data (re-)processing for medical research, essentially due to the value of health as an objective of public interest (recital 53 GDPR).

5.3 Secondary use of health data for research - a legal perspective

A key area of overlap between research ethics, the science of big data mining, and legally imposed constraints under European data protection law (in the guise of the concerns the secondary use of health data for medical research. In this context, where data originally collected for one purpose, e.g. individual diagnosis or treatment, is used for another purpose (e.g. to allow a detailed comparison between the particular patient and others, to draw wider conclusions about the origins of the disease), a number of fairly stringent conditions need to be satisfied of both a legal and ethical kind. Thus, both the GDPR and normative codes of research ethics (such as the Declaration of Helsinki) often insist on the need for fresh consent from the patient to the research use. In addition, under the GDPR strict safeguards must be observed to ensure the fairness and security of the processing, including the principle of 'data minimisation', under which the data must be "*adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed*" (Art 5(1)(c)).

In this regard, it appears the GDPR may contribute rather well to the building of trust-relations as a key component, as mentioned earlier, to research involving the use of data, especially sensitive data of a medical kind. The patient subject retains control – the ability to veto the data processing (by refusing consent) as well as knowing that the researcher (data controller) owes an ongoing obligation to use the data in a fair and careful manner. The GDPR also equips the subject with a series of additional rights (under Arts. 12-20) including to withdraw their data from the research at any time. Under these circumstances, it is suggested that researchers, working with relatively few subjects, who take the trouble to build up ties and to involve these in the overall research aim (e.g. better treatment for a given disease, from which the subject or relative may themselves suffer) has a good chance of being able to acquire and use relevant data in an effective (as well as legally compatible) manner.

At the same time, it may be queried how far this legal framework is favorable to larger scale research, particularly where the researcher is positioned remotely from the subject, and has little (or perhaps no) direct contact, instead receiving the data via a third-party intermediary. Here the hurdles, including the need for reconsent to different research uses and the guaranteeing of the subject's rights under the GDPR, may pose considerable logistical and organizational challenges. In this kind of situation, it appears that data privacy and autonomy concerns could lead to sub-optimal research outcomes, though this is admittedly difficult to quantify.

A further interesting question, in the specific context of 'big data'-analytic medical research is whether the risk-based approach to data-processing found in the GDPR may inhibit such research, even where, in line with normative codes of ethics, the interests and concerns of the research participants are safeguarded to the letter. This arises in view of the need, under Art 35 GDPR, for data processing operations "*likely to result in a high risk to the rights and freedoms of natural person*" to be subject to a rigorous prior 'data protection impact assessment', potentially including the need for approval by the relevant supervisory authorities. Arguably, this would apply if proposed data research is likely to generate knowledge that would create a dilemma not adequately addressed by the research plan. This is certainly a risk with unsupervised data-analytic processes of the kind used to make sense of large volumes of data, which discern probabilistic correlations rather than causal relations. In particular, it may lead to cases where science can predict, on the basis of a person's data, that they have a high probability of contracting a given disease, but (lacking firm causal knowledge) not do much to stop it: here, the dilemma would be what to tell the person.

In summary, it can fairly be said that, while the GDPR contains important provisions, contributing to safe and ethical use of medical data for research, it has the potential to make both the approval and execution of such research quite complicated. While this may result in Europe lagging behind other parts of the world, where such legal restrictions do not operate, it is not clear that it will (or to what extent the legal constraints will be enforced with regard to research). For example, in the last scenario, a rule requiring data researchers to privilege data-analytical processes that generate actionable causally-grounded knowledge, could also provide a (scientifically) useful steer. Careful, ongoing analysis, which takes account of diverse data-analytical research methods, and their respective strengths and weaknesses (including risks to the data subjects and wider society) will be required in order to progress towards a balanced legal and ethical solution.

6. Conclusions and steps forward

From the above said follows that the initiatives towards creating a European digital ecosystem of trust, including trustful research environments, are quite a few, spreading across regulatory, societal, technological, geo-political, and legal fields. Such initiatives encompass mechanisms integrating ethics-by-design, privacy -preserving technologies, the phenomena of data altruism, data intermediaries, self-sovereign identities, instruments towards web decentralisation. An important infrastructural and normative link has been established thus enabling the creation of trustworthy setting facilitating safe data sharing. The avenues already opened for research, both by the FAIR principles and the legal grounds provided by the GDPR, have found due reflection and productive adoption. A remarkable sign is that such initiatives mainly pay tribute to the core values of European society, namely fundamental rights and ethics.

The further the story goes, the more challenges emerge. In particular, it becomes evident against the background of integrating the stringent GDPR requirements into research settings, especially when health data is concerned - an important asset both for individuals, healthcare and associated industries, the public and the state. The attempts to address such challenges are quite prominent, such as solutions around explainable AI, innovative data control mechanisms, efforts to address data biases and discriminatory capacity hidden in data and algorithms. Such aspects are quite important, merit due reflection and interdisciplinary approach, what goes beyond the realms of this paper, but bear rich potential for further exploration.

Acknowledgment: This work was supported by the European Commission through the H2020-INFRAIA-2018-2020 / H2020-INFRAIA-2019-1 European project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042). The funders had no role in developing the research and writing the manuscript.

References:

- Amoore, L. and Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, *48*(1), 3-10.
- Andrienko G. L., Andrienko N. V., Boldrini C., Caldarelli G., Cintia P., Cresci S., Facchini A., Giannotti F., Gionis A., Guidotti R., Mathioudakis M., Muntean C. I., Pappalardo L., Pedreschi D., Pournaras E., Pratesi F., Tesconi M. and Trasarti R. (2021). (So) Big Data and the transformation of the city. Int. J. Data Sci. Anal. 11(4): 311-340
- Andrienko N. V., Andrienko G. L., Fuchs G. and Jankowski P. (2016). Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. Inf. Vis. 15(2): 117-153
- Appadurai, A. (2020). The COVID exception. *Social Anthropology*.
- Asikis T. and Pournaras E. (2020). Optimization of privacy-utility trade-offs under informational self-determination. *Future Generation Computer Systems*, *109*, 488-499.

- Aufrère, L. and Lionel M. (2018). Pour une protection sociale des données personnelles. Working Paper Projet EnCommuns. Accessed August 13, 2020. https://hal.archives-ouvertes.fr/hal-01903526
- Comandé, G. (2020). Unfolding the legal component of trustworthy AI: a must to avoid ethics washing. In Annuario di diritto comparato, ESI, 2020, 39-62, (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690633);
- Couture, S. and Toupin, S. (2019). What does the notion of "sovereignty" mean when referring to the digital?. *new media & society*, *21*(10), 2305-2322.
- Danezis G., Domingo-Ferrer J., Hansen M., Hoepman J.-H., Le Métayer D., Tirtea R. and Schiffner S. (2015). *Privacy and Data Protection by Design – from Policy to Engineering*, European Union Agency for Network and Information Security (ENISA).
- Daucé, F. and Musiani, F. (2021). Infrastructure-embedded control, circumvention and sovereignty in the Russian Internet: An introduction. *First Monday*, *26*(5).
- DeNardis, L. (2014). *The Global War for Internet Governance*. New Haven: Yale University Press.
- Desrosières, A. (1998). *The politics of large numbers: A history of statistical reasoning*. Harvard University Press
- Domingo-Ferrer J. and Blanco-Justicia A. (2020). Ethical value-centric cybersecurity: a methodology based on a value graph. *Science and Engineering Ethics*, 26(3):1267-1285.
- Domingo-Ferrer J. and Soria-Comas J. (2021, to appear). Multi-dimensional randomized response. *IEEE Transactions on Knowledge and Data Engineering*. https://arxiv.org/pdf/2010.10881.pdf
- Domingo-Ferrer J., Blanco-Justicia A., Sánchez D. and Jebreel N. (2020). Co-utile peer-to-peer decentralized computing. In 20th *IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing – CCGrid 2020*, pp. 31-40. IEEE.
- Eiss, R. (2020). Confusion over Europe's data-protection law is stalling scientific progress. *Nature,* 584:498.
- Fiore M., Katsikouli P., Zavou E., Cunche M., Fessant F., Le Hello D., Matchi Aïvodji U., Olivier B., Quertier T. and Stanica R. (2020). Privacy in trajectory micro-data publishing: a survey. Trans. Data Priv. 13(2): 91-149
- Frey B. S. and Oberholzer-Gee F. (1997). The cost of price incentives: an empirical analysis of motivation crowding-out. *The American Economic Review*, 87(4):746-755.
- Giotitsas C., Pazaitis A. and Kostakis V. (2015). A peer-to-peer approach to energy production. *Technology in Society* 42: 28-38. https://doi.org/10.1016/j.techsoc.2015.02.002
- Hacking, I. (2004). Between Michel Foucault and Erving Goffman: between discourse in the abstract and face-to-face interaction. *Economy and society*, *33*(3), 277-302.
- Haggart, B., Tusikov, N. and Scholte, J. A. (Eds.). (2021). *Power and Authority in Internet Governance: Return of the State?*. Routledge.
- Hummel, P., Braun, M., Tretter, M. and Dabrock, P. (2021). Data sovereignty: A review. *Big Data & Society*, *8*(1)
- Irion, K. (2012). Government cloud computing and national data sovereignty. *Policy & Internet*, *4*(3-4), 40-71.
- Isin, E. and Ruppert, E. (2020). The birth of sensory power: How a pandemic made it visible?. *Big Data & Society*, *7*(2).
- Koloskova A., Stich S., and Jaggi M. (2019). Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proc. of the 36rd International Conference on Machine Learning – ICML 2019*, pp. 3478-3487.
- Linkov, I., Trump, B. D., Poinsatte-Jones, K. and Florin, M. V. (2018). Governance strategies for a sustainable digital world. *Sustainability*, *10*(2), 440.
- Matwyshyn, A. (2019). The Internet of Bodies. *William & Mary Law Review* 61 (1):77-167. Accessed August 13, 2020. https://ssrn.com/abstract=3452891
- McMahan H. B., Moore E., Ramage D., Hampson S. and Agüera B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proc. of the 20th Intl. Conf. on Artificial Intelligence and Statistics – AISTATS'2017*, pp. 1273-1282.
- Nanni, M., Andrienko, G., Barabási, AL. *et al.* (2021). Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. *Ethics Inf Technol*. https://doi.org/10.1007/s10676-020-09572-w
- Pellungrini R., Pappalardo L., Pratesi F. and Monreale A. (2018). A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. ACM Trans. Intell. Syst. Technol. 9(3): 31:1-31:27
- Pratesi F., Gabrielli L., Cintia P., Monreale A. and Giannotti F. (2020). PRIMULE: Privacy risk mitigation for user profiles. Data Knowl. Eng. 125: 101786

- Pratesi, F., Monreale, A., Trasarti, R., Giannotti, F., Pedreschi, D. and Yanagihara, T. (2018). PRUDEnce: a System for Assessing Privacy Risk vs Utility in Data Sharing Ecosystems. Trans. Data Priv. 11(2): 139-167
- Sanfilippo, M., Frischmann, B. and Standburg, K. (2018). Privacy as commons: Case evaluation through the governing knowledge commons framework. *Journal of Information Policy*, *8*, 116-166. https://doi.org/10.5325/jinfopoli.8.2018.0116
- Schneider, G. and Comandè, G. (2021a). Can the GDPR Make Data Flow for Research Easier? Yes it Can! By Differentiating!, in Computer Law & Security Review, 41 105539;
- Schneider, G. and Comandè G. (2021b) Differential Data Protection Regimes in Data-driven Research: Why the GDPR is More Research-friendly Than You Think, in German Law Journal, forthcoming
- Schollmeier, R. (2001). A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer architectures and applications. Proceedings of the First International Conference on Peer-to-Peer Computing, 27-29.
- Teli, M., Bordin, S., Menéndez Blanco, M., Orabona, G. and De Angeli, A. (2015). Public Design of Digital Commons in Urban Places: A Case Study. *International Journal of Human-Computer Studies* 81: 17-30. http://dx.doi.org/10.1016/j.ijhcs.2015.02.003
- Toussaert S. (2021). Upping uptake of COVID contact tracing apps. *Nature Human Behaviour*, 5:183-184.