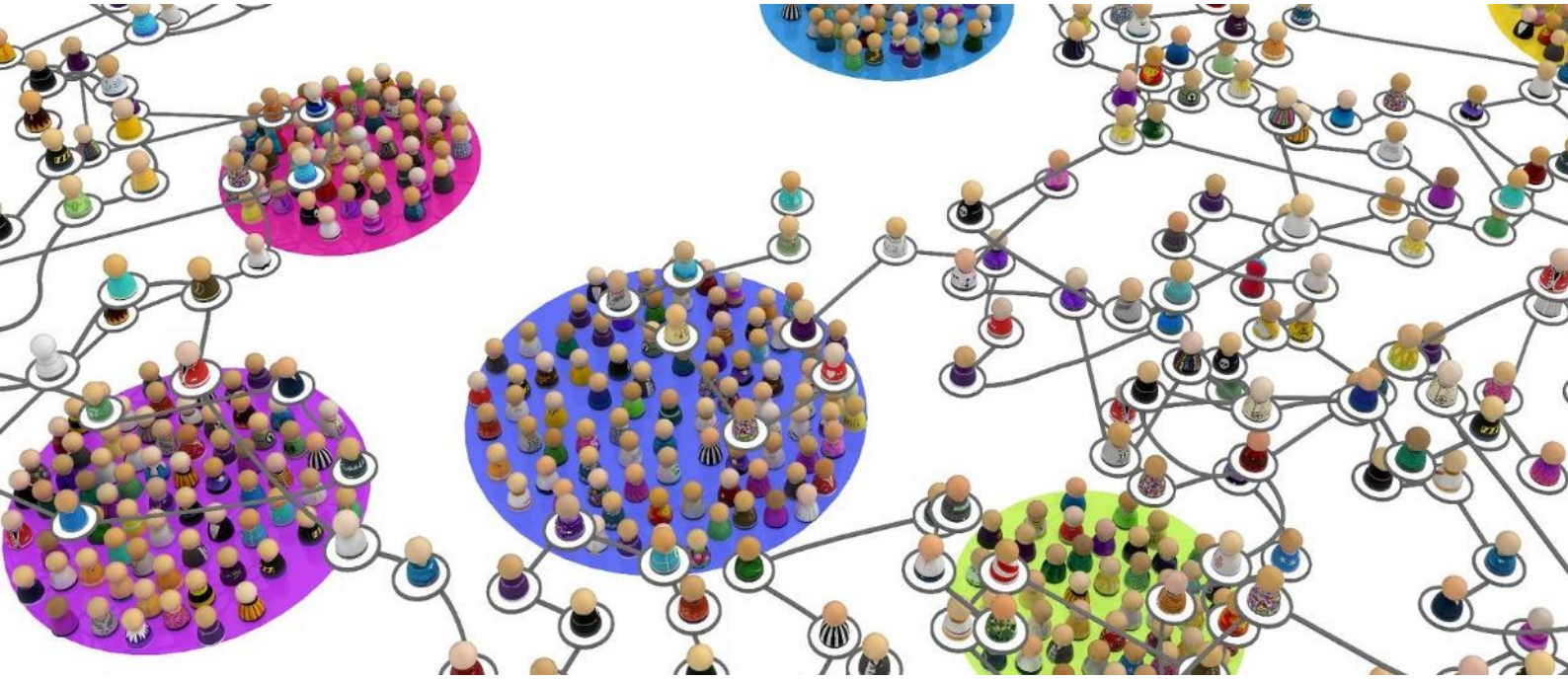
SHERPA

Report of Interview Analysis

Kalypso Iordanou, Eleni Christodoulou and Josephina Antoniou



Deliverable 2.2. - October 2020

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Document Control

Deliverable	D 2.2. Report of Interview Analysis
WP/Task Related	WP 2
Delivery Date	Month 30
Dissemination Level	PU
Lead Partner	UCLanCY
Contributors	
Reviewers	
Abstract	
Key Words	Ethics; Human Rights; Big Data; AI; education; interviews;

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
0.1	17.10.2020	Kalypso Iordanou, Eleni Christodoulou and Josephina Antoniou	Doris Schroeder	First Draft
0.2	29.10.2020	Kalypso Iordanou, Eleni Christodoulou and Josephina Antoniou		Second Draft

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Contents

Executive Summary

Thirty-five interviews with stakeholders across Europe took place in the context of SHERPA's Work package 2, Stakeholder Analysis and Consultation, Task 2.2. "Interview Stakeholders". The aim of the interviews was twofold. Firstly, to gain stakeholders' views regarding the recommendations that have been developed in the SHERPA project, particularly regarding the set of *Guidelines for Users and Developers* (T3.2), the *Regulatory Options* (T3.3.) and the *Terms of Reference for a New Regulator* (T3.6) (14 interviews). Secondly, to obtain an in-depth understanding of what stakeholders consider to be the main ethical issues that come out of Artificial Intelligence (AI) and big data; the way those ethical issues are currently addressed, and their suggestions on how those ethical issues can be addressed efficiently (21 interviews). The distribution was as follows:

Figure 1. Activities involved in Task 2.2.

The findings of the (14) interviews on stakeholders' views about SHERPA outputs , will provide internal feedback to the project and serve as a means for corrective action in the future. The findings of the (21) *Exploratory* interviews will contribute towards SHERPA's objective for development of a set of recommendations for the responsible development of SIS.

Thematic analysis was used to analyse the interview data. The (8) *Guidelines* interviews highlighted a number of positive aspects, but identified some issues and challenges to consider for their successful implementation. The (6) *Regulatory* interviews, highlighted specific topics, such as what needs to be regulated and how such regulation should take place, concluding with a number of proposed areas that a potential new regulator should focus on.

The *Exploratory* interviews (21) revealed 11 main ethical issues identified by stakeholders in relation to big data and AI to involve lack of transparency; 'information asymmetry' and lack of public understanding; biased data and lack of critical thinking; loss of human agency and dignity; failure to protect privacy as a fundamental human right; surveillance, manipulation and coercion; unethical monetisation of data; lack of accountability and product liability; loss of human jobs and mistreatment of employees; health and environmental risks; exacerbating socio-economic inequalities and the digital divide.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
AHR	Aequitas Human Rights NGO
AI	Artificial Intelligence
DMU	De Montfort University
EBS	European Business Summit
EUREC	European Network of Research Ethics Committees
GDPR	General Data Protection Regulation
(AI HLEG)	High-level Expert Group on Artificial Intelligence
IEEE	The Institute of Electrical and Electronics Engineers
MS	Mutual Shoots
NEN	The Royal Netherlands Standardization Institute

SIS	Smart Information Systems
TRI	Trilateral Research Ltd
UCLanCY	University of Central Lancashire - Cyprus
UT	University of Twente

1. Introduction

1.1. Background and objectives

The aim of the interviews was twofold. Firstly, to gain stakeholders' views and suggestions for novel proposals for the responsible development of SIS (WP3). Secondly, to obtain stakeholders' feedback for evaluation, validation and prioritisation of the proposals developed in the project (WP4).

To achieve these objectives, a total of 30 interviews were planned. In the Consortium meeting in Cyprus in October 2019, it was decided that the interviews should focus on three thematic areas:

a) *Guidelines* (Short version of two sets of *Guidelines for Users and Developers* - D3.2)

b) *Regulatory Options* (T3.3) & *Terms of Reference for New Regulator* (T3.6)

c) *Exploratory Interviews on Ethical issues on AI/Big data*

The first two sets of interviews – a) on *Guidelines for Users and Developers*, and b) on *Regulatory Options* and the *Terms of Reference (ToR) for New Regulator* – aimed to evaluate the major outputs and recommendations developed in the SHERPA project. The third set of interviews (*Exploratory*) aimed to explore stakeholders' views on ethical issues in relation to AI/Big Data, as well as how best to address them, using open-ended questions, thereby contributing to the formation of recommendations.

The content of the interviews was decided after receiving feedback from Task Leaders in the project, particularly regarding T3.2. *Guidelines for research and innovation of SIS*; T3.6. *Regulatory Options*; T3.3. *ToR for New Regulator*; T3.4. *Assessment of standardization potential*, and T3.4. *Prioritization and finalization of recommendations*. The results of the interview analysis will feed back to those tasks.

The report is structured as follows. We begin by providing the methodology, including information about the data collection and data analysis process, before moving on to the findings of the data analysis. The empirical section consists of three main parts. The first part presents the results of the *Guidelines* interviews, the second part presents the results of the *Regulatory Options/New Regulator* interviews, and the third part the findings and future suggestions that emerged from the *Exploratory* interviews. We conclude by highlighting the core results of the data analysis, as well as their implications for policymakers.

2. Methodology

2.1. Ethical Approval and Data Management

The Task Leader applied and secured ethical clearance from the Cyprus National Bioethics Committee (see Appendix A). Each partner who conducted interviews made sure that all the required measures were taken in order to be compliant with the ethics guidelines for conducting research in the country where the interviews took place. Prior to the interviews the interviewees' written consent was secured, using the information sheet (Appendix B), adapting it first to the aims of each interview, and the consent form (Appendix C) that had been prepared by the task leader. The interviews were transcribed by the partner who conducted the interview and few cases by a professional transcriber. The transcripts of the interviews were anonymized. Each partner uploaded the transcripts and the consent forms for the interviews, in the project's secure drive, to which only SHERPA partners have access.

2.2. Data Collection

2.2.1. How many Interviews were conducted and by whom?

Overall, 35 interviews were conducted. One partner (EBS) pursued more interviews (7) than the 2 that were initially allocated to it, because of increased interest from potential participants. The 35 interviews involved 8 interviews focusing on *Guidelines*, 6 interviews focusing on *Regulatory Options and ToR*, and 21 *Exploratory* interviews. Table 1 shows the number of interviews pursued by topic, and the project partner who conducted the interview.

Table 1. Number of interviews pursued by Topic and Partner.

Guidelines[1]	
DMU	2
NEN	1
UT	1
MS	1
Fsequre	2
EUREC	1
Total	**8**
Regulatory Options/Terms of Reference[2]	
TRI	2

[1] Represented in the report as G1, G2, G3, etc.
[2] Represented in the report as R1, R2, R3, etc.

UCLan Cyprus	2
AHR	2
Total	**6**
Exploratory Interviews[3]	
UT	1
NEN	1
EUREC	1
MS	1
EBS	7
AHR	5
UCLan Cyprus	5
Total	**21**

2.2.2. Participants

Participants from different stakeholder groups were recruited. Representatives from the following stakeholder groups were included in the sample:

- Data analysts and AI experts from industry and the private sector

- Experts from academia, professors and senior researchers

- Policy makers

- Professional body representatives

- Civil Society Organization representatives

2.2.3. How participants were recruited

The task leader suggested the inclusion criteria, namely diversity in stakeholder groups, gender balance and ethnic diversity. The participants should also be professionally involved in some aspect of relevance to the interviews. The Consortium partners discussed with the task leader their plans for recruiting participants to make sure that the recruitment criteria were met.

Partners recruited stakeholders either using their network or by finding the contact details of experts with highly relevant profiles on the web. Prospective interviewees were sent the information sheet and asked if they were interested in being interviewed. Once a positive response was secured, a mutually convenient time was agreed to conduct the interview. The interviewee was asked to sign and return the consent form to the interviewer, prior to the interview.

[3] Represented in the report as E1, E2, E3 etc.

2.2.4. How the Interviews were conducted

All the interviews were conducted virtually, using a video conferencing service (e.g. Zoom, Skype, MS Teams). The duration of the interviews was on average between 40-90 minutes. The interviews were recorded and transcribed. Two interviews conducted in Greek were transcribed in full and the relevant parts used in the report were translated by the authors of this report.

2.2.5. The content of the questions

For the 6 *Guidelines* interviews the Task Leader of Task 3.2. (UT - "Develop guidelines for research and innovation in and with SIS") provided the interview questions. There were two sets of questions focusing on either the set of *Guidelines for Developers*, or the set of *Guidelines for Users*. Appendix D presents the list of questions that were used for these interviews.

For the 8 *Regulatory Options/ToR of New Regulator* interviews, the Task Leader of T3.3. (TRI - "Explore regulatory options") and T3.6 ("Propose terms of reference for a new regulator for SIS") provided the questions. Appendix E shows the interview questions on *Regulatory Options*, and Appendix F includes the questions for *Terms of Reference of New Regulator*. Both sets of questions were used in each interview which focused on *Regulatory Options* and *ToR of New Regulator*.

For the 21 *Exploratory* interviews, the T4.2. ("Stakeholder evaluation and validation") Task Leader developed the questions, in collaboration with the Co-ordinator of the project, Prof. Bernd Stahl. Although there was some degree of flexibility in the interview discussions, the majority of interviews followed the same questions. These revolved around 3 core aspects:

- the main ethical issues that come out of AI and big data and their relation to human rights;
- current efforts to address these ethical issues, and
- suggestions of activities that should be undertaken in the future to deal with ethical issues that have not been yet adequately addressed.

Table 2 presents the questions that were used in the Exploratory Interviews.

Table 2. Questions used in Exploratory Interviews*

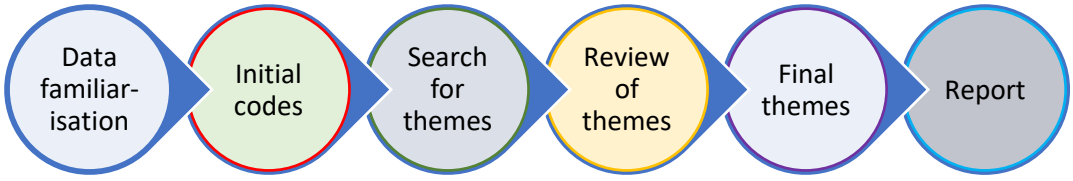
1. What are the (3-5 main) ethical issues that come out of AI and big data?
2. How do those ethics issues relate with Human Rights?
3. How are those ethical issues currently addressed?
4. What are the limitations of the current efforts addressing those ethical issues?
5. What ethical issues haven't been addressed so far?
6. What are the three most important activities that should be undertaken to deal with the ethical issues that haven't been adequately addressed yet?
7. What do you think guidelines for developing and/or using AI systems should look like?
8. Which changes in regulations could protect human rights?
9. How can standards help dealing with AI / big data analytics?
10. How could a new regulator help and what should it look like?

*Some interviews had an additional question on education: "How can education contribute in dealing with the issue of AI/Big Data and ethics/human rights?"

2.3. Data Analysis

The data obtained from the interview transcripts was analysed using thematic analysis, and in particular the framework provided by Braun and Clarke (2006). We followed the six stages of thematic analysis (2006, p.87): (1) Initial data familiarisation; (2) Generation of initial codes; (3) Search for themes; (4) Review of themes in relation to coded extracts; (5) Definition and final naming of themes; (6) Production of the report (see Figure 2).

Figure 2. Six Stages of Thematic Analysis



Thematic analysis can be defined as: 'a method for identifying, analysing and reporting patterns (themes) within data. It minimally organizes and describes [the] data set in (rich) detail' (2006, p.79).

Although thematic analysis always involves a degree of interpretation (Boyatzis, 1998), due to the main objective of the interviews, which was to let the participants' opinions and expertise inform the findings, we took a data driven, inductive approach in which codes and themes were generated from the data (open coding), rather than having a codebook prepared in advance of the data analysis (deductive approach). Our themes were however informed from the results of a similar research process which had been conducted earlier in the project, with similar questions but using the data collection method of focus groups instead of individual interviews.

To ensure better organisation, classification and interpretation of the data, the analysis was supported by Nvivo, a qualitative data analysis computer software (Version 12). For the coding process, we followed closely the instructions and advice of Braun and Clarke (2006); Bryman (2008), and Charmaz (2004), namely:

a) Code as thoroughly as possible as 'you never know what might be interesting later' (2006, p.90). This point is also suggested by Bryman (2008) as well as Charmaz (2004) who suggest a line by line coding and for researchers not to be alarmed by the proliferation of codes; this is important so as not to lose any detail or potential interpretation of the data.
b) Code the data extracts in an inclusive way (with the surrounding text) so 'that the context is not lost'.
c) The same text can be coded for several different codes (and then themes).
d) Keep in mind that contradictions, inconsistencies and tensions are an inevitable part of the data and the researcher should not feel that they should smooth these out or ignore them.
e) Retain accounts that depart from the dominant themes (e.g. when there was a suggestion that was only mentioned once).

We took a strong qualitative focus rather than a quantitative one: we identified patterns of meaning expressed whilst also including useful inputs even if they were only mentioned once. In line with qualitative analysis, the focus was on answering the 'what, how and why' questions, the kinds of issues raised, what these mean for improving our understanding of ethics and human rights related to SIS, and the insights and experiences offered by experts and stakeholders, rather than how many times issues were mentioned.

For the *Regulatory* interviews 111 codes emerged and for *Guidelines* interviews 103 codes emerged. For the *Exploratory* interviews 107 codes emerged (of different hierarchies, 'parent', 'child', 'grandchild' etc.). The coding of text segments was not mutually exclusive, for instance one paragraph could include up to 4-5 different codes. Appendix G offers an example of a visualisation from stage 2 of the data analysis of the *Exploratory* interviews, showing the codes that emerged regarding suggestions for addressing ethical issues. The findings are presented anonymously. Therefore, we use numbers to refer to the interviews.

3. Analysis of Findings: Guidelines Interviews

The 14 interviews that focused on discussing the proposed *Guidelines for the Development and Use* of *Smart Information Systems* (SIS) technology, featured a variety of stakeholders from different sectors, such as academia, industry, policy, etc. All the stakeholders had expertise with technology either from a research perspective, a development perspective or an organisational perspective (e.g. management).

The interviewees were asked to review two documents: the first on responsible development of SIS, and the second on responsible use of SIS (this document focuses on organisational use of SIS). The proposed Guidelines are publicly available on the SHERPA project website (*https://www.project-sherpa.eu/guidelines/*). In particular, the *Development Guidelines* deal with how developers can construct an ethical AI or Big Data system, and the *User Guidelines* deal with how to ethically use an AI or Big Data system, especially adapted to governance and management of organisations that use these technologies as part of their services.

Both Guidelines documents were shared with stakeholders prior to the interviews. The documents aim to stand as **practical guides** to incorporating the high-level ethical and human rights requirements into processes of development and organisational use of SIS technology. The interviews focused mainly on reactions to the Guidelines in terms of their **overall *"feel"***, specifically their **clarity**, **engagement**, **presentation**, **operability** and **usefulness**. Discussion went further into **suggestions for improvement** both in specific sections and generally, as well as into comments regarding the **practicality of the Guidelines**, to evaluate the overall practicality goal. The interviewees also had the opportunity to **compare the documents with similar guidelines documents**.

The following sections present, primarily, the reactions of the stakeholders on aspects of clarity, engagement, presentation, operability and usefulness of the Guidelines. Further, they present reactions on comparison with similar guidelines that the stakeholders had previously interacted with, as well as suggestions for improvement. Also, the section presents the stakeholders' feedback on the practicality goal of the Guidelines, to conclude with a summary of the main findings, focusing especially on how to use the finding to move forward with the guidelines as practical tools.

3.1. Clarity: Findings

Feedback on the clarity of the Guidelines is given both generally and regarding specific sections. The general feedback is quite encouraging, with stakeholders finding the Guidelines '*well explained*' (G 3), and the language '*clear*' (G2).

Specific feedback on the clarity of individual sections varies. For instance, Section 2 is believed to be '*as long as it needs to be*', and clear from a high-level requirement point of view, as it is '*directed at those people in an organisation that need the high-level requirements and don't need the detail*' (G3). Opinions are split when it comes to the Introduction. On the one hand, it is '*one of the most valuable parts, the one with the table on the second page […] you really get a feeling of what this document is about*' (G2), but there are also comments that more work is needed on this introductory section:

> *My point of view is that the introduction misses the purpose and a target. Who is this document for? I open this document; I want to understand is it for me or is it not for me?* (G1).

Although a separate glossary is available, several of the stakeholders commented on the need for clarification of specific terms and acronyms in the Guidelines, so that readability is enhanced.

An example of a term that needs clarification was *account* in Section 3.2.2, i.e. the phrase *accounts for the ethical guidelines for the system in the implementation plan* is characterised as '*ambiguous*' because it is not clear whether '*it means implement them or address them or give a rationale*' (G3).

Another example of the use of terms, mentioned by multiple interviewees, which may have hindered readability of the Guidelines, was the example of using the terms *positive liberty* and *negative liberty*:

> *for me it was helpful […] but I don´t know if other readers have the same background […] the positive liberty and negative liberty, if those are more generally well-known concepts that I am just not familiar with then, that was something that I had to […] do a double read of*. (G5).

> *maybe at some point you just have to read it in more detail, like what is positive liberty, negative liberty. If you are a developer then you are not familiar with those terms* (G2).

In addition to clarification of terms and acronyms, it was mentioned that the organisation of the content can hinder clarity in specific sections:

> *it is hard to see why certain things fall under a heading. Autonomy is mentioned in 1. Human agency, liberty and dignity but it is also linked to privacy and data governance, most notably in relation protection of data* (G6).

In general, the interviewees commented that the feel of the document in terms of clarity was positive throughout, however, the above-mentioned comments on further explanation provide important **feedback for future revisions**.

3.2. Engagement: Findings

The second area investigated in the interviews was the level of engagement with the documents, i.e. to what extent they hold the readers' attention and interest. The comments received were again both general and related to specific sections, and were both positive and negative with regards to the engagement aspect. In their negative comments, the interviewees often propose **mitigation measures or solutions**; we have collected their suggestions and present them as part of the overall feedback for this section.

On the positive side:

> *I found them engaging, but then I would, because it's what I do, it's about ethics. I'd found the introduction and the explanations of how the principles were derived and relevance to human rights principles good* (G3).

On the negative side, feedback identified some specific sections as not very engaging. For example, Section 3 was characterised as '*not very interesting*' because '*it provides a lot of info but has limited link with the proposals*', with the interviewee asking whether it is necessary (G6). Other stakeholders

seem to agree that Section 3 '*didn't add too much*', because it presents a set of general frameworks, and usually interested parties '*know approximately what these frameworks are*' (G1). However, Section 4 seems to generate a more positive response from the stakeholders, as it is '*much more interesting*' (G1).

Another issue with engagement that is picked up by multiple stakeholders is the length of the document; there seems to be a general consensus that the Guidelines would be better if available in a hierarchical, interactive manner, e.g. through a digital platform rather than as reports.

> *If you have it as a, sort of, interactive thing, that might also work, because at the moment, as I say, I don't see developers with a deadline and everything just sort of reading […] a 40-page document* (G4).

Finally, there was a general comment on how engagement could be improved by targeting specific professional functions in specific sections, so that the items that are relevant for specific professionals are easy to locate, i.e. proposals for specific areas, e.g. service design, implementation, deployment etc. should be easy to locate.

> *I have a general comment on how proposals could be organized for better readability and engagement. Imagine I'm a manager and I'm opening… I'm a manager responsible for service design. What is in my immediate span of control is everything related to service design. What I would like to have in the Guidelines is an understanding of which proposals should be implemented at which stages. So what I would like to do, I would like to open my stage – service design, or open my stage – deployment and implementation and find all the proposals related quickly to this stage* (G1).

Nevertheless, the majority of the comments with regards the aspect of engagement were positive.

3.3. Presentation: Findings

The third aspect investigated in the interviews was that of the presentation of the Guidelines. A recurring comment from the interviewees had to do with the images and graphics. The stakeholders' views on images and graphics are not consistent. For instance, there were some negative reactions from stakeholders who '*don´t think the images help*' but '*think they actually hurt a little bit*' (G5), or that the set of graphics '*doesn't have explanatory power*' (G1). However, opposing views from other stakeholders state that '*graphics are always useful and they haven't been overused […] too many can be distracting and unnecessary, but […] what you've got is appropriate and relevant and they're in the right places*' (G3).

In terms of the document structure, interviewees reported that the hierarchical structure, with many layers of sub-sections, can become overwhelming:

> *the high-level requirements are numbered, human agency is one, technical robustness and safety is two, privacy and data governance is three, but those aren´t the same numbers that are used for the chapters or even the subtopics. […] And then you also have the proposals, and the proposals are numbered, but then there is other numbered bullets throughout each of the subchapters and so to me it was just, there is a lot of different numbers* (G5).

Interviewees argued that the hierarchical structure is necessary in order to accommodate all the content, but that makes the document too long for practical use:

> *I find it a bit on the long side. If there was […] something like a management summary or something like the most important things on two pages and then you could take a decision based on that whether you want to go further, hopefully* (G2).

The stakeholders suggested that to preserve the hierarchical structure without leading to confusion, it would be useful to have the Guidelines in an interactive format, i.e. digitally. This was a suggestion for improved engagement as mentioned above, but it has also been mentioned as an improvement for the presentation of both sets of Guidelines, by multiple stakeholders.

> *it would be great if there was […] an interactive format […] - the seven high level requirements could be realised maybe in a way like the UN Sustainable Development Goals, like you have let's say a fancy framework where it's very clear on one side what are the important points and then you can click on it and then find out more about that topic* (G).

> *… an online format and I've given the example of the government's Data Ethics Framework, where you've got each aspect or proposal as just the key titles and then you click on the link to see what's behind it.* (G3).

3.4. Operability: Findings

Since the overall goal of proposing the sets of Guidelines for responsible development and use is to create practical guides for developers and organisations, it is important to investigate the aspect of operability. Again, there are varying opinions about operability between interviewees. Initially, there is agreement on the significance of proposing both sets of Guidelines since '*any guidelines for organizational users are also important for the developers to acknowledge*' (G7), but there are opposing opinions on the operability of these document for all. It seems that non-developers would find the guidelines (for organisations) more operable, '*because this is a guideline for all kinds of businesses*', and for the purposes of organisational use '*they are okay at that level*' (G4).

> *But then it just raises some questions, yes I want to do this but how? And this is then answered later in the document. So yes, I feel it provides some understanding for somebody who is not a developer because then you would understand what are the steps, what is data understanding, what is data preparation. But for somebody who approximately knows what this is then there is not too much about ethics let's say. That was my understanding […] I would take section three out and really focus on section four (G2).*

The interviewee proposes shifting focus to Section 4 rather than Section 3, which appears more actionable. Section 3 was also mentioned by some interviewees during the discussion of clarity, and the repetition of the comment calls for an action to review the specific section according to the interviewees' feedback.

The developers may not find the Guidelines '*very straightforward to operationalise. They have the feel of a position statement rather than specific guidelines*' (G8). A future improvement would be to enrich the *Development Guidelines* with more practical examples to enhance their operability potential. We will further revisit potential suggestions for improving the practicality of the Guidelines in the following sections, as this aspect was discussed separately in the interviews.

3.5. Usefulness: Findings

Finally, in terms of "*feel*", it is important to document the reactions of the interviewees in terms of the usefulness of the Guidelines, both for developers and organisational users.

Both developers and organisational users find that the Guidelines are useful. For organisations *'it's sometimes very difficult to see what's sensitive data and what's not sensitive data, what data makes sense to collect… very often it's a "let's collect it all" approach because they are not really sure*' and guidelines can be useful since they can provide the necessary answers, since '*there's a lot of communication needed inside companies*' (G4). For developers, the guidelines "*all make sense*" and '*in many ways it´s clear enough to be well used*' (G5).

One of the reasons that stakeholders agree on the usefulness of the Guidelines is possibly that '*they cover the different areas of use*', for example, '*you've got your governance area and deployment implementation*', allowing '*different people coming from different perspectives [to] find something relevant in there for them*' (G3).

The discussion on usefulness will continue in subsequent sections as there will be suggestions on how to make the Guidelines more practical, and hence more useful. Before moving into the practicality of the Guidelines, the interviewees were first asked to compare the proposed Guidelines with other similar documents. The opinions are presented next.

3.6. Comparison to other relevant Guidelines Documents

Part of the interviews focused on comparing the proposed Guidelines with similar guidelines documents that they had previously come across, and most interviewees said that they had read several guidelines documents as part of their professional roles. The first element discussed was the "*feel*" of usefulness of the Guidelines, that we have presented in the previous section. This is an important aspect to help adoption of these Guidelines by professionals:

> *Usefulness is better than most guidelines because it is more concrete. Many guidelines on ethics and AI are at the level of HLEG principles. This is more at the requirements level* (G6).
>
> *I, definitely I think the usefulness of these guidelines are definitely improved […] compared with OECD guidelines for example* (G1).

Then, there is the element of detail and consequent understandability:

In terms of detail, interviewees stated that *'… it's good that they're so detailed, because […] just very high-level stuff, there exists enough of this kind of thing'* (G4). Detail also supports implementation of the responsible principles, in contrast to other such guidelines documents, '*especially, say, coming from the ethics perspective' that are 'much more conceptual and abstract and I think what your guidelines provide above and beyond that is more detail and some help to people involved in actually implementing them*' (G2).

In terms of "understandability" it is harder to compare the Guidelines to HLEG principles, or similar high-level requirements documents, because '*the target group is different. Most guidelines are for administrators or C-suite. But this is more at the level of a manual, so it is hard to compare. It is better to compare to Prince II or ITIL manuals. If that is your ambition then you have to add text and convert*

it into a study book'; in such a case '*more is needed (examples, best practices) to really implement well. Or link it to existing literature of ITIL and COBIT*' (G6).

The element of practicality is also different than the EU guidelines, especially the fact that the proposed documents have a clearer focus on implementation:

> *I think that´s very helpful to me, this is much more purpose focussed on implementation than what the EU put out. The High-level Expert Group is, there's is much fluffier und fuzzier and it´s more about the 'why' and less about the 'how' or the 'what'. So that to me is very helpful* (G5).

In terms of practicality, it is important to compare the documents with EU regulation on personal data protection, GDPR. Responsible guidelines documents that aim to propose a practical way of incorporating ethics into development and use of SIS, must consider solutions related to GDPR implementation. The stakeholders picked up on some of the similarities:

> informed consent is already a topic that´s in GDPR, personal data use and reduction and elimination and security concepts are already part of GDPR. […] the topic of data minimisation is already there. So, the concept of having a data protection officer is already there (G5).

The positive comparative feedback was followed by a discussion on improving the Guidelines further, by looking at general suggestions but also very specific suggestions.

3.7. Proposed suggestions for improvements

The stakeholders interviewed were asked to provide suggestions for improving the Guidelines documents. Some of these suggestions were focused on specific sections and some were more general. We have selected a few important suggestions for specific sections, and we present these first, in ascending order according to the sections they refer to. Then, we will provide a summary of the general improvement suggestions collected from the interviews.

The stakeholders' feedback is positive overall for the structure and content of Sections 1 and 2, with the exceptions of some concerns about clarity that were documented above. The suggestions presented next, begin with Section 3. In some cases, a need is identified for more explanation and detail in specific sections, e.g. in 3.1.4 and 3.1.6:

> *In 3.1.4 "Modelling" sub-bullet 2. "Generate test design" it should be added that this includes defining by what kind of measure model accuracy is calculated. Since this is often an average over some subset of data, there are risks related to optimization procedures over-emphasizing the majority data, leading to unfair treatment of minorities. In 3.1.6 "Deployment", it should be added that deployment of new AI systems often requires adding integrations which process existing data in new ways and/or combine data sources in new ways, as these may have ethical implications* (G7).

Similarly, in Section 4, there is a need to add detail, especially detail on specific risks that are not mentioned:

> *Section 4.2 is too brief […] most cybersecurity issues have no particular dependence on AI development, but there are some AI-related risks that should be mentioned: Proposal 14: It should be ensured that the deployed model is not leaking sensitive information about the training dataset* (G7).

We should mention here that, although it was considered important to add detail to incorporate additional risks, there are opinions that some of the requirements mentioned in Section 4 are '*redundant or vague*'. An example of this is identified in proposal 43, where the need to evaluate '*trade-offs*' is detailed, but '*a development team will always be making trade-offs […] and can therefore always claim to meet this proposal'* (G8).

Continuing with Section 5, comments highlight the need to clarify section 5.3, because:

> *explainable AI systems are very capable, and even "black-box" algorithms can be forced to reveal some degree of explanation by querying the model with strategically modified input parameters and studying how that changes the outcome* (G7).

This leads to a proposed addition to the text that '*any prescriptive decisions provided by an AI model should be accompanied by an explanation whenever possible'* (G7).

There is also a suggestion for a '*clearer structure*' of Section 5, which '*could help the reader navigate these easier'* (G7). Finally, there is also some scepticism on whether Section 5 should be removed from the main document and kept as '*an appendix to a guideline that is reviewed periodically*' (G1). The suggestion is based on the fact that the specific section deals with topic-specific examples, which could be updated as technology evolves.

Next, we move away from specific sections and record some of the general suggestions proposed by our interviewees. Some of these suggestions have to do with specific potential risks that were not tackled by the proposed Guidelines:

> *Because […] when an attacker gets in, it's very often through unpatched servers or something. So, that's definitely something that should, at least, be somewhere mentioned* (G4).

> *One viewpoint perhaps not sufficiently addressed is that the potential harm from collecting personal data is not only to the person but to society* (G8).

The issue of potential harm from collecting personal data is reinforced by other interviewees:

> *sometimes it's just you click through and click through and click through and you don't find, actually, where you can deactivate the cookies. Sometimes it's very nice, you can have the top, "Reject all". I think that might be something where the regulator might be a bit - if you could have the very easy opt-out* (G4).

Finally, a general suggestion was made on motivating implementation of these Guidelines by linking them to the decision-making process in an organisation:

> *In the decision support system, I mean, if you talk about organisational decision making, I mean you could write something about binding it to, for example, to role for production toll gates […] Usually in big companies you have checklists what has to be ready or not ready. If you bind it to that one, it's very easy and it' s part of the process. It's not something that you have to do on [the] side* (G4).

3.8. Evaluation of the practicality of the Guidelines

This section is based on conclusions drawn from the guidelines-related interviews. It explores **the practicality potential** of the Guidelines, which was one of the main goals set for developing them. As a general reaction, the Guidelines are considered sufficiently practical, with a good level of detail, and examples:

> *So yes, really good job and as I say, having been involved in developing guidelines on previous projects and looking at numerous sets of guidelines, I think these have been addressed really well. They've been well chosen, well explained, detail given, examples given where necessary. I'm very impressed* (G3).

> *I don´t know if we could take this and implement it as it is, but I think this is easier to implement as it is than what the EU has published so far … Clearly there seems to be more thought in this in the sense how an organisation might function and who the appropriate stakeholders in a company might be for some of these different things and I really like that* (G5).

The practical elements identified are specific sections, mostly Section 4 and Section 5:

> *section 4.2 for like accuracy, reliability and effectiveness of the system, or, you know the reproducibility and follow up, even the bullet points under that, those are written in a way that would be much more actionable than anything I have seen* (G5).

> *section 5, […] practitioners would usually then only look for their section that applies to them. Let's say if I was working in medical AI then I would read only that section […] I think it's good that there are sections for different disciplines so that everyone could find whatever is interesting to them* (G2).

> *I really like the special topics for considerations actions. I think that consolidates a lot of different worn off conversations that I have seen on, you know, mass surveillance systems or on other topics* (G5).

Some specific suggestions to improve practicality include adding perspectives of specific professional roles, as well as specific post-development phases of the technology's lifecycle that are not captured by the current Guidelines:

When I start reading the document, I don't have a perception that my role is understood and that my daily challenges are understood. What I would like to see instead or what would maybe benefit is that if I open the document and I see that this document is meant to be read and used by CTO or by Data Protection Officers or by this role in the organization (G1).

I think the thing that´s missing there that is AI-specific is that even after deployment there is a lot of life cycle management of AI models that is going to be very relevant for the AI ethics topic. So, for instance model retraining, […] the model results will change over time, the model bias might be introduced, or bias might be introduced into the model and so for the purposes of reproducibility, I think post-deployment is an entire step that is really missing (G5).

There are also suggestions for improving the Guidelines in terms of more practical ways of incorporating human rights and ethical values into development and use processes, but the challenge of this suggestions is understood by stakeholders as; 'the difficulty in implementing this is a universal difficulty in implementing AI ethics. When you are trying to take an ethical concept and make it a business reality there is going to be a lot of problems' (G5).

3.9. Summarising feedback from the *Guidelines Interviews*

The feedback from the *Guidelines Interviews* has been organised according to different aspects.

Primarily, the interviewees were asked to focus on the general 'feel' of the documents, particularly on clarity, engagement, presentation, operability, and, usefulness. Regarding clarity, there were mostly encouraging comments, with some suggestions on improving the documents by providing specific clarifications. Feedback on whether the documents were engaging varied between types of stakeholders, but the interviewees proposed mitigation approaches, including the use of a hierarchical digital platform as the medium for interacting with the Guidelines; a suggestion repeated multiple times. In terms of presentation, the interviewees did not reach a consensus regarding presentation elements, such as graphics, but reinforced the suggestion that presenting the Guidelines in a digital, interactive format would be an improvement. In terms of operability, the interviewees highlighted the need for both sets of guidelines, i.e. for development and for organisational use, and provided suggestions for improving this aspect mainly for the development guidelines by adding more actionable items. Finally, in terms of "*feel*", the reactions of the interviewees in terms of the usefulness of the Guidelines were documented. The consensus was that the Guidelines are equally useful for organisational users and for developers. Usefulness is the dimension of the first aspect investigated.

Additional aspects included comparison of the Guidelines with other relevant document with positive overall feedback, and, specific suggestion for improving the document sections. Finally, feedback sought aimed to cover the practicality of the documents and the potential for implementation. According to the interviewees, the Guidelines are considered sufficiently practical, with a good level of detail, and a good use of examples.

Even with the identified challenges and proposed updates to the two Guidelines documents, the overall reactions and comments are positive, especially on the practical nature of the Guidelines, which was one of the main goals of this task. Nevertheless, a number of suggestions have been picked up for further improving the two sets of Guidelines.

4. Analysis of Findings: Regulatory Options/Terms of Reference for New Regulator Interviews

This section presents the analysis of the Regulatory interviews, where participants were asked to comment on a proposed new EU regulator and propose potential approaches. Thus, the analysis primarily collects the regulatory issues identified in the interviews, and then proceeds to highlight the aspects that, according to the participants' feedback, could be the basis for successful EU regulation. The Regulatory interviews targeted mainly legal professionals and policy makers.

4.1. Potential Challenges for Regulating SIS

Before discussing the challenges in regulating SIS with the interviewees, we asked whether there is a need for an EU Regulator for AI. There was majority agreement that there is a need for a new AI regulator at an EU level, to address '*the increasing opportunities of companies to work with new technologies*' (R1). However, some interviewees felt that a new AI regulator is not needed since existing regulation already deals with the various aspects that the proposed AI regulator would be dealing with, i.e. '*IT developments are already regulated by law…the same applies, for instance, regarding human rights, ethics and so forth*' (R3).

Since this proposal follows the GDPR application in Europe, there is the potential challenge of falling into similar pitfalls, so it is important to '*avoid mistakes as with GDPR application, where supervisory authorities in many cases lacked the funding and resources to properly apply the regulation*' (R1).

Another challenge for regulation is to clearly select what its definitions are, i.e. which technologies exactly should be regulated:

> *when we're talking of artificial intelligence, we're not actually talking about robots. We're talking about algorithms which have been created by humans, which form parts of AI, perhaps more complex system* (R7).

In terms of regulation, interviewees felt each specific aspect of AI should be clearly defined and separately considered for regulatory purposes:

> *AI is two words but if we look inside the concepts and practice there are plenty of different technological devices, systems, methods, applications and each of these is working in some specific area. Take the examples of image detection, pattern recognition. When we deal with case law and legislation, where there is a problem with natural language processing, it's a completely different world* (R2).

Definition of AI was seen as a first step towards transparency, which is a main requirement for any successful regulation:

People talk a lot about blackbox AI, or AI's blackbox, this is not true. AI itself has many different practices and models. Also the field itself is rapidly developing in terms of providing more transparency on how the models work. We should look critically at how we define AI in this sense (R3).

There is also lack of understanding on matters of accountability when it comes to regulating such technologies, including accountable persons, but also accountable products:

> *the person who's perhaps responsible is a person who's chosen that dataset without considering whether it has inbuilt biases. So, there's a human being there somewhere who can be liable and responsible […] The Product Liability definition for defective product is a product which does not perform with the safety that is expected by the general public or by the user* (R7).

Interviewees highlighted that regulation must also consider the value of democracy, and how bias can compromise that:

> *… it has an impact on democracy […] the ability of people to make decisions […] After the Cambridge Analytica case, […] it makes me really scared to think that people get very biased information …* (R4).

Feedback from interviewees further highlights that even when democracy is satisfied, i.e. when the majority is taken into account, there is still a risk of discriminating against the minority of the population, and relevant regulation must consider that challenge:

> *And even if the government is declaring fairness and non-discrimination, some minorities are going to be discriminated because of the definition, because they're a minority, because the rules are going to be taken by the majority* (R6).

Wellbeing of citizens is a primary concern, but also a challenge, since the bubble effect can affect citizens' digital reality and hence their perceptions regarding certain facts. If these have to do with medical data, for example, then wellbeing may be affected:

> *… then he went to Google and searched and he said, 'You know, I'm looking and searching on Google and everything here is saying that this drug is great even for kids without formally being diagnosed as ADHD'. And I thought that, well, maybe there are search results, but they apparently are pushed to, let's say, page 17, 27, 123, very down on the list, deep into the list. You have to be very patient in order to get the other data …* (R4).

The example given by this interviewee calls attention to one of several potential "traps" technology usage may cause. In addition to non-obvious challenges, interaction with technology often generates a more direct and visceral response from the users. Therefore, wellbeing also has to do with a positive user experience overall when interacting with technology, and a challenge for EU regulation is to find the balance between regulation to remove threats and challenges for wellbeing without harming citizens' positive experience with technology:

> *At the EU-level, we shouldn't only think in terms of how to make things more impossible to do, but how we can support uptake in a human-friendly way i.e., provide transparency for citizens. In the end it is their data that is being used* (R3).

In summary, stakeholders have identified several challenges that a potential EU regulator for AI would need to address. Such challenges include, but are not limited to: avoiding repetitions of errors from past application of regulation across the EU, such as the introduction of GDPR; ensuring common understanding of AI across the EU; resolving issues of accountability; safeguarding democracy while at the same time ensuring fairness and non-discrimination; promoting wellbeing through direct interaction of citizens with technology but also by safeguarding against potential technology "*traps*".

4.2. Opportunities for successful EU regulation

Having identified some of the challenges, the interviewees discussed their opinions and suggestions for successful EU regulation.

One opportunity identified, which could potentially also become a challenge is that '*every country has their own [regulatory] structure*' and pushing the regulatory decision-making power to national levels may be an opportunity for success, i.e. '*every country should have their own final say*' (R3). Even if AI is regulated at the EU level, there still needs to be active involvement of authorities at national levels, to effectively regulate in the Member States.

National strategies are often driven by innovation goals and efforts to boost their economy through technological innovation. The opportunity for innovation and economic growth is very important to motivate for AI regulation, such that *responsible* innovation is supported and safeguarded:

> *[AI] national strategy document is not very different in the way of thinking to an economic strategy document […] AI is considered there as a means of production and raising production outputs…* (R6).

Another opportunity is making use of past experience by considering existing regulation for technology, such as the application of GDPR and lessons learned so far across Europe. GDPR provides a plateau to which to link the new AI regulation, as '*one complements the other*' and if they are somehow linked '*enforcement would be improved*' as long as there exists '*a specific legal basis*' (R1). The opportunity of past experience in regulating technology is also highlighted by other interviewees, e.g. '*we can have some guidelines based on past experiences*' (R5).

In terms of understanding AI, the challenge of defining this was identified by the interviewees. This challenge brings forth the opportunity to utilise expertise in the area, e.g. to put forth '*a body of experts on different real issues*', in particular, referring to technological issues and their effects on society, which is an ongoing challenge in regulating technology, since '*sometimes the Commission, Parliament and other institutions find it difficult to understand what the technical point we are discussing is. An authoritative body of people that is able to advise the institutions might be helpful*' (R2).

Additionally, with the support of the experts, the regulation of AI provides a great opportunity to look into aspects of technology that '*could enjoy stricter regulation'*, such as for example '*bias and discrimination by the [AI] algorithms*' (R4), impacting society positively in terms of ethics and human rights.

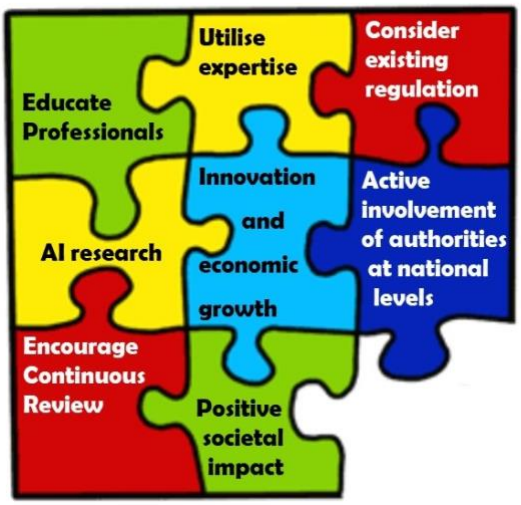
Moreover, the evolving nature of technology, and the ongoing learning and understanding that needs to take place to adequately address key issues as part of regulation is recognised by interviewees. However, this presents the opportunity for continuous review and improvement, and the responsible regulatory bodies at both EU and national levels should be '*open to the periodic review process, because the gaps are often there and are discovered during something that we conduct research upon'* (R5). The review process may bring forth the need to change or update certain roles or processes, therefore, the success of the regulator is dependent upon the body's flexibility to adapt to such review recommendations, i.e. *It depends on how the Board will start to function, its flexibility*' (R3).

Finally, the novel nature of the technology and its application and use in society, provides an opportunity for education; whether this is more formal with appropriate certification, or informal in the form of seminars by the national regulating bodies, it is essential that professionals who will be developing or using AI technology have an opportunity to be educated:

> *at the business level… the [regulatory] body should be able to go out, organise seminars, explain terms and conditions, rules, etc. and should be able to address society as a whole, i.e. unions, business networks and citizens* (R1).

3.9. Summarising feedback from the *Regulatory Interviews*

A number of challenges have been identified by the interviewees including accountability issues, issues of democracy, fairness and non-discrimination, need for ensuring transparency, positive user-experience and the wellbeing of citizens.



At the same time, the interviewees considered, in addition to the potential challenges of regulating AI, a number of opportunities for positive change, including positive societal impact, more research and innovation, economic growth, better educational opportunities for professionals, more active involvement by the national authorities, etc.

In summary, the challenges that a new AI regulator needs to face bring forth many opportunities for positive societal, economic, and technological impact. These opportunities, according to the feedback from the interviews, are graphically presented in Figure 3.

Figure 3. Opportunities for successful EU regulation

5. Analysis of Findings: Exploratory Interviews

This section presents the findings that emerged from the *Exploratory* interviews with stakeholders. Interviewees were asked to give their insights on what they consider as the main ethical issues that come out of AI and Big Data, and how they intersect with human rights; their feedback on current ways of addressing these issues, and their suggestions on how these ethical issues can be best addressed in the future. As 5 out of the 10 questions involved suggestions for addressing ethical issues in the future, more space has been given to these related themes.

Eleven ethical issues regarding AI and Big Data emerged from the analysis of the interviews (see Table 3 below). One issue which cut across several of the eleven themes discussed below is the intersection of ethical issues with human rights and democracy. Even though interviewees were not explicitly asked about the relation between SIS and democracy they often framed several ethical concerns within the framework of human rights, the need to safeguard them and prevent violations that weakened the democratic institutions and processes within a society. Human rights were not only used to frame ethical issues, but as one interviewee put it: '**ethical issues forced us to rethink the dimensions of human rights and rearticulate them in new ways**' (E7). Therefore, in light of the ethical challenges regarding how Big Data and AI technologies work, interesting conceptual issues and dilemmas emerged about '**how to rethink human rights**' (E7) that did not exist before we were confronted with dilemmas related to, for instance, understanding automated decision-making that has serious consequences for human life and well-being.

As the majority of interviews took place during the Covid-19 pandemic, there were references to this topic in around a quarter of the interviews (5/21). These were included in the findings when they were relevant to the main themes that emerged.

Table 3. Eleven ethical issues (themes) of AI and Big Data emerging from the analysis of interviews

No.	Ethical Issue
1.	Lack of transparency regarding who owns the data, how it is used and for which purpose
2.	Lack of adequate public information and understanding of ethical issues
3.	Biased data, algorithmic bias and lack of critical thinking
4.	Loss of human agency, dignity, autonomy and intervention
5.	Failure to recognise and protect privacy as a basic human right
6.	Surveillance, manipulation and coercion
7.	Prioritisation of financial over ethical interests
8.	Lack of accountability and product liability
9.	Loss of human jobs and mistreatment of employees
10.	Impact of Big Data and AI on health and the environment
11.	Exacerbating inequalities within and between countries

5.1. Eleven Ethical Issues related to Big Data and AI

5.1.1. Lack of transparency: who owns the data, how is it used and for which purpose?

One of the most prominent concerns expressed by the interviewees was the lack of transparency with regards to Big Data and AI. Lack of transparency was discussed in depth and in the following ways:

- how and why (algorithmic) decisions are taken
- who has access to the data
- which purpose the data is going to be used for
- the implications of the data collection, use and storage on the user

Interviewees argued that what was needed was:

> **transparency with regards to who is getting the data, who is using the data and for which purpose will the data be used.** I would say that is almost **fully hidden**. Can you think about why some people want all your context from your cell phone? You don't get a real idea on that and I think it would be appropriate to say that we need this and that, or for example, we need to be able to use your camera or your microphone **for this and that purpose** (E20).

Lack of transparency, it was argued, was particularly worrying given ***algorithmic decision-making***. '**Not having an explanation for why a decision was made**' was seen as a fundamental ethical issue, '**especially when AI algorithms are used to make decisions about people**' (E18) such as whether they are imprisoned, given permission to act upon something, being classified as fit for a particular job, or when one's data was used for commercial reasons such as for advertising (Interview No.13). As one interviewee put it:

> It is not that when you have a human you always get the answers. But having just the algorithms spit out an answer and not even the operators being able to say **why or what was the weighting of the different characteristics**, **that's awful** (E18).

Transparency was seen as going hand-in-hand with explainability and informed consent: for something to become understandable to the public, accessible information needs to be provided that is explained clearly and adequately '**in a way that makes people actually understand how the technology works**' (E21). It was important to '**make the processes transparent so that not only experts but also all people understand**' (E2). Otherwise, app developers could be '**misleading people**', and this is especially important with regards to vulnerable parts of the population such as children (E20). Lack of explainability and transparency, interviewees argued, meant a violation of the human right to have access to information regarding technology that may affect an individual's life, as well as weakening democracy:

> Humans should have the right to understand why a decision concerning them was made and how it was made…If no one knows how and you can't look into the algorithm and just say "Ah, okay they took this, this and these features and this feature got 20%, and this feature was weighted more, and we can change that". You can't do this. And once you can't do this, then I think it's a very big problem, [to] human rights, and to democracy in general (E18).

> It can be debated the extent to which there is *any* **human liberty** or *any* **freedom online**, especially with the lack of **algorithmic transparency** (E13, emphasis added).

Here, it was seen as the responsibility of companies, especially Big Tech[4] to ensure that there is transparency, and that when data is used it is not only always with people's consent, but also that this consent is given after adequate information has been provided in order for the user to be able to make an ***informed decision***. Otherwise, people would not be able to make informed judgements of whether they want to use a particular application or not:

> you often don't know where the things come from... Again, to make a **judgment** if it's okay to use something like that, it would be helpful [to] have at least an idea **who invented it, for which purpose** et' (E20).

> people have the **right to know** who has their data and what they're doing with it. But I think they have the right …not to just give permission but to report and **consent** the use of their data and I think **one of the big ethical issues is around the extent to which people are able to do this** (E4).

There was an acknowledgement that there are efforts for greater transparency in social media platforms, for instance by Facebook, especially after the Cambridge-Analytica scandal,[5] but that there

[4] The phrase 'Big Tech' was used by many participants; this term refers to the biggest, most dominant companies in the information technology industry.

[5] The Facebook–Cambridge Analytica data breach refers to the harvesting of millions of Facebook users' personal data by Cambridge Analytica, without their consent, mostly for political advertising. Data collection began in 2014 and was officially disclosed by a former Cambridge Analytica employee in 2018. It is the largest known leak in Facebook history. The participant in this quote was specifically referring to the US 2016 election scandal where data was harvested to promote support for presidential candidate Donald Trump at the

was still a lack of understanding by the '**majority of people**' on how transparency works and it was still hard to '**get people to engage and pay attention**' to such matters (E13). Therefore, what emerges is a dual responsibility; that of the companies to provide the information, but also that of the user to be able to take initiative, '**meaningfully engage with these technologies**' (E12), and as much as possible educate themselves in order to be able to safeguard their best interests.

5.1.2. Lack of adequate public information and understanding of ethical issues: 'information asymmetry'

As discussed in the previous section, lack of transparency was seen as directly connected with lack of adequate, clear and accessible information. This was again discussed both in terms of the responsibility of the developer or company as well as the responsibility of the user, though more emphasis and weight was usually given to the former. This was particularly the case given what one interviewee called the **'information asymmetry'** that exists between developers/Big Tech and users/the public. The ethical questions surrounding these issues need:

> more and more attention, like just to fight against **the information asymmetry** between what people think they know and what is actually happening... a good example of that recently [is] with Alexa and other like Google Home and other assistants, and there will be a scandal about the fact that people didn't know that these devices were actually recording more than they were supposed to. And this is something that for people who are working in the **AI industry,** that sounds pretty like normal because this is a feedback to improve algorithms themselves. But let's say **the majority of the people didn't know that and I think we should be much more educated on these issues** (E10).

In other words, it was not just the ***responsibility of the company*** to educate the user, for instance on '*how AI works or how data mining works*' (E15) but it was also a basic ***human right of the user*** to be provided with adequate information and understanding of (un)ethical implications of issues related to Big Data and AI. This was applicable to '**any area - AI in healthcare and police and military**' for example (E12).

As for the theme above (5.1.1), interviewees pointed out that some progress has been made, but that this was still rather limited. Aspects '**like data collection and the question of freeform consents**' have, they argued, been given some attention, but this was '**not enough**' (E10):

> So now people are more and more aware of cookies on websites. But I think that they're still not able to make a distinction between one website and another. Or you know, on some websites you have like very basic cookies, just like tracking to know which pages attract the more attention and on other platforms you have much more developed systems with a sense of, for instance, what we call prices discrimination strategies combination systems, which is like dynamic pricing when you go and try to get a plane ticket or booking for something (E10).

Inadequate information also remained an ethical issue when it came to ***data ownership***, i.e. who ultimately owns the data. Again, although interviewees acknowledged current attempts to provide a

expense of Hilary Clinton. Further information can be found here:
https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html

clearer definition through GDPR (General Data Protection Regulation), there was still a big gap in practice, as people did not properly understand the ways in which their data was being used:

> So even though data ownership, I mean, even though there have been some attempts to kind of make data, know **who owns the data**, the data is yours, you know, GDPR has tried to do that, I don't think that a lot of people understand really **how big data is being used** (E13).

Interviewees also pointed out that the question of data property is intertwined with the question of human rights, in the sense that it was a fundamental human right to own your data (E10), but also to be able to protect your own interests when personal data was being monetized by companies:

> I'm a big advocate that you know – you are the **owner** of your own data. And that's where data about yourself or data that you are collecting that you have for a project, that somehow you have ownership of that data because it's yours or because you decided to collect it in a particular context. And as soon as other people start to take that data and **start to monetize it and utilize it, maybe against your best interests, then I think this is a big ethical issue** (E4).

Information asymmetries can be traced down to **power asymmetries** given these three factors, a) the position of power Big Tech companies have due to their knowledge, expertise and technical capabilities, b) poor, inadequate or inefficient legal frameworks which allowed them to behave in unethical ways, and c) on behalf of the user there was to some extent a '**lack of motivation and time in order to understand how decisions are being made'** (E18).

A fourth factor that was identified to greatly and unprecedentedly exacerbate the power asymmetry is if governments combine their power with private companies such as Big Tech:

> there is now this extra possibility of **huge power gap** if the state power combines with the private companies…because of the technologies that we are using…now the private companies have **all of this information about us** that other agents have not had before. Also, if and when they collaborate intimately with governments, then, basically, the **two large powers become merged, you know, that state's power of using a monopoly on coercion and violence**. If it combines with the private companies' information about us, that is possibly an **unprecedented** level of powers in our society (E12).

Ultimately, the effect perceived here was a democratic deficit, as citizens cannot protest about or strive to protect something they do not understand or are not aware exists in the first place. It means having:

> the potential to move people **away from democracy** basically…you have another step, another very **high wall that prevents people from actually criticizing the government** and the **decision-making** processes or whoever is taking decisions on their behalf (E18).

5.1.3. 'Maybe the machine is wrong': biased data, algorithmic bias and lack of critical thinking

However, more access to information and more transparency would not simply solve the ethical issues at stake here; interviewees noted that what was also lacking was a ***critical approach*** to the information provided to them, regarding data collection, data management, how decisions were made etc. It was important, they noted, not to blindly trust the SIS and assume that all data are accurate, unbiased and reflect '**the truth**':

> I think another problem is that many people would just look at the answer they get as like **the truth**. Like that the result is equal to the truth, and I don't think there will be enough place for **critically thinking** about **what's the uncertainty. Maybe the machine is wrong.** I think all come to the same issue. Being able to **critically look at how the decision was made** and fix it (E18).

> the problem is that we tend to have this **wrong sense of objectivity** because it's a machine and not a human being (E9).

An argument that emerged was that when these three worked together - transparency, access to adequate information and critical informed approaches to both the data given by automated systems and to decisions made by users - the result would be a more ethical approach to SIS. More often, there was an explicit link made between transparency and critical approaches directly reducing the ***possibility of bias and injustices***, but also the ability to ***identify*** it in the first place (e.g. E5; E8; E16; E18):

> a huge discussion is the black box...a **problem of opacity** with algorithms…how are we able to explain if the eye is making a **prejudice**, is being **racist**, if even the programmer cannot state **how exactly** this algorithm came into this decision (E5).

Biased data was seen as '**a major ethical issue'** (E19) by the majority of the interviewees. Some participants noted how there was inequity in terms of both the ***quantity*** as well as the ***quality*** of data available '*for underrepresented groups and for vulnerable populations*' (E19). They spoke of the lack of diversification, accuracy and reliability of the data, and of '**discrimination in data**' such as '**gender bias**' and '**minority bias**' (E9 and E13), or not taking into account people with disabilities (E4). People were seen as '**bringing these biases to autonomous systems because [they] are training them**' (E16). As a result, machines were making '**serious ethical mistakes**' and exacerbating stereotypes (E16). One participant, when referring to Big Data, noted how '**it's easier than ever to ingenerate fake news, fake content**' whilst checking '**the veracity**' of information was a rather difficult task and '**a major stumbling block**' for the ethical application of Big Data (E19).

Biased data was also seen to refer to the **i*mplications*** of algorithmic bias and biased data on society and human rights: it worsened inequalities, discrimination, prejudices and injustices and led to errors as well as problematic policies. It meant, for instance, that when there were improved services offered to society, some groups who were better represented in the data, benefited more than those who were not:

> data that has been collected to a larger extent from certain groups than from others. Therefore, all those potential positive effects of **personalization and improved services only apply to those whose data is best represented** (E19).

> If you are making biased decisions, then it leads to systems that are **not equal** and are not **inclusive**…you end up developing algorithms which are for the **majority** and then they don't serve the **minority** (E4).

Interviewee E4, from their experience in learning analytics and developing algorithms that were designed to help failing students, talked about how '**most of the models were good at recognising the typical student, but they didn't recognise somebody that had a different learning pattern, for example because they had a disability or were having accessibility issues**' (E4), thereby creating issues of inequality through exclusion.

Another interviewee referred to the issue of algorithmic bias as a type of '**algorithmic governmentality**', due to the capacity of algorithms to give inaccurate, poor quality and biased data that were detrimental to the life and well-being of an individual (E10). Examples included credit scoring in the US and China which meant people were refused access to certain services or declined a job, '**not because of their resume or their past experience**' but due to '**very poor quality data…sometimes not even referring to the right person**' (E10). Another example was of algorithmic discrimination: '**discrimination of facial recognition…that doesn't recognize black faces as well as it does white faces**' (E11). Here, an interviewee mentioned a specific example from a Dutch welfare fraud case where:

> the lower court in the Hague found that the Dutch government's use of an algorithm to sort of predict the likelihood that people were cheating on social security benefits, […] was contrary to the **European Commission of Human Rights** because it didn't meet the proportionality test (E11).

In some cases, it was considered that algorithmic biases meant people were being unfairly treated by the **justice system** due to sexism or racism:

> data can **privilege** some people but worse than that can **de-privilege** others… facial recognition…Obviously, they've been trained by the white male and it doesn't do so well with **women** and that it doesn't do so well with **people of colour**…there's actually been cases where black people have been wrongly misidentified as criminals and this is a **huge ethical issue** because…first of all you don't always have people who are applying **human judgement** in these situations, especially if A.I. becomes more prevalent and relied upon. You know it could be that people get into the system and then **they can't get out of the system** based on a misrepresentation because of bias in the way that…data sets are being developed and then utilised for their algorithms (E4).

Biased data was seen as a consequence of certain groups not having ***equal access*** to devices that collected the data, but more importantly as a result of bias in the ***design stage*** or bias in ***data collection*** processes. The process of data collection entailed certain ***human or automated decisions*** of inclusion and exclusion, reflecting biases that are dominant not only in the current period of '**fake news**' (E19), but sometimes reflecting ***longer-term patterns of discrimination*** that are remnants from many years ago and still present in certain datasets:

> many of these systems rely on much older data sets that have been collected even before such a thing as informed consent around data sharing was developed…in some cases, these data sets go back a century and that means that they **carry with them all of the societal biases of the past century.** And we know that they have been quite a lot over the century, but they have been slowly dismantled and reduced, but then now resurface in the new uses of these datasets (E19).

5.1.4. Individuals vs. Machines: loss of human agency, dignity, autonomy and intervention

We discussed in Section 5.1.1. how the lack of transparency enabled automated algorithmic decision-making that was problematic, to go unnoticed. We have also examined in the previous section, how,

according to interviewees, automated decision-making could lead to unrecognised discrimination, not least because of biased data and an uncritical approach that treats data produced from SIS as '**objective**' and as the unbiased '**truth**'. One other important implication of automated decision-making pointed out by the interviewees was the ***loss of human agency and autonomy*** to machines. Individuals were seen as losing both '**control of their data**', and the ability to act and make choices without external parties making decisions for them, and this was seen as a serious ethical issue that needed to be addressed (E1). There was a call for '**keeping humans in the loop, especially for decision-making processes**' (E10), and to halt the '**loss of respect for human dignity, because the value from machine intelligence is given too much value**' (E15 and E6). Interviewees spoke about the loss of human agency because their ability to choose or make decisions has now been taken over by SIS or '**has been predetermined**' for the user:

So, you know, I think I'm selecting a link. But…that link has been curated to appear on that feed…at the top of the page so that I'm most likely to click on it. So is that **freedom** or is it not? So I think that's also another way that it links to **human rights**. And most importantly, and I think this is one of the big reasons why this needs more regulation, is that it really impacts **democracy**, so, and the whole democratic process. So, you know, because it also affects like **human agency**, I think I have the **choice**. But then maybe I don't have that choice because **that choice has been predetermined** for me. And that becomes especially important, even when we think about things like **targeted advertising**. I mean, you know, the whole scandal with the US 2016 election (E13).

The participants in the interviews expressed deep concerns about the way AI and Big Data shift autonomy and agency away from the individual to either machines or institutions: it meant '**relinquishing human control**' which could have '**an impact on our autonomy**' (E9). Ultimately, this was about loss of both ***power and control***. So, on the one hand human agency and autonomy was lost to machines and robots, and on the other, individual power was being manipulated for political or financial reasons. This required a change of perspectives, being able to also view chatbots, AI and robots as ***actors*** with influence, albeit non-human ones:

> we have the right answers to think about weak AI, but we don't have the right answers to talk about general artificial intelligence and autonomy. Since the **robots are going to interact and influence our behavior, we need to be able to look to them as social actors**...[to] look in a more clear way at internal human actors and say how they interfere in our behavior, how they interfere in this interconnected network. And this is applicable, very much applicable to the online scenario with chatbots for instance…So we need to be able to look into those nonhuman actors, and law has a lot to learn with anthropology, sociology, but also philosophy...' (E5).

Yet, as participants put it, this is a field of interaction where sometimes it is impossible to fully know the effect of '**non-human agency**' (E8) on humans:

> we're not sure how much of an AI input has formed our beliefs, our attitudes, our decision making and so on (E8).

> Because when we talk about **deep learning**, it's very hard to state **where the decision of the robot or algorithm came from**. So, this **autonomy** on the decision making of robots and algorithms is going to create a huge challenge also for **law**, how to create this **liability** in this context (E5).

The latter quote raised two related issues: that of legitimacy and liability, especially in pursuit of justice and legal matters. Is it ethical for a human to take responsibility for an act, based on data whose legitimacy is questionable and whose decision-making processes are difficult to both trace and contest? As one interviewee put it '**who has the legitimacy to take [a] decision, is an algorithm legitimate enough to take specific decisions for me?**' (E10) This was discussed again in the context of algorithmic governmentality and the risk that it was posing to ***human dignity*** and the ability of a human to be '**claiming for justice**':

> if algorithms are taking decisions for us, we need to know when, why and how and especially how we can change it if something's going wrong, **how we can change data** about us, as it can lead to significant **negative** consequences for us (E10).

There were references by interviewees to the importance of the '*right to challenge decisions*' (E11), and that this ***right to resistance, this human intervention*** is being undermined by the loss of human agency. One interviewee, in particular, spoke at length about the '**right to be considered innocent until proven guilty**' in relation to legal matters. They argued that AI is:

> certainly part of the decision making process and is diminishing that right to be considered innocent, especially if the A.I. is given to much credence…because quite often artificial intelligence is put there so that there can be less human input (E4).

One's right to due process and a fair trial were seen as a fundamental human right. Conversely in the case of '**no human intervention…these things can just get into a system and it can really affect your human rights**', for instance, when an innocent person cannot prove that the algorithm has misidentified them as guilty due to '**a faulty**' or '**a biased training set**' (E4).

5.1.5. Failure to recognise and protect privacy as a basic human right

The ethical issue related to privacy was not only a common theme across the interviews, but one predominantly presented as a fundamental human right. Concerns were raised not only because of the shallow understanding of the dangers related to the lack, loss or threat to privacy, but also to the fact that people were not '**aware that privacy is a right'** and therefore not *protecting* '**the very normal right that is privacy**' (E20). One interviewee argued that often:

> people don't care because they have this idea that "I don't have to hide anything", but I would say that is not about if you have to hide something or not, but if you have **rights** or not. You know, privacy it's a right that you have and although there is no need to hide something, it's still that you have the right that something stays as private as possible. I think people don't really **understand** the issue and they are not really very much **taught** in understanding that (E20).

Some participants pointed out the ***multi-faceted*** nature of privacy and the importance of appreciating and protecting these dimensions. Privacy, one interviewee argued, should be seen on the one hand as '**a collective right**', where digital technology affected '**our environment and health**' collectively, but also as an individual right which impacts both '**individual physical and psychological integrity**' (E3). Another interviewee spoke of '**a broad perspective**' that took into account:

> not just informational privacy, data protection, but…physical privacy, mental privacy and decisional privacy as well...It's basically the **impact on our private lives**, on how we live our lives as individuals, not just our data (E9).

In the context of the current pandemic, there was a differentiation made between giving up privacy '**for protecting me and others**' (so for health reasons), versus giving away information to enhance targeted advertising – the focus here was that either way the user should be given the **choice** to decide (E20).

The way in which ***different countries*** responded to the pandemic also pointed out country differences in respecting and recognising privacy as a human right. One interviewee, pointing to '**China and Korea**' argued that the human right to privacy '**is just not being recognised in some places**' and that even talking about this as a topic '**is really hard**' (E1).

One interviewee pointed out the impact of the lack of privacy on children, and how parents' behaviour was affecting their children in unpredictable ways, either due to their own lack of understanding or factors outside their control:

> A lot of **parents don't understand**, like, the **digital footprint** that they are leaving behind for their kids. And what that could mean…in decades from now, like, who knows what the future holds…even if I did have a kid and I'm pretty data conscious, there's **no way I can protect my kids' privacy** because all of my communication which happens either through my computer or my phone is mediated through **third party applications**. And that data might be mine, but it's on someone else's server. So there's no way to protect young kids, you know, privacy. And that's a basic human right (E13).

The common use of the word '**protection**' vis-à-vis loss of privacy would suggest an emphasis on security and/or safety. However, only two participants mentioned these two aspects together with privacy (E7 and E4), and E7 did not expand on this. The argument made by participant E4 was that privacy, safety and '**personal security**' is compromised when AI is being developed, as this process:

> often requires a lot of data – you know quite **personal data** about people that they don't want to get into the public domain and then they often are because there aren't enough **safety policies** around the uses of this data and then there are breaches and then that can compromise people's **personal security** (E4).

5.1.6. Surveillance, manipulation and coercion

Loss of privacy was directly connected to the theme of surveillance, manipulation and coercion. The argument made here was that the negative imposition on human agency was so extensive that it ultimately signified manipulation of users. Surveillance signified a direct violation of the right to privacy. Both were viewed as detrimental to human dignity (E10, see also Section 5.1.4 above) and democracy (see also Sections 5.1.1, 5.1.2 and 5.1.4 above). Interviewees spoke about users being '**surveilled**' or '**manipulated**', about being '**nudged**' in particular directions, for either **political or financial gains**.

A similar concern to that of privacy emerged, where interviewees were concerned about '**government surveillance**' (E15), and users not finding CCTV cameras and surveillance worrisome: '**the discourse which is developed around surveillance is very much troublesome**' given that technologies are intrusive and yet people do not take this as seriously as they should. This went back again to the

perceived problematic rationale of '**why would you care about surveillance if you don't have anything to hide?**' (E10).

One interviewee mentioned the term **'surveillance capitalism'**, in the context of personal data being commodified for profit at the expense of people's personal freedom. This tension between **freedom and safety** was discussed, given that in the name of safety, people were being surveilled so extensively that one could talk about '**the return of potential authoritarianism'**. There was a call to take freedom '**seriously**' and not '**destroy any right**' that humans have in the name of safety (E10).

Interviewees also explicitly talked about the:

> …notion of **coercion**…. Obviously behavioural science can do with this time…to configure environments in a way that will maybe coerce people to do more shopping of this product or that product… But when we get into this kind of artificial intelligence and what can be done with the data, I think anyway, it is questionable how you're nudging people in these day-to-day scenarios…it really goes back to this issue of privacy… (E4).

Coercion was unsurprisingly linked to human rights abuses, and diminishing personal liberties negatively impacting the democratic process. It was argued that '**not enough**' is being done in regards to regulations on '**the coercive use of data**' and that even transparency regarding who has the data, and which data, does not automatically mean that people are '**fully aware**' of '**the psychological principles and aspects**' that are being adopted (E4). The same interviewee argued, one of the problems is that '**people are quite used in a way to being manipulated**' (E4).

So there was perceived manipulation of behaviour, e.g. voting behaviour for satisfying particular political agendas, while there was also monetisation of data for profit-making, and sometimes the two were interlinked. It was argued that politicians are often '**ruthless about this**' using AI to influence voting behaviours (E18). This also involved wider policies, going beyond voting behaviour, where '**dissemination of information is used to in some sense manipulate or nudge individual or social decisions**' like '**whether or not you are going to get vaccinations**' or '**what kind of advice you follow**' (E12).

Nudging behaviour in particular directions was also criticised as a strategy adopted by some companies in order to avoid contact with customers, prevent customers from leaving feedback, or in order to allow better access to users' data in covert ways. The latter point was particularly important given that a lot of users are not aware of the degree of data that is collected from a mobile app or website, '**or whether their geolocation is being used or not in default settings** (E10). Therefore:

> Some websites and some developers use negative/positive technologies to nudge negatively people and discourage them to change either the default settings or to be able to reach a call center. And I think this point is very important (E10).

One interviewee spoke about the direct relation between data having the ability to unethically influence the behaviour of users, with reference to the GDPR (General Data Protection Regulation):

> I absolutely love that the GDPR says that your data, your personal data is the extension of your person. And I absolutely think that's true, because just like if you push your physical body you fall over, if you push someone through their personal data **they tend to be influenced and they alter their behavior** (E6).

5.1.7. Prioritisation of financial over ethical interests

Monetisation of data was seen as an ethical issue when '**commercialisation of big dat**a' occurred, especially when this was done without adequate ***transparency***. The prioritization of making money over ethical issues has already been mentioned in the section above, with data often collected in ways the user is not aware of, and then monetized directly, or where data is used to ***manipulate*** the behaviour of the user into making particular decisions, e.g. specific purchases through targeted advertising. We have also seen in Section 5.1.1. how lack of transparency allows companies to monetise data at the expense of users' ***privacy***: '**Privacy is a very key point here. No one knows what happens to the data, is it sold, is it not?**' asked one interviewee (E16) while another noted how monetization of data is a big ethical issue when this is **'against'** people's **'interests'** (E4). These dimensions feed into the wider theme which is the tension that emerges between ethical issues and economic imperatives.

This tension applied also to the design and development of smart appliances, where due to a '**profit motive**', companies benefited not only much more than the individual, but ***to the expense of the individual***. An interviewee, speaking of smart washing machines, spoke about how:

> the technologists who are developing an AI system have such **a strong profit motive** to make it work and make it **not transparent**. I mean, I can only come up with an example of making a smart appliance. It is sold to the consumer as something powerful for the consumer, because now you know when you run out of detergent, for example. Or now your smart information, your smart washer can figure out what the weather is, so it knows the best time of day to run the wash in order to use the sun, which sounds very nice, but the reality is that a washing machine, a smart washing machine, which is more expensive than a regular non-connected washing machine **was really designed to benefit the corporation. Because it helps to gather data as well...it is designed to be far more beneficial to the corporation than it is to the individual.** And the fact is, the individual was actually going to pay the higher price, and it is going to be sold that 'this smart appliance is for your benefit because it can make sure it's not on at the same time as the dryer' or whatever it is. But the reality is that **it was developed for the benefit of the corporation** (E15).

The latter quote shows how prioritization of financial over ethical interests is a theme that intersects with transparency and therefore, with the ability of the user to make an informed choice. Continuing the discussion, this interviewee notes how important it is to give '**the whole story**' to the user, and not just selective information in order to manipulate customers into making a purchase:

> And so I think that if the consumers were told that, that will be a determining factor, the choice to purchase it. And **they are not told that.** They are told that this new technology is really cool because it could bring down the consumption levels and maybe that's true, but **it's not the whole story** (E15).

Some participants presented this financial-ethical tension as related to the ***customer needs vs. scientific innovation*** dilemma, and argued that it is an ethical issue that is not '**being addressed by many at all**' (E1). In other words, they felt there was a problematic focus on a business model that prioritized innovation and left out, or left behind what the customer might really require, and viewing the customer as an individual with human rights. Instead, what was really needed, the interviewee argued, was a business process that starts from customer needs and follows this '**through the ethics via design concept**' and then moves on to the innovation aspect.

5.1.8. Lack of accountability and product liability

The absence of moral behaviour and a moral responsibility by the companies, together with contextual conditions such as the legal and regulatory gaps, led to discussions regarding the lack of accountability and product liability in SIS. This theme was a common one, and there was a call for companies to become more responsible (including being more transparent) and accountable, but also for more structures to be put in place to deter and prevent malevolent behaviour that abused human rights:

> one of the big topics is the question concerning **responsibility and accountability**…And I think that those questions then break down into issues concerning transparency and explainability of AI. And that also then again entails fairness and the fair implementation of AI (E8)

> the intransparent (sic) character, which is also a bit of a culprit for lack of **accountability and responsibility**…. (E9).

The behaviour of companies (see Section 5.1.7. above) was often discussed within the wider context of ***legislation*** and ***regulation*** frameworks favouring big businesses or so-called Big Tech, again at the expense of the individual. Referring to the US context, one interviewee pointed out how '**the data legislation is really left open**', frameworks '**take a very business-centric perspective**' and that both these points '**favour business, rather than any individual**' (E1). The call for more liability i.e. responsibility to be entrenched in legislation, was discussed in several interviews (e.g. E5; E9; E11; E15). Part of the reason for both legislation and regulation frameworks operating favourably for businesses was put down to strong '**lobbying pressure**' satisfying the business model that was '**just so lucrative'** (for instance in Israel this came from telecom companies) (E3). Another reason was that countries are more concerned about the '**economy**' and being '**utilitarian**' rather than focusing on '**the importance of human dignity**' and deontological issues (E5).

Interviewees argued that companies should be made accountable for their actions, and called for more guidelines that '**put[s] the responsibility and accountability squarely on the shoulders of those who are profiting**' (E15). Gaps in legislation, standardization and practical difficulties in regulation were also identified:

> And then when it comes to **legal** gaps, for me, the biggest one is the gap of **transparency and accountability**. So since there is no wide **obligation** for, for instance, **documentation** of which data you use and how you use data, on which standardification (sic) methods you applied, it's very difficult for **regulators** to even just go and check whether or not existing law was respected, because there is no transparency on that. And that doesn't have much to do with the black box, because even for black box systems you can still have documentation of which data you use and in which matter (E9).

> **accountability** of private corporations…**I have the right** to know how is this product affecting my life. How is the company benefiting from my data? How can I use this product in a way that it was advertised to me? That is really **a product for me and not a product made to gather information about me**. And right now, I feel like there's **no regulation** on how honest a company really needs to be about that and so it leaves a lot of individuals **vulnerable** to being tricked, to being duped. Because it's not really saying like 'look at this great feature in this new app'. It's entirely developed to get information about you. It looks like it's for you, but it's not really. And I feel like **we shouldn't have to fight so hard** not to be tracked (E15).

The previous quote also serves to illustrate how ensuring product liability and preventing human rights being violated - for instance, the right to information - seemed for many to be a '**hard**' struggle, rather than something natural that is expected to be upheld in a democratic context. One interviewee spoke at length about the ***lack of customer service*** regarding a '**smart**' product that was bought, or a digital service that was used. The argument here was that there was no option for contacting a human being ***to explain data collection aspects***, and sometimes the information was so overwhelming and inaccessible to the average user that it rendered it '**meaningless**'. This happened both online and in street shops:

> I doubt very much that the sales person on the floor is **trained** to respond to questions about what data is really being collected here (E15)

> It's not thought of enough when they sell a product that uses AI or when they sell a connected product… there's **no option** on my Spotify that I can call up Spotify and talk to a **human being** and ask them 'can you explain to me, please, how you're picking songs for me'? No one's going to answer my questions. There's **no help desk** to help every single one of us when we click on a web site and were presented with an **impossible to read consent form**, consent for cookies on every single website we visit…with thousands of different forms they might as well not ask for consent at all, it's completely meaningless…I can't call up any particular company or website, and say I 'm sorry I can't understand your consent and terms, can you explain that to me? There's no **accountability** and because of that it's a **meaningless choice** (E15).

Ultimately, it remained a significant ethical issue for participants that companies were not made to account for their actions, for their obligation to customer service, and for users' right to have questions answered to be respected.

5.1.9. Loss of human jobs and mistreatment of employees

The core concern in this theme was the ethical issue that emerges when ***human jobs*** are lost and human expertise is replaced by machines, especially in the context of the financial crisis caused by the ***pandemic*** (E10). Only one interviewee argued that this was not an ethical issue and merely required '**reskilling of the human resources**' (E14). The rest of the interviewees who touched upon this issue saw the impact of unemployment as a serious ethical challenge:

> AI is very, very **disruptive** and it's very interesting for us to talk about this first big ethical and social problem because we are facing a **pandemic** right now that is putting this under the spotlight. So many companies are already **substituting human labour force for robots** right now in the middle of this pandemic (E5).

Interviewees spoke about the negative effects that this transition to machines and robots had on people's lives with '**all their personal life and income**' being dependent on their jobs (E).

In particular one interviewee discussed extensively the human impact, not only in terms of people becoming unemployed, but also the consequences on ***human rights*** in terms of how people who remain employed are put under constant pressure, and unrealistically compared to machines:

> There's also, of course, the loss of jobs…I use the example of Amazon as the company…suddenly humans are **being compared to robots** and say you can't keep up like a robot does…for the people who are the pickers, who pick the orders and select them and put them into the box…They are held to **inhuman almost standards** and they're going to

> eventually be compared to 'why isn't the performance as spot on or as regimented as a robot can do it'. And that is a really unfair comparison. **And it's really unfortunate that any human should be compared to the performance of a robot.** Even though financially, for those of us who buy from Amazon or are interested in my order gets here by tomorrow by 3 pm…it is a more important issue, that the people who work for Amazon are not treated like that and as a consumer it is perfectly reasonable that I have to wait three days for my order because the **human rights of the workers** come first (E15).

There was an acknowledgement that this issue was discussed in the AI community but that there was still '**no sound solution that makes people confident**' regarding these negative consequences of AI on the job market, and how to prepare the economy and the labour force for the disruption caused (E2; E5 and E8).

5.1.10. Impact of Big Data and AI on health and the environment

Although the theme of the impact of Big Data and AI on health and the environment was not very common across the interview cohort, it was discussed by three of the interviewees, two of whom spoke about it at great length (E12 and 3).

Regarding health, the impact was discussed firstly in terms of the impact of biased data on health care: '**If the tools are biased then you are creating discrimination in the right to healthcare, right to health…a systematic violation of these human rights**' (E12). Secondly, health was discussed in terms of the impact of electromagnetic waves and the '**low level radiation**' it emits. According to the interviewee who discussed this issue, and noted that this is aspect is absent from both the SHERPA ecosystem and in wider AI discussions:

> this entire AI infrastructure, all of it, relies not only on electricity which you mentioned…and the amount of pollution that is going to be necessary to produce AI. But no one mentions the waves themselves. AI does not work without the transmission. And the transition is radiation. It is low level radiation but radiation (E3).

The same individual discussed extensively the negative consequences they were concerned were occurring as a result of AI systems and wireless technology, aspects they claimed were often not discussed, due either to vested interests or the invisible nature of their impact:

> So there is an entire framework necessary to make wireless technology work and that is a framework that operates with electromagnetic waves. Electromagnetic waves are a form of radiation, a form of pollution. They will be naturally produced in our environment by the sun also, and yet, we have been spewing ever increasing amounts of electromagnetic radiation into the human environment since 1995. And because these waves are invisible we often do not think of **the impact on human health and on human immune systems**. And of course, independent scientists have produced many studies, thousands of studies that demonstrate that these waves may in fact be **dangerous** and the laboratories of the telecoms have also produced many studies that prove the opposite.
>
> And so under the idea of the principle of precaution we probably should err on the side of caution and remove electromagnetic wave **pollution** from the environments of those that are most vulnerable, the elderly, children, people in institutions where they are being treated for mental health issues etc. So the first issue is **health** and no one ever thinks of the

> infrastructure and the hard infrastructure first when they talk about technology, ethics and human rights (E3).

The issues of health and environment were linked by the interviewee, when referring to environmental pollution. An additional note was explicit reference to the recycling aspect of the hardware involved in AI:

> There is a lot of **hardware** involved in AI. I mean an enormous amount of hardware. Much more so than we have now with our current phase of digital technology. The amount of hardware will **double**. And so consequently the impacts of this hardware are going to be **enormous**. That plus the data storage plus the roll-out of 5G (E3).

5.1.11. Exacerbating inequalities within and between countries

The final ethical issue that emerged as a theme referred to the '**really concerning gap**' between countries and people i.e. the digital divide in terms of access to, use of and benefit from Big Data and AI. Firstly, there was concern regarding a **'knowledge gap'** between people – this referred to the difference between those who had a sufficient understanding of data mining and how AI worked, and those who did not understand SIS and just had to '**trust it'**:

> it's concerning that only a very tiny elite, highly educated, very intelligent group of people in the world understand smart information systems. And the rest of us, just…have to trust it. And I think that will only increase (E15).

There was a call to acknowledge the digital divide as a serious ethical issue, especially in vulnerable parts of the population, and particularly in today's digital era, as well as to move in a direction that bridges the gap and offers all humans an **equal right to access** scientific innovation and progress. Speaking specifically of the European context, one interviewee argued that:

> all European citizens through the international human rights treaties should have the **right to access scientific progress**…And I believe that access to digital technology and the removal of the **digital divide**, at least within Europe, is **absolutely critical, it is a critical ethical issue**. Because society is moving in a direction that increasingly incorporates digital technology and everything we do, we know this from the Covid response of the past few months. The issue here is that in our headlong pursuit of access to scientific progress we forget that particularly **vulnerable members of society** may be specifically impacted in technology in ways that we have not thought of, or that we are aware of but don't consider serious enough initially (E3).

Secondly, there was a concern regarding Big Data and AI widening the gap between the **poorer and wealthier regions** of the world, as the latter are already in a more advanced position in terms of having the necessary structures and resources to be able to benefit from it:

> the countries who are going to benefit from AI are going to be probably the **wealthier** countries because they're the ones who have the **infrastructure** to support it, like for **smart cities**… And they will only move **ahead faster** and they will leave **further behind** those poorer

regions that lack the **infrastructure or the manpower or the education, the skills** to support smart information structures (15).

5.2. Mapping Current Work to Address Ethical Issues

When discussing current work to address ethical issues in AI and Big Data, interviewees tended to remain at an abstract level. Figure 4 below is a broad mapping of the approaches that were presented. Where an interviewee gave a more in-depth example, this is presented as a case example in the speech clouds below. In the majority of cases, the limitations that were identified intersected with suggestions for improvement, as the latter were presented as a way to rectify the former, and so these are presented in Suggestions below (Section 5.3).

Some interviewees were quite critical of the EU approaches: '**it is frankly all over the place**', noted one. But this was justified by arguing that this is a fairly new topic and so things are progressing through smaller '**bits and pieces**' (E1). As was expected, GDPR (*General Data Protection Regulation*) was most commonly referred to, mentioned in 14/21 interviews. Feedback on GDPR was quite positive, with one interviewee calling it a '**model**' for other democracies across the world to follow:

> Europe has set up, I think, a really interesting mechanism of data protection for individuals which goes a long way in buying time for the European Union to try and set up their own digital environments that would adhere to European human rights law for example. So I would argue the **GDPR is a step forward**, and in fact states like California in the United States, Chile even Israel have examined the GDPR in great detail. It serves as a, I would say, **model for many democracies worldwide** at least at the federal level (E3).

Other interviewees **praised the GDPR** for increasing transparency; making the use of personal data more secure (E4); requiring some sort of risk assessment for high-risk data and applications (E11); being easy to understand in terms of its principles and improving privacy (E1). So it could '**be seen as a bit of a success story**' (E1). A particular example was given of the way GDPR adapted in France, where it gave people their autonomy from algorithmic domination in the sense that it protected '**their right not to be subject to a specific algorithmic treatment, so involving a human anytime you want for administration purposes only**' (E10).

Another interviewee talked about both the advantages and disadvantages of GDPR, calling it '**a wonderful first step**' and specifically discussed how it enabled '**the right to be forgotten, the right to request data**'. The interviewee gave a personal story to illustrate this:

> I recently, just for the exercise of it, wrote to Facebook and asked to have all of my data. And I did get it. And I think without the GDPR, that wouldn't have happened. So that is good that is addressing that…it's helpful (E15).

But this interviewee pointed out that GDPR '**in real life what that looks like is all the consent to cookies form you see on every website. That doesn't really solve the problem**' (E15). They also pointed out how companies are still not made accountable, and do not offer adequate product liability (see Section 5.1.8).

Other interviewees also argued that GDPR was ***limited*** in the sense that '**it tries to set some kind of compliance framework**' but ultimately '**leaves it to individual countries**' (E1), and '**different courts in different member states apply very different standards**' (E9).

Beyond the GDPR, within the EU context references remained vague, though there were some passing references to the creation of the ***High-level Expert Group*** **(AI HLEG)** of the European Commission and to ***European courts***. There was also reference to current work within the ***EU-funded research streams to raise awareness*** and integrate ethical principles related to Big Data and AI that reduce bias and promote inclusion:

> the inclusivity and diversity goals are really prominent in nearly all of the **European court halls**. So if you look at any of the **Horizon 2020** it's pretty much mentioned there and I think even something like that is quite important… especially these issues of bias and inclusivity and these things are always being thought about when **new research** is being undertaken (E4).

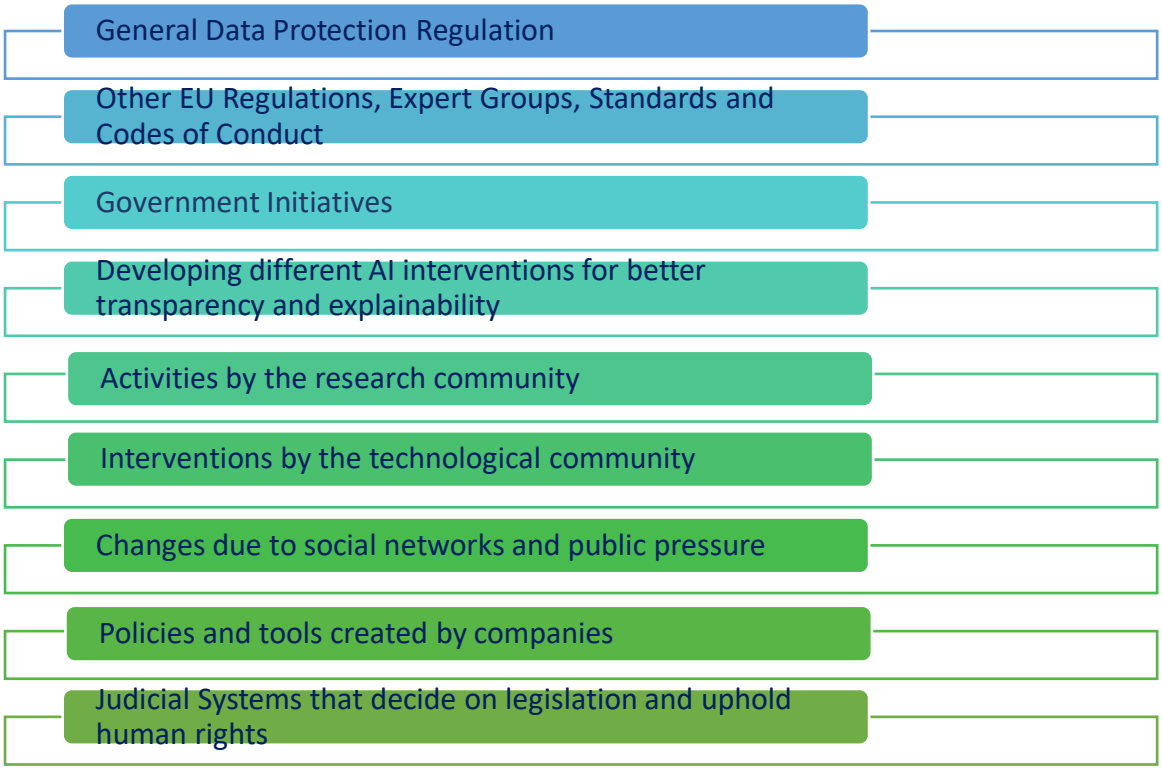
> there's **EU projects**, there's profit and non-profit **research** institutes...universities are going crazy with all of these AI topics...there's a lot of work going into **guidelines** and **standards** and those kind of those white papers in that sense. But all of these things are more of a **recommendation** character...there's a lot of things happening on the EU level which on the one hand is very nice. And I think that is very important. I think the big dollop of ethical questions has probably been addressed because there is a lot of attention that has come to **philosophy and ethics** in that sense throughout the course of the last few years, especially also with this **High-Level Expert Group** in AI (E8).

The same interviewee raised concerns regarding the extent to which various initiatives such as the AI HLEG, or '**other guidelines and other institutions**' or '**the IEEE report**' have moved forward in terms of offering practical guidelines that can be adapted into everyday scenarios, and how far they '**have moved with operationalizing their standards**' (E8).

Reference was also made to work which raises awareness about aspects of inclusion in education that bridged the digital divide: '**new accessibility regulations that are designed to make online education more accessible and... somehow everybody knows about them. So I just see that kind of role of awareness building**' (E4).

Communicating to governments was one other activity that was being pursued as part of increasing awareness. One interviewee mentioned how they were '**going around and explain[ing] to governments that software is just like any other manufacturing artifact that you can go ahead and make sure that proper procedures are filed, best practices filed. All the things that you can do for a car, you can do for a program. They just didn't know, nobody told them, and it's just a big magic thing for them** (E6).

Figure 4. Range of current efforts and actors identified that address ethical issues



'We are developing an AI intervention for social media, like to address misinformation... it is a part of the plan to include to make our tool transparent, to include those explanations. So it's something that goes hand-in-hand with any kind of like machine learning, kind of like interventions. So I think it's a code of conduct, at least within the academic research community developing these tools' (E13)

'I think the GDPR is a decent example of how it is being addressed with legislation. I also think that personal privacy, in the mobile data sense, is actually being addressed by people like Apple, by all technology companies' (E1)

‘How we address it is by systemic cooperation with the employees’ representatives. So that's what we do internally within the companies talking to the representatives of the employees, mainly with the trade unions, and trying to address the migration strategy towards the new IT technology or artificial intelligence by close cooperation with them. So all necessary training is made available’ (E14)

‘In France we have what they call the 'Conseil d'État, which is a supreme administrative court, which are sort of the gurus of fundamental rights and public law. So you have people at the Conseil d’ État who were thinking about the fundamental rights angle and whether legislation is sufficient to guarantee a good balance of fundamental rights’ (E11)

‘There’s one project called My Data, which I think is a project looking at these issues of kind of document wallets so people can keep all of their personal data in one place and start to manage it in the way that they can see what they’ve given permission for, who’s using it – so which companies have access to the data, how it’s being used and then importantly they’re able to retract permission. So even if it’s been given permission for it to be used in one way, if it’s then sold on to somebody else it’s required that the people would be giving permission also for the secondary use of their data. So this means that if you have a single point of entry you’re able to go and monitor the data use much more carefully. (E4)

A couple of interviewees talked about Big Tech companies in fact taking the lead in ethical issues and ‘the only ones taking this issue seriously’ (E10), though the majority thought that the efforts of these companies had made were still very meagre. Some were quite pessimistic, arguing that in many cases ethical issues are ‘**not addressed**’ at all (E16), while others made references to how data is currently regulated, for instance, how network operators are able to collect certain data only to the extent that they are able to maintain their ‘**business relationshi**p’ (E2). Interviewees spoke about the existence of legal frameworks ‘**and so corporations are already regulated, it's just a matter of: is it adequately regulated specific to its task?**’ (E6).

A gap was identified in the future energy market, where the interviewee argued that ethical issues in this case were not addressed:

> We don't know which AI technologies will be used and how they will be used. We have pilot applications and they are successful and relevant. But they are still pilots and not organized so that interoperationality is ensured (E2).

Overall, participants spoke of two types of approaches:

- voluntary, non-mandatory guidance that aims to raise awareness about ethical issues, and offers guidance and suggestions on how they should be addressed and is often general in nature, not specific to a particular sector and span '**a whole range of issues**'.
- The mandatory approach that through legislation and regulation aims to address these issues.

As one participant summed it up:

> there is already a lot of legislation out there. So we should never forget that there is no like legal vacuum. There is that machinery directive, liability legislation, GDPR, etc. Also just human rights legislation, which is already enforceable (E9).

5.3. Suggestions for Addressing Ethical Issues

This section presents fourteen recommendations that were suggested by the interview participants to deal with the ethical issues of Big Data and AI. For suggestions where there was no consensus, both 'sides' are presented; for instance, some interviewees preferred more radical changes to legislation than others. Education was the most prominent suggestion across the interviews and is presented in a more detailed way.

5.3.1. Recommendations: Legislation

a) Do not reinvent the wheel – work within existing legislation

Some participants suggested that relevant stakeholders should work within existing legislation and '**take stock of what exists already**' (E11). There was a specific suggestion by one interviewee for the European Parliament to perform '**stocktaking of what already exists to deal with algorithms …identify true gaps**' that exist in, for instance the consumer rights legislation and then fill in the gaps rather than starting to adopt new measures before necessary research was done (E11).

Another participant similarly noted that they were '**absolutely convinced**' that the required law is already '**in place**' (E3). It was 'simply' a matter of applying it to '**new circumstances, so good law, good EU law, like the *EU Charter on Human Rights* evolves as our standards of decency evolve**':

> So, if you look at an article like physical integrity you can apply quickly to AI if one approaches the problem with the understanding that the law is already in place and simply needs to be applied to a new problem or a new set of challenges. (E3)

It was argued that if the approach that taken calls for a '**whole new set of laws**', for instance to protect human rights, which the participant believed '**is what lobbyists do'**-

> then we are never going to get anywhere. So…one of the contributions the SHERPA project would make is to argue that **the law is in place**. We simply need to adapt our application of it through **regulation** and through **court case interpretation** (E3).

Keeping within this line of thought, of not reinventing the wheel, there was also a recommendation from interviewees to follow the ***successful aspects*** of the GDPR (see also Section 5.2. on positive feedback on the GDPR). One participant suggested that just as there were new privacy guidelines, new EU guidelines '**about the way AI should work and should be used**' might have good outcomes, despite it being a '**top-down**' approach (E18). Some specific characteristics of what these entail can be inferred from what interviewees identified as positive dimensions of the GDPR:

> it is principle based. It is driven from the individual. It is actually not that hard to follow. It has been left up to individual countries to legislate, which by and large, they have done in a very similar way. Then, it is monitored and reported (E1).

Conversely, it was seen as important to reassess the GDPR, with the objective of identifying its weak points and addressing these for a more improved updated version: as one interviewee noted, '**there are loopholes**' in the GDPR, and '**identifying what those loopholes are and finding ways to overcome them**' is a crucial future step (E4).

Another suggestion was to use the ***existing Human Rights framework*** as a ***universal*** homogeneous reference for AI development, as the ***common denominator*** which can help overcome cultural and social differences across the globe. This was suggested as a way to overcome the difficulty of dealing with '**different cultural aspects**' when **'an international AI approach'** is discussed (E5).

Related to the globalised nature of today's world and the fact that technologies travel across countries with vastly different legislations, one interviewee proposed a form of **'international governance'** of technologies:

> in a way that **minimizes the harm** that technology developed in one region can do in the **rest of the world**, because that region does not have the systems in place to address these ethical issues. So that has, over the previous years, for instance, been a serious issue with US developed technologies coming into the EU and then clashing with **EU legislation** (E19).

b) Further legislation is needed

There were however, interviewees who suggested that further changes in legislation may be needed, including new legislation:

> in general there is always ground to change legislation because legislation is made by people. Whenever and especially when there are constant changes, such as internet security that is constantly changing our reality, obviously legally we need to find a way to adequately respond to these both educationally and politically (E17).

Some calls were general, like the quote above, but others were more specific. Firstly, there was a call for further legislation for environmental protection**:** the proposal here was to have more legislation and more '**strict application**' of environmental legislation (see also 5.1.10) related to electromagnetive wave emissions, the recycling of used material (especially hardware), and data storage (E3). It was also suggested that a ***regulator*** could help make this more efficient.

Secondly, legislation was seen by some interviewees as the '**only way**' to promote ***transparency***, improve ***accountability*** and ensure that companies behave responsibly:

> as long as you don't say **very strictly** that this is the thing that is **absolutely essential to do** otherwise you are not allowed to use something or put something online, no one will care about that. It's really a thing that traces back to **legal issues** I would say (E20).

This included requiring companies to show **'a data trail'** if the user requests it, as well as legislation or regulations that protect human rights through ***diversifying*** Big Data and preventing ***discrimination*** (E13).

Another specific suggestion was to re-examine the 'grey area' of ***autonomy*** and ***responsibility***:

> So, in law and particularly in military affairs and chain of command **responsibility** you can always find a human who is responsible. But increasingly what machines cannot do now is **interpret**...AI machines can do reason, they can do many, many things, but they are still unable to interpret, something that human beings do without thinking. So, consequently we need to perhaps anticipate the ability of machines to interpret and begin thinking about how we might **adjust the way we regulate autonomous machine behaviour**. So that´s one area that we need to think about it, because for the moment we can still pinpoint the human, legal corporation…But we may not be able to do that for much longer… **by 2030 we need to have in place legislation that specifically addresses this grey area.** So, even though the principles of law are in place, we are creating a new physical entity with intelligence that doesn´t quite fit in and most people won´t be comfortable with existing law, so I think that is one area where we have to look: autonomy and responsibility (E3).

In terms of human rights, it was suggested that existing legislation should be reviewed to reveal the extent to which '**we need additional legislation that …concretizes these human rights in the context of AI…legislation might be needed to say, OK, this human right in the AI context means that you have an obligation to do this or that you have a right to do that'** (E9). In other words, further legislation may be needed to make more explicit the direct links between SIS and human rights violations.

5.3.2. The use of standards to deal with AI/Big data analytics

Regarding the use of standards to help deal with AI /Big data analytics, the majority of participants argued that this was a desirable way forward, though with particular considerations put in place. Standards were seen as helpful, as they '**would send out a clear message to people**', they would enable standardization procedures, they would improve transparency and increase trust towards technologies (E19 and E12). They were seen as '**a powerful approach because …manufacturers really go for standards. They follow standards, like ISO standards**' (E15). It was also seen as the ideal way to '**move forward**', as they acted as a form of **communication** which made it easier to spot those who did not follow them, as well as ***raised awareness*** generally:

> If we can talk the same language, we can develop the standards. We can move forward…We can say you don't keep those standards. (E16)

> these standards help a lot with bringing these topics to the awareness of different stakeholders…awareness concerning ethical and data protective issues …on the developmental level, on an educational level, on the implementation level, political level and to somewhat also to society, because all of these EU projects always have a category of you need to inform stakeholders, also inform society, which is I think a good thing, but I don't know whether this is enough. So I think these standards are a very good first step (E8).

Some participants talked about certain conditions that need to be in place in order the use of standards to be successful:

- Standards should not go too far; standards for Big Data and AI were compared to food standards which have been successful in improving food quality and reducing the risks of contamination, and so they could also be helpful '**as long as they don't overshoot it**' and bring in regulations that would seem absurd like those requiring '**bananas to have a specific kind of curve**' (E19)

- To have a compliance framework to ensure that people will comply (E1); to go beyond codes of conduct and '**have a set of standards that are legally binding and there's repercussions**' if they are not followed (E13)

- For stakeholder organisations first to have more discussions about what '**path**' and '**direction**' they should follow and only once this is '**clear**', should organisations '**start thinking about how this can be supported by standards**' (E2)

- To approach transnational organisations and seek their assistance and expertise but keep small and medium sized companies in the loop as well; '**they shouldn't be seen only as the under regulated space, they should be seen as potential allies**' (E6)

- To focus not only on '**interoperability between systems**' but also '**interoperability between […] users aspects as well**' (E4). The interviewee here argued that standards usually:

  > ensure inter-reliability between different systems. But I think that what's important here is that we're not just talking about the technical interoperability but also **user interoperability**. So if there are standards that are defining how artificial intelligence should be created, we've got data – how this data should be made transparent. How it should be made so that people can query it …. for example, in explainable A.I. if they can see the problems, the decisions that are being made…on what basis something was decided by the machine and do they agree… (E4).

- To take a **holistic** approach. Here the participant was part of the ISO working group on privacy by design for the international standards organisation and the argument was that some people are there representing industry, and they are more focused on doing as much as possible without breaking the law, while others are more genuinely and broadly interested in ethics. The conclusion of the participant here was that '**any standards that are developed for AI, I think industry will do everything in its power to circumvent those standards. So, I think standards will always be catching up to industry**' (E3). This was related, the participant argued, to the fact that standards are divided into '**many, many subgroups**' which '**weakens their impact**'. Therefore the suggestion was to:

> Approach standards from a **holistic angle** and rather than allowing AI standards to be broken down into dozens of subgroups, you know one privacy by design, another environmental pollution, perhaps the greatest service **SHERPA** could render in the development of standards is to force the issue of an **architecture of standards where each subgroup has an impact on another subgroup**. Standards will be very ineffectual with AI unless a **holistic approach** is insisted upon (E3).

Other participants spoke about the **limited** use of standards, given that they are not legally binding (e.g. E8 and E17). Some also pointed out that standards should not be given priority. Instead:

> the most important is for both academics and private research centres to align on what they want for the future…a general idea of which kind of project we agree to produce and which one we don't agree to. So it's a bit more vague than just specific standards (E10).

One participant (E3) touched upon the relationship between standards and regulations and suggested that the two interact with each other in the sense that '**hard regulation determines the standards in many cases**', while in others '**the ISO will develop a standard and it will become hard regulation after the fact**'. The participant argued that given that industry prefers voluntary standards, but '**the ideal would be that the regulator would work to have binding standards across Europe**', '**perhaps the best service**' that could be provided was to make **'hard standards'** as a kind of middle way compromise. The participant noted that '**the problem with standards is that industry will invest more in legal counsel than the public sector can ever afford to do**', and that this asymmetry of resources needed to be kept in mind (E3).

5.3.2. The use of guidelines

The predominant response of interviewees regarding the use of guidelines for developing and using AI systems, was that they are '**a very good step**' (E20) and would be '**absolutely**' ***helpful*** (E21). Suggestions were varied, and revolved around guidelines:

- being '**specific enough to be executable and testable**' (E19)
- ***operational*** in everyday scenarios
- offering a ***simple***, clear process that developers should follow, and if the issue is too complex, to provide guidance on alternative expertise that could be contacted
- ensuring that those who do not follow the guidelines are held ***accountable***
- possible ***evolving*** into standards
- being inclusive of ***good practices*** e.g. on '**how [do] you test for fairness, what kinds of tools can you put to the dataset in order to check whether there is bias**' (E21)
- being '**a combination of GDPR, putting the customer first and Asimov's rules**'[6] (E1)
- being **accessible** to users of AI systems, with adequate and clear explanations
- being able to provide ***risk/impact assessment*** for high-risk applications
- being **'holistic'** (E3) in the sense of including not just developers, philosophers and lawyers, but also educators, health professionals, environment experts as well as 's**omebody who is an expert on lobbying in the AI industry'** (E3).

[6] Isaac Asimov was a science fiction author who conceived of three laws related to robots in the 1940s. The three laws can be found here: https://www.pcmag.com/encyclopedia/term/asimovs-laws

Emphasis was given to the importance of guidelines that not only '**touch upon these big aspects of transparency, explainability, responsibility and so on**' but that '**also give kind of these practical operational guidelines**' that are operational in real everyday scenarios. Otherwise, the interviewees argued, guidelines become just like many other documents '**produced by some people in the ivory tower that cannot be at all used in the outside world**' (E8). One interviewee explicitly suggested avoiding writing the guidelines '**in a legal way**' so that '**every developer who has nothing to do with law can read**' and understand them (E16).

One participant differentiated not just between the ***developer/user*** side of guidelines, but also between what they saw as a list of considerations vs. a list of questions. Their preference was for a **'workflow type of guideline rather than a trustworthy guideline'**. The workflow type, according to the interviewee, involves:

> a list of **questions** that you should ask and when should you ask them, which ones and at which point, I think it is a very important question…If we use the EU's trustworthy AI guidelines as an example, it is nice to have a list of **considerations** for the developers to keep in mind because it is harder to demand ethics and answers from developers, but it is easier to say, 'Hey, did you think about this? And if you found it problematic, did you think some actions?'.
>
> For the **users** I would say it's a completely different thing. I feel like the users of AI system should have guidelines that **explain** to them how does the AI system work. What kind of **reliability** should they expect, what kind of **limitations** they should keep in mind?. (E12).

A related suggestion was brought up by one interviewee who recommended the inclusion of ***open-ended questions*** in order to encourage self-assessment, and ***self-reflection*** on behalf of both developers and users:

> it's helpful to have **open questions** that force you to think about whether you took certain things into account. And if maybe, you know, the guidance can spell out the different ways in which you can take that into account and what you can do about it, but not just asking like: 'Did you consider?,' but rather, you know, **'What mechanism** did you put in place to ensure that?'. And then give some options. So really forcing the developers and users of the AI systems to **reflect on**, how they're making use of a technology and how they consider tackling the ethical issues (E9).

Distinctions were also recommended depending on which '**sector**' or '**use case**' the guidelines were for (E9). Another recommendation was to have further discussions regarding **'electronic personhoods'** and including this in guidelines, especially given that '**AI is evolving very rapidly…all guidelines must think about this next chapter**' (E5).

5.3.3. The use of a new regulatory body or regulating officer

Suggestions regarding how regulations could help address ethical issues in Big Data and AI varied. Some interviewees argued strongly for the need of a new regulating body, others talked about which specific changes in regulation they thought should take place, while others described in more detail how they thought a new regulator could help, and what it should look like.

The new regulator, it was argued could '**create an agency that really focuses on this question**' of AI and Big Data (E12); it could act as a '**focal point**' (E4), offering '**explicit significance**' (E17), with extensive authority that could also improve accountability:

> So once it became known, people would know there's a **single entry point** for understanding what the **regulation** in practices around big data and artificial intelligence are important and if I had a question then I would know that there would be someone that would be able to field that question. And then I think the regulator should somehow be **monitoring** what is going on and looks for good and bad practice…in some way they would be the people that would understand what is and what isn't allowable. They would be reassuring that the guidelines would be kept up to date… I just see that kind of role of **awareness building** and also **imposing** – making sure that things are in post (E4)

Having the power to '**impose**' was something that several interviewees touched upon; any regulator, it was argued, '**shouldn't have an advisory role only**', it '**should have authority**' to '**impose fines**' (E13).

It was also suggested that the regulator should be a '**trusted entity**' (E2) and should make an '**anti-lobbying statement-promise**' (E3), so as not to fall prey to vested interests of companies. One participant argued that the '**biggest threat**' is located at the intersection of the government and private sectors, and so there should be '**clear regulation rules**' on how these two sectors '**should and can collaborate**' (E12).

Suggestions included an AI Commissioner (E1) at an EU level and national data commissions. The latter would be able to decide not only what type of information is stored on the cloud but also make decisions on the kind of things that are really needed. They would decide on questions such as: '**do we need self-driving cars, or do we only need them in certain circumstances and with certain protections in place?**' (E3).

Some participants spoke specifically about the need for a dedicated regulating officer and gave the following recommendations on how this could look:

> that could and should be a person who really understands the technology, and understands the implications of regulations, and understands the industry so well so he/she can at least have some sense of how this regulation will actually affect companies. Regulation is often well-intended, but then it becomes a consultancy checkbox certification of some sort that actually has very little impact, or some heavy-handed regulation that doesn't really help much... basically there are so many sorts of **knock-on effects and potentially unintended side-effects of regulation** that it has to be **really well informed by industry partners, informed by NGOs, informed by scientists and engineers** (E21).

A similar suggestion was made by E3, who strongly recommended a regulating officer; the main function of this person would be to make '**noise**', be visible and resist pressure from lobbies:

> I would love a regulating officer, that´s a great idea, because…their career will be dependent upon **how much noise they make** in their position. And I like the idea of having a regulator for AI. I think it´s really important. I look at the work that the CNIL[7] does in France, the data protection officer at the EU level, and although their roles are limited they are **quite visible**

[7] The CNIL, Commission Nationale Informatique & Libertés, is the French Data Protection Agency.

> **and I think it´s really important to call attention to it**. It's a **symbolic gesture** as much as a **practical administrative gesture**…I like the way special rapporteurs for the UN function, in other words they have the ability to set their own agenda within the constraints of their mandate…it´s very, very important that…[there is] some sort of an **ethical statement** that the person adheres to when they agree to accept the position… I would make that **anti-lobbying-statement** as it were, you know for every hour spent talking to a lobbyist that person needs to spend two hours speaking to civil society groups' (E3).

Another interviewee discussed at length the proposition of '**ethics oversight committees for AI**' instead of a regulator, which would be able to evaluate the specific uses of AI systems by each state or country (E7). This, it was argued, '**would be a big step towards protecting human rights**'.

Yet, there were some who '**were sceptical about a new regulator**' (E16) or thought that a new regulator, specifically at an EU level, would not help much due to differences in the way countries perceived ethical issues related to SIS:

> I think it'll be very hard to kind of force Member States to have the same kind of approach or regulator because that's another thing that we've countered in AI ethics generally, is ethics and even fundamental rights vary considerably depending on the Member State, and traditions and culture and so forth. So I think it'd be hard for the European Commission to adopt a system where there's like a European regulator. Probably it'll be necessary to give Member States freedom to adapt their own regulatory structures (E11).

Instead, what was proposed was a mechanism that would reinforce coordination across the Member States, similar to that currently taking place with telecoms:

> so you have a system where national measures basically have to be run by a special task force of the European Commission to make sure they're **consistent with some baseline principles**. And they don't go too far off the mark from what other Member States are doing. So, I think there'll be a need at the European level for a body that is in charge of making sure that national measures are consistent. And the best example of that is telecommunications regulations…**it's going to be difficult I think to impose a one size fits all approach to all the Member States**. So, a compromise is…you give Member States freedom, but you have a strong kind of coordination mechanism to make sure things don't get totally out of whack (E11).

5.3.4. Transparency in human and algorithmic decision-making

As discussed in Section 5.1.1., the lack of transparency with regards to Big Data and AI was one of the most prominent concerns expressed by the interviewees, and this went hand-in-hand with lack of explainability and informed consent. There was a strong recommendation for transparency in terms of the how and why of human and algorithmic decision-making; who has access to the data; for which purpose the data is used; and the implications of the data collection, use and storage on the user. One interviewee focused on three things that every developer should self-disclose: '**information about the data on which the system was trained, the set of algorithms which are used and the decision space**' (E16).

In a single case, there was a recommendation that even transparency should be regulated, as sometimes too much transparency could lead to more harm than good:

> It might cause **malicious actors** to exploit the algorithms and things like that. There is also certain types of **research** and certain **types of code** that maybe shouldn't be published because by doing that you are basically helping malicious actors to use AI for malicious purposes. So, **regulation** has a big part (E21).

One interviewee who spoke extensively about the need to have more transparency emphasised the need to provide the user with accessible and clear information regarding the process of ***personalisation***, the data that was collected and how this leads to ***targeted advertisements***. There was a strong call for:

> clear communication of how big data is used to **target i**ndividual users. And they should be tested, you know, like this should be user tested. So, for instance, like one example could be... like if I see a specific advertisement, it should be evidence-based...I think the **algorithmic transparency** is so important...and understanding what is **personalization**...so much of the content we see is catered and personalized to our taste. But then also we don't know how that kind of piece of intelligence was derived. Based on what data? Based on my browsing history, from when? Based on which platform, like, you know, I don't know. And some of it might be wrong (E13).

The interviewee argued that this recommendation was '**especially true**' for Facebook:

> because Facebook owns so many things… it's got WhatsApp, it has got Instagram. It recently bought, I think it was Giphy. So…just imagine how much **data** Facebook has on me and it can **target advertising** across all of these platforms. ... I think it is important to kind of make it that clear that this is **why** you're seeing this... a **specific explanation** that tells me exactly like, why this is the way it is **for me**, for instance. ...that's really important to kind of know that my version of what I'm seeing on the screen is **not the same as your version**. And that my **version is personalized** and it's personalized because I make these **choices** and these choices have led to these **assumptions** about me (E13).

Transparency was also recommended as a **'safeguard'**, a protection against human rights violations, but also one that enabled people to '**understand what's going on**', and that in fact they had the right and agency to '**challenge individual decisions'** (E11). Making '**machines transparent**' was also seen as going hand-in-hand with reducing the possibility of bias and acting as a force for upholding human rights in society (E16).

5.3.5. Increased accountability and liability for companies

The suggestion to increase the **a*ccountability*** of companies, which in turn would lead to more ***responsible innovation***, was common across the interviews. This could be done via education in schools, through increasing public understanding of the ethical implications of SIS, as well as via legislation, guidelines and standardisation (e.g. E5; E13; E21). A specific recommendation was to have ***product liability*** in terms of the service that companies provide, '**that puts the power in the hands of the consumer**', to hold them accountable to provide a good but also transparent ***customer service***

(see also Section 5.1.8), where users could easily get their questions answered regarding how their data is collected, stored, used etc. (E15).

Another suggestion was to have '**open questions that force you to think about whether you took certain things into account…So really forcing the developers and users of the AI systems to reflect on how they're making use of a technology and how they consider tackling the ethical issues**' (E9). This was related to '**document obligation**', in other words forcing companies to be transparent about their operations; this would act as '**an indirect incentive for organizations to respect human rights because they know that some information will become public. So they better do good on it**' (E.9).

5.3.6. Raise public awareness and engagement on ethical issues related to Big Data and AI

Raising public awareness and increasing public engagement on ethical issues related to Big Data and AI was another core suggestion espoused by several interviewees. It was seen as important to '**inform people**' of the possible implications of AI on the '**democratic process**' and how it could potentially threaten it, both on an individual level and a societal level (E8). For some, engineers[8] had a '**bigger responsibility to be more in the public domain**' and to engage in public discourse (E21). Their responsibility involved explaining to members of the public '**what machine learning is**', '**how things work and what are the real risks**', but at the same time '**what are not real risks**' (E21). The latter was especially important, it was argued, in order to avoid fear-mongering and unnecessary scares, for instance, of people imagining that in the next decade '**we will have evil superhuman AIs**' (E21).

Public broadcasting on these matters via platforms such as ***YouTube*** was one specific suggestion. Another was to use more ***national news agencies*** so that a bigger audience is reached. One participant suggested the development of ***stakeholder forums*** which included not just scientists but also the average consumer:

> who for some reason has a smart dishwasher that talks to her coffee pot refrigerator…She doesn't understand how or why. Give that person the right in a stakeholder forum to say 'You lawmakers help me understand. Put the power in my hands', to be able to say, 'Why is it like this and how can I buy one that's not like this? How can I turn this off? How can I have the right to have a product that functions without having to have software updates?'. So, to have a stakeholder forum where you gather the people at the lowest level who are the data subject, as we all are, and allow them to ask the questions they want answered. And eventually, go to the technology experts and say, 'OK, how can we come up with a solution to this?' (E15).

What was suggested therefore was a ***bottom-up driven*** process, where the individual consumer has both the authority, the means and agency to ask questions, and get them answered, but also form an integral part of the process, i.e. interventions being direct responses to their own issues or questions. Public discussions were seen as vital, especially in terms of **getting feedback** to guidelines or legislation related to AI:

> the idea of the public hearing, of public feedback, of a process of public consultation before the roll-out of any guidelines, or as you are doing with this project, to me seems essential (E3).

[8] In the case of engineers, it was suggested that in order to increase trust, it should be independent engineers and scientists that should be explaining to people how things work, rather than only employees of Big Tech companies.

A **citizen science project** was another interesting recommendation to increase **public awareness and public engagement**. Citizen science projects refer to the opening up of science to the wider public, whereby the general public actively engages in scientific research activities.[9] This was seen as a way of better involving those people who were already interested, but as the interviewee points out, this would not solve the issue of reaching out to those who seem uninterested in AI-related matters:

> people who are interested and want to be involved can get involved in something about that. For example, to monitor where do you see AI in your environment and in your country, and then it could be a global challenge to people. But again, it would be the people who are interested or the teachers who are interested in involving their students (E18).

As an another tool for increasing public awareness and understanding, one interviewee proposed the need for '**data advocacy projects**' and **'data advocates'**, the latter being able to act as mediators who would be able to make citizens better understand the ways in which technology and AI impact their rights, and so raise awareness of the ways in which they '**could ultimately infringe upon their rights**' (E7). As one user put it, it was vital to '**bring awareness to society…on how, on what levels and with what gravity, AI actually has an impact on each and every individual human being**' (E8). In other words, a user-friendly perspective was proposed rather than a business-centric one.

Related to improving the end user's understanding, the need to provide clear, simple and accessible communication in guidelines was emphasised by some participants:

> sometimes those guidelines are very robust texts and maybe they could work better on the kind of design layer to help with this awareness and capacity building…they could communicate this better to people (E5).

This would increase both public awareness and public engagement, as well as improve inclusivity. One interviewee spoke specifically about the elderly, and how it is a matter also of fairness for them to be able to understand these aspects (E5). Disentangling some of the '**buzzwords**' that are used in guidelines and recommendations would also help to ensure consistency and '**find a common ground**' across sectors and across people:

> So, for example, what do we mean when we talk about "transparency"? What do we mean when we talk about "explainability" and what do we mean if we talk about "fairness"?… I think it's very important to kind of find a **common ground** across all of these recommendations and possible future regulations in which **everyone is on the same page** of the understanding of the complexity or narrow narrowness of this one concept (E8).

5.3.7. Be inclusive, ensure fair and reliable data and avoid discriminations

Participants argued for an approach that was ***inclusive*** and ***fair***. It was argued that although there was increasing attention given to issues of exclusion, bias and discrimination, still the issue was not sufficiently addressed. There was a call for more ***diverse*** and inclusive ***representation*** of opinions in in human and algorithmic decision-making and data, but also for '**a universal design that will suit**

[9] For a more detailed explanation of citizen science is, please see here: https://theconversation.com/explainer-what-is-citizen-science-16487

people and give everyone the same possibility to use the system' (E16). Inclusiveness included '**all stakeholders**', e.g. trade unions, municipalities, the public.

One participant spoke at length about the need to address gender and diversity issues in STEM subjects (science, technology, engineering and mathematics). As long as certain groups of people are excluded, then biases in SIS will remain:

> it's really important to tackle gender and diversity issues in STEM subjects and computing subjects, in particular in data science, because we kind of know that they are ongoing problems, and I think that **until we have diversity in those subjects we're still going to have these kinds of issues of bias**... You can't really fully tackle it until you've have people from **all different walks of life participating in creating and generating and collating the data sets, collecting the data, being part of the data but then also structuring it and defining the A.I** and thinking about these issues in terms of how the applications are done (E4).

In practical terms, the participant also suggested '**a requirement for diversity testing**' of AI solutions to mitigate the problem.

One interviewee pointed out that the focus was on gender equality and racism, ethnic differences, and differences depending on sexual orientation, but what was not considered was '**bias between people who are fatter than others or people who could be considered more ugly than others**', arguing that the use of this subjective criteria is often '**what society does**' (E10).

This theme included changes in regulations that could better protect human rights, especially those of minorities and other vulnerable groups in society. This involved ensuring that data treats all humans equally, is unbiased and is part of ethical systems that are sensitive to the cultures of different communities. One interviewee called this an '**open ethics sector**':

> we need to be aware of dealing with minorities and various biased data that we use for training. In open ethics we have an approach for this that is called **open ethics sector**… it can help us to deal with minorities in different ethical frameworks and ethical codes. Every culture can have its own ethics. So how can we design a system that will fit the culture of a specific community? (E16).

Another suggestion was to better address ***misinformation*** through '**legal and technical solutions**' that '**build a protective chamber**', though there was no specific suggestion of how to address this in practice beyond the need for further explainability (E6).

5.3.8 Focus on consumers/users and on AI for the public good: 'do not leave people behind'

The argument here was that governments and companies which have a '**public service mission'** should embrace an approach whereby AI is used for the ***public interest*** and with social implications in mind: '**what can AI do to do good in society**', for instance related to work on smart cities (E11). The benefits of AI should ultimately be about ***improving the planet*** in relation to keeping it safe and in terms of climate change and environmental pollution (E2 and E3; see also Section 5.1.10). As one interviewee put it, the '**most important thing is to focus not on technology or how the market is organized but on the. Do not leave people behind**' (E2).

Related to '**not leaving people behind**' was a suggestion to mitigate the negative impact on human jobs and the human job market (see also Section 5.1.9). The recommendations suggested the creation

of ***national strategy plans***, such as those that Germany, France and Singapore already have, with '**a very clear ethical perspective**'**,** the main objective of which would be to prepare the economy and the human labour force to adapt to the '**disruptive**' effects of AI and to emphasise the importance of reskilling: '**how to capacitate…humans for the AI era**' (E5). Each national strategy plan would take into account the internal specificities of each country and be guided by ***ethical principles*** such as the ethically aligned design of IEEE, the *European Guidelines for trustworthy AI*, and the Asilomar Principles, the main characteristics of which are fairness, reliability, privacy, data protection and security (E5).

5.3.9 Embrace 'design thinking'

Underlying the above theme of focusing on AI consumers/users and the public good was a clear position that policymakers and companies needed to include the perspective of the ***consumer***. As one participant remarked: '**we need some real involvement with real people**' (E1). Following from this focus on the individual, emerged another suggestion from several interviewees: that of ***design thinking*** and co-designing: '**design thinking is all about starting with the customer need before you design anything**' (E1). This included asking questions such as **'what would a design thinker's approach to this be?**' (E1), and recommendations for ***further research*** into '**design issues**', especially how these intersected with '**negative use of persuasive technologies**' (E10), for instance manipulating people in particular directions (see also Section 5.1.6).

One interviewee also noted how paying more attention to the '**design phase**' may solve a lot of problems concerning AI. This included not just privacy, but also '**security by design**' and '**ethics by design**'. The objective is not to focus only on '**the damages**' but be proactive and think about how we can achieve '**a value sensitive design**' taking into account human values in a systematic and comprehensive manner (E5).

'**Co-design**' was suggested in terms of who is designing new AI applications, referring not to just diversity within the team of researchers, but also a better representation of '**the real end user**', so they could be cooperating in this process. Co-design requires getting:

> the **end users** involved to understand from **their perspective** what **their capabilities** are, what **they prefer** to be made, what they don't want to look at and so how this can be **safer** for everybody (E4).

A stronger focus '**on putting the customer and the individual first**' was also a suggestion, specifically for the SHERPA project to integrate in its work (E1).

5.3.10. Adopt a transdisciplinary and global perspective

A suggestion proposed by the majority of participants was to adopt an approach that is multidisciplinary, transdisciplinary and global. For instance, there were calls for more **'intellectual diversity'**, meaning more transdisciplinary approaches that include '**more philosophers and social scientists and anthropologists and ethnologists focusing on this research**' (E10). Another interviewee spoke about the need to bridge the **'real language barrier'** that exists between practitioners from the technology domain and philosophers (E7). When speaking of multidisciplinarity, some interviewees blended this discourse with that of inclusion in terms of gender, nationality, ethnicity etc:

> I'm a big fan of **multidisciplinarity**. So definitely there should be people from all kinds of different backgrounds. So legal, ethical, philosophical, engineering, data science, but also different genders, different cultural backgrounds, different Member States (E9).

Another participant specifically talked about how ***legal regulations*** should not be '**too vague**', but rather be understandable ***across disciplines*** (E7). This approach was not only presented as desirable, but described even as a **necessary 'duty' of good *practice*** in current times:

> So I think **we have to be** very interdisciplinary **right now**, because if I as a lawyer or a jurist, want to regulate, only looking for the legal perspective, I'm **going to do it completely wrong**. **I have the duty to be interdisciplinary right now**. So I have to understand anthropology, sociology, philosophy, but also technology, I have to understand about what is AI itself (E5).

A participant who said they were a philosopher argued that there are a lot of insights from the discipline of philosophy that are not understood properly, and that this is partly due to philosophers who '**don't speak well to audiences outside of philosophy. So what is needed is this kind of translation of all the work of different disciplines**' (E7), so that there could be better and more effective ***dialogue and exchange of expertise and knowledge****:*

> we need that whole kind of **ecosystem conferences** where people are talking about these issues, and **review boards** where people are doing the actual work, and they can have an i**nterchange**. And so I think that's one of the things that we don't have currently…we need more interdisciplinarity, like **genuine interdisciplinarity**, which is easier said than done (E7).

Adopting a ***global perspective*** was seen as good practice, and by some even as the only way forward. One interviewee strongly argued against having '**a non-global view'**, saying this '**is the same as not having a view at all**' (E1). The argument presented was that even if the EU decides to adopt a comprehensive framework on AI, but other powers such as China or the USA do not adopt it, then it will be ***ineffective*** and not much use beyond the academic level: '**What good is it?**' asked the participant. More attention, it was suggested, should be given to ***international organisations*** and to attempts to get '**the global stakeholders on board'** (E1).

A global perspective was particularly important when it came to regulations and legislation related to ***accountability***, as there was no control over companies' behaviour if they could just move to another region to escape ***liability***:

> as long as it is not global, something that is basically online can just **move somewhere else** and provide services from that location (E12).

> one thing that for sure needs to be taken into account is that this needs to be a **global body** because the **internet is global**. These **issues are global** (E13).

5.3.11. Provide more funding to advance the research field of ethics and AI

Interviewees called for more funding to advance the research field of ethics, with one pointing out that they thought funding was particularly scarce when it came to Europe, with SHERPA being '**one of the very few exceptions**' (E10):

> I can quote like three research centers in the US tackling these issues, two in Canada, four in the UK. In Europe this is very, very much nascent in my opinion. And we cannot sincerely think that we will take the lead on AI ethics if we don't **invest in research**, like serious research in AI ethics….talking about research centers, academia, we need **much more funding** (E10).

Given that some AI interventions are led by private corporate entities, and it is not always feasible to test whether their ethical interventions are effective or if it is mere '**ethics washing**', or following some other political or financial agenda, it was recommended that more ***public funding*** was invested in ***testing the 'effectiveness'*** of these interventions:

> more funding…could be placed in actually testing whether the explanations provided to the end user, actually make sense to the end user (E13).

Others saw space for a closer relationship between academics and private companies. More funding, noted one participant, would enable more research and more support offered to private companies when reflecting and deciding on '**what is acceptable and what is not acceptable**'; it was acknowledged that these questions are difficult to answer, but that it was time that they are no longer put '**under the carpets**', if we want to improve our democracies (E10). The same participant argued that '**the most important**' issue now was not necessarily creating specific standards, or new organisations that would ensure responsible innovation, but:

> it is for both academics and private research centers to align on what they want for the future. And that's not necessarily a common vision that would be shared by our companies. That would be completely impossible. But more like **a general idea of which kind of project we agree to produce and which one we don't agree to**. So it's a bit more vague than just specific standards for it, but the standard is easy to change if it's too technical and too precise. So it's more about, this is the way we want to work to change. This is how we will contribute to it to have a product. And if consumers aren't happy with that, please let us know and we will rediscuss it. But just some transparency about the high hands of these developments, because I think we need to **regulate research** from the beginning (E10).

5.3.12. Create regulations, laws and rights for non-human entities

One interviewee proposed the creation of '**robot rights**', given the context of AI gaining power and autonomy. The argument made was that as robots and AI gain more and more autonomy, their interaction with human beings will also increase, and so will their influence:

> we need to think about robots rights because they will interact much more with human beings. They will enter in a kind of social technical sphere, interacting with human beings, influencing human beings. And because they are getting closer to some of those, let's say **human characteristics**, the law should not also underestimate the necessity for us to bring some, let's say some **law perspectives also to non-human entities**. And in that point, I shall say as a lawyer - I'm jurist by formation - … that the law was not meant to regulate animals…robots, synthetic beings. And now law…must make a kind of **philosophical shift** to comprehend non-human beings in this regulatory framework. And it's very, very hard to make it (E5).

Such adaptations to the law therefore, would require a rather novel and different approach to the one that has been regulating human beings so far. The interviewee emphasised the '**necessity to analyze**

those autonomous beings from a deontological approach'. It was argued that robots were non-human, but nevertheless remained ***social actors*** '**that are already influencing our behaviour and interacting with us socially**' (E5) The interviewee accepted that adapting the rule of law to this new context '**is going to create a huge challenge'** in terms of legislation and liability, but it was something that had to be considered given these changing times.

5.3.13 Recommendations regarding the ethical approach to be adopted

- Acknowledge the complexity and sensitivity of the issues and do not expect clear-cut specific answers, but **specify which ethical philosophical approach** is to be used: '**what kind of ethics are we going to apply on automation and AI development?' and do not underestimate this question as there are different ethical paths and perspectives (e.g. deontology vs utilitarianism), and depending on which one is followed, then the output of the development is going to be very, very different**' (E5).

- Create a **'systematic and practical way'** of utilising conversations around ethics of Big Data and AI. Similar to some interviewees' approach to legislation, this suggestion argued that '**we should not waste time on rediscovering the wheel**', but rather study in more depth which kind of questions concern an already vast applied ethics literature, and see how they apply to the context of SIS (E12). Questions could involve:

  > What kind of methods should be used? What kind of ethical analysis should be incorporated into the innovation process? At which stage of the innovation process? What philosophical tools are available? (E12).

- Adopt an **'entire ethics ecosystem'** approach, via an '**oversight model**' that organises oversight for new emerging technologies: the ideal route would be to have laws and regulations and oversight that's '**responsive to our needs, but not overly restrictive or doesn't miss the mark**' (E7). The rationale here is to shift away from the tendency to focus solely on the '**legal compliance**' model which is '**just how laws typically function**' to a more '**oversight committee**' model i.e. a normative framework and an '**entire ethics ecosystem**' (E7).

- Be ***constructive and not defensive***: the argument of the participant here was that whereas a lot of attention has been given to '**negative**' and '**defensive ethics**', that focuses on the type of threats that AI could potentially bring, instead a better approach would be ***pro-active ethics*** that focuses on how AI can bring positive changes to the world by way of positive action rather than reaction. Pro-active ethics was specifically defined as ways in which we can:

  > actually and actively use AI solutions for **a better world**. Not just preventing bad things from happening deriving from AI solutions, but actually using them to improve the world, what some people have called for instance, **tech for good**. I think this is very much important because this system, this is the area of reflection while we will think about the uses of AI and what kind of future we actually want to build, not the negative one (E10).

- Encourage respect for **human dignity when building robots**: Following on from the suggestion above of **educating** young generations to appreciate human dignity, it was seen as vital that

robots are created in an ethically responsible way i.e. so that they do not '**resemble human beings**' as that would create a risk, especially for young children, of creating a false bond, getting **emotionally attached to machines** that imitate human behaviour without them realising that they are machines without emotions (E15). It was seen as crucial to have a clear vision of the ethical questions related to human dignity and to preserve its importance (E5; E6; E8).

- Differentiate between **immediate and future needs** and focus on the former first: the interviewee here acknowledged the importance of both immediate and future concerns but argued that it is '**very counterproductive to mix them up**' as this could '**take attention away from**' the more immediate and pressing contemporary concerns that should be urgently addressed. The argument was that '**often in ethical discussions, people get really interested in those sort of philosophical science fiction questions because they are very interesting about the future of humanity**', and although:

  > we should be mindful of…the risks to humanity of artificial general intelligence…robots taking over the world and AI getting out of hand and so forth…at this point, that **should not be the immediate focus** because that sort of technology is still miles away…that can take attention away from sort of the more immediate, what we call **the operational AI ethics** issues …'What do I do about this new system at airports to check facial recognition against passports?', you know, stuff that's **really happening now** (E11).

- Avoid ***ethics washing*** where a company reports to be doing something ethical merely to give the image or impression that they are behaving in an ethical way. It was argued that an '**abundance of ethics principles**' may lead to ethics washing, with companies merely '**showing that they have these principles, but then principles not actually being far enough developed and far enough detailed in order to be able to actually action them or…to hold anyone accountable**' (E19).

5.3.14. Education: who, how and when?

The recommendation of '**education**' was the most dominant suggestion throughout the interviews, with 16/2121 interviewees suggesting it. The specific education code had also the most references attributed to it in the NVivo software (54). Education was seen as having an all-encompassing role that spanned ages and audiences. It was viewed as almost a panacea for solving a lot of related ethical issues, and although there was not always a specific suggestion of *how* this education should look, interviewees were confident that it should be on '**different levels, for different types of audiences**' (E18).

Regarding *who* should be educated, this spanned a broad spectrum, including:

- Policy-makers and government ministers
- Law students, lawyers and judges
- Developers of technology
- Users of technology (general public)
- Civil Servants
- Journalists
- Teachers

- Children (early childhood)
- Students (primary, secondary, tertiary education)

Education was seen as aiding people to interact with machines '**in an intelligent way and make informed decisions**'. It enabled an '**informed understanding of what these things can do and cannot, and how they do it, and what are their limitations**'. Otherwise the danger is that people, for instance, in key positions of power and responsibility end up being '**servants to technology**'. Therefore education was viewed as a tool that ultimately increased the public's agency and autonomy, as well as ability to make informed decisions (E18).

It was acknowledged that for adult citizens, lifelong learning would only work if citizens were themselves ***motivated***, usually through a significant turning point or a life-changing event . Parallels were made to how certain topics like the Coronavirus are now relevant for people as they have a motivation to learn about this issue that affects them, whereas before the pandemic it would be difficult to motivate people to learn anything about this virus:

> I don't think we can shove information into somebody's throat **if they are not interested** in that information. So, right now many people are interested in the Coronavirus; two months ago, or three months ago, you couldn't make them learn about it even if you wanted…It's not that everybody will be interested suddenly, **unless something happened** (E18).

The suggestion given here was to have information that is both accessible and interactive so as to increase the chances that the general public will learn about it:

> have the **information available, accessible, understandable** to different levels. For journalists, but also for civil servants, for general lay public. So, I would have the information or whatever information you think you can make, or tool kit to help teach about this, and then a website to interact, like an **interactive** website about it (E18).

Continuing education was also seen as '**really important**', and it was argued that this required a change of '**mindset**' where one is seen as able to adapt and to learn new skills, for instance '**rescaling in order to be able to interact with AI technologies**' throughout their life:

> it would be **extremely beneficial** for everyone to work with the assumption that they are never done learning. You have to **keep learning and being educated throughout your life**. You are never completely expert on everything. With that mindset which has been part of certain kinds of professions, especially in medicine, spreading out that kind of mindset to the population as a whole, I think it might be a good solution for that (E19).

Regarding the ***formal school education level***, there were differing views as to what education that would help address ethical issues of AI and Big Data could look like, and how it could be implemented.

One interviewee argued that the education system is already overwhelmed and so it should ***not be added as a new topic*** into the curriculum. Instead, it should be embedded into the education system '**as part of other skills**' with a focus on skills rather than new content. Their suggestion was to teach it in ***citizenship education***, for instance, and for the teaching to be tied in within an existing learning context:

> I think the educational system has so much on its shoulders and so much topics and so much skills that the curriculum can't have much more. It should have less. So, the way to incorporate that into education system is as part of other skills you want to teach. It can't be like a new topic. There are enough topics and there are enough skills. But if you want…in **citizen teaching** or whatever, to teach some of the things you will need to teach anyway in most countries, to teach them in the **context of that sociocentric issues**, I think that's a good idea. Just to say, "We need a curriculum about this topic", that will not work. There are enough topics not being taught and it will be just another one of them (E18).

Another suggestion on how education could help was through ***data literacy, information literacy, digital literacy or media literacy*****.** It was argued that although a considerable amount of work has been done already in schools, this has happened '**in a rather haphazard way**'. It was important to acquaint students with the dangers of technology and especially with:

> the fact that their **data can be mined** in the first place. Being unaware of data mining practices and exactly how much of your data you are giving up and where it is going, what it is being used for, being ignorant of that, that is what has led to, for instance, those quizzes on Facebook being able to mine so much of your data that they were able to influence elections. So, improved **data literacy** for those accessing these technologies…**IT education in primary and secondary schools around most of the world has been drastically disregarded**…it hasn't been as important as reading, writing, mathematics in terms of skills and knowledge that you will continue to apply for the rest of your life, that will make a drastic difference in self-reliance, in ability to engage with the world. So ideally, I would see a **huge improvement in the education and provision of IT skills** (E19).

One participant noted how it was '**absolutely essential'** to have this type of education: '**Fundamental information literacy…should be a major part of the school curriculum**' (E21).

Critical thinking was another competency specifically linked to educating citizens (this time the reference to education was a broader one, but included a specific reference to schools and universities) about the ethical issues, how they could deal with them, how they could improve their understanding and how to make informed decisions. The education aspect was also directly linked to democracy and ***education for democracy***. It was argued that there is:

> **a responsibility to educate citizens**. We really want educated citizens in our future world and as Big Data is becoming a bigger part, it is also of course a topic that people should have some idea on how to deal with that. **It's really a democracy issue** (E20).

Similar to other participants, the suggestion here was to have a closer focus on teaching students how to think (critically) but also how to develop their argumentation skills (E21), so offering not just '**knowledge of science…but knowledge about science**'. It was argued that students:

> need to be able to make judgments about the information and judgments about behaviours. I would say that when we take that seriously, we want **education for democracy** and that we really want citizens with **critical thinking** (E20).

Similar to education for democracy, was the suggestion of ***education for human rights***. This type of education, it was argued '**contributes to cultivating this specific sensitivity of being suspicious at any time, and being aware of the consequences, that the information that you may be giving**', could

potentially be used or abused by different actors, either for targeted advertising or for manipulation, ultimately leading to '**various human rights violations**' (E17).

In a well-functioning democracy, companies should be kept accountable by public pressure, and in order for citizens to be able to do this, '**a key ingredient**' was to be able to '**actually understand how these things work**' through **education**. This was especially important, it was argued, given that in today's world '**AI algorithms might determine your life choices and your opportunities in life**' (E21).

A more specific suggestion was for teachers to promote students' critical thinking skills and awareness regarding the dangers of Big Data and AI through ***real-life examples*** of problematic practices, such as how AI can be used to manipulate political agendas. However, a necessary requirement was that the students are given an in-depth understanding of the events by their teachers, rather than superficial discussions:

> As long as **the person doing the teaching** has a **full understanding** of exactly what happened, rather than the sometimes shorthand references that are given to Cambridge Analytica or Russian Bots (E19).
>
> That's where ethics become **realistic**. When you have **real scenarios** (E21).

Another specific suggestion for education was to improve understanding of the ***distribution of labour*** in companies, as this would help identify who is responsible for specific tasks and decisions. This, it was argued, could shed light on '**how the scientists behave or how the company need to deal with legal issues'** making it easier to assess whether a company was behaving appropriately or not (E20).

Teacher education was also suggested as a specific recommendation. Teachers, it was argued, should be trained in how to include ethical issues of SIS during their training, or later or in their professional development. But there should also be '**computer science or technology teachers as they are really educated as experts in this field**', teaching such topics in secondary schools (E19).

Regarding *when* education should start in the ***formal level of education***, again the majority argued that it should ***start at primary school and continue all the way through university***. One interesting argument for having education on ethical issues of AI and Big Data at the school level and before attending university was that many new innovators or leaders of Big Tech have been '**college dropouts and that makes it more important that this should be part of education before university**' (E19). One interviewee recommended that '**ethics and philosophy should be part of early curriculum**' but that that the ideal place to teach computer ethics is either at high-school level or tertiary education level (college or university) (E21).

Another interviewee pointed out that it was particularly important to educate children about human values and the importance of human connection, especially given children's vulnerability to creating false bonds with machines that resemble humans:

> education of what AI is, what smart information systems are, and what robots are and are not, should be started at a **very early age, for children**…I have two little kids. And I wonder in the future, will they be made **to be attracted to and drawn** in by the flashy technology, some sort of robot or a kind of information assistant. And will they learn to rely or not, instead of looking for human connection? And so I think that it needs to be a real early on activity to educate children about human beings, **the value of human beings versus the value of technology** (E15).

There was a call, therefore, especially for the younger generations to be taught to have respect for ***human rights*** and an appreciation for ***human dignity***. This was related to the fear that given they are '**growing up now with the presence of AI already established**' they would end up ***losing trust*** in humans, not acknowledging the '**human value to something**' and have a '**preference for decisions made from AI**' (E15).

Motivation was also acknowledged as an issue for students and teachers at the formal school level. For instance, it was argued that even in citizen science projects, it would be those teachers already interested in the topic who would try to involve their students (E18). So one of the challenges of education was how to have a broader outreach and not just talk to those already convinced of the significance of these issues.

At the ***university level*** one interviewee gave an example from a course they were preparing that was also relevant to adult education, geared at those with more work experience:

> we are currently setting up a masters in AI ethics, but it's not aimed to people who come straight out of the undergraduate degree. So, with a bachelor's degree going on to do a masters. This is aimed at…people who have been in a job for a long time and now suddenly have to deal with AI and Big Data algorithm and their ethical implications (E19).

Another suggestion for education at the higher education level, which is similar to the one given above for education at the school level, was to not treat ethics as a separate course. This often happens currently, and is a reason why lessons '**sometimes go wrong'**. It should be made an **integrated** part of other courses:

> So, like you take courses on Interface Design, C++ and ethics. That means that ethics forms a very small part of the curriculum, that the course would be very different to what the average student is looking for or wants, and it means that it could be compartmentalized as something different. So, what I would propose instead is not to have the ethics course but additionally **making ethics part of the other courses**. I mean not setting it apart as "there is only this one place where we look at the ethical implications of things". **Ethics should be a continuous integrated process of developing technology** (E19).

Others presented a slightly different view, arguing for the inclusion of specific and mandatory separate '**ethics courses in engineering and AI education**' (E9).

Despite education being a dominant recommendation across the data, there were some instances of ambivalence or caution regarding the extent to which we are perhaps expecting too much from education, or putting '**too much pressure**' on developers. For example, a participant argued that it is '**too much**' to expect from developers, and that '**it is not reasonable to expect developers to also become ethicists and solve ethical problems**' (E12). In another instance, the interviewee who gave the citizenship education suggestion above also said, '**I don't think that kids are the right audience for this**' (E18). Therefore we see a case of ***ambivalence*****,** even within a single interview, as to the extent to which education in the formal school setting should be implemented.

6. Conclusion

Eleven main ethical issues related to SIS emerged from the *Exploratory* interviews with diverse stakeholders: lack of transparency; 'information asymmetry' and lack of public understanding; biased data and lack of critical thinking; loss of human agency and dignity; failure to protect privacy as a fundamental human right; surveillance, manipulation and coercion; unethical monetisation of data; lack of accountability and product liability; loss of human jobs and mistreatment of employees; health and environmental risks; exacerbating socio-economic inequalities and the digital divide.

The findings from the analysis of the *Exploratory* interviews suggest that it is important to adequately understand these ethical challenges or risk 'misapplying' (as one interviewee suggested) the ethical resources available to solve or address some of these problems. The large number of ethical issues that emerged from the interviews, and the wide-ranging contexts in which they were discussed, indicate that there is still a lot of work that needs to be done to improve the way ethical issues related to SIS are addressed. This requires global cooperation and a multi-actor response that involves all levels of society, from the state, to private companies, the public, the media, schools and educators. Similarly to a previous report where exploratory questions were asked in the format of focus groups (SHERPA Deliverable 4.2. *Evaluation Report*), participants tended to focus more on the limitations, challenges and inadequacies of existing efforts, and on the ethical issues raised, rather than on the progress that is being made. 20/21 *exploratory* interviews argued that ethical issues are currently not adequately addressed; this is particularly worrying given that the findings suggest that the issue at stake is not whether changes are needed but firstly, the extensive and multi-level nature of these urgently needed changes, and secondly, how exactly these changes should or could materialise in the near future.

The overwhelmingly negative language used included references to the '**return of potential authoritarianism**', '**surveillance capitalism**', '**ethics washing**', '**manipulation**', '**coercion**', '**surveillance**', '**loss of human dignity**', '**destroying**' human liberty, and violating human rights. Reference to asymmetries also featured in various ways in the findings: whether it was in terms of information asymmetry – the knowledge gap between developers/Big Tech and users/the public; power asymmetry between those in positions of authority for instance big corporations or governments and judicial systems (with little accountability), and individuals who have little agency to accept decisions and little understanding of the breaches of human rights that are affecting their daily life due to inaccurate and biased data; and socio-economic asymmetries within and between countries that are exacerbating the digital divide and as a consequence, the burden of the negative implications of unethical use of SIS. The human individual vs. machines was also presented as a worrying issue, not only due to the loss of human dignity, autonomy and intervention of algorithms, but also to the implications this is having on the loss of human jobs, or to employees being under constant pressure to perform better, their efficiency being unfairly and unrealistically compared to that of a robot.

Lack of transparency and lack of adequate public information/understanding of ethical issues were two of the most prominent ethical issues that were identified, so it is perhaps not surprising that increasing and improving education on ethical use and production of SIS was the chief recommendation to policymakers. Education was extensively discussed as the best way forward. The provision of education as a recommendation spanned different target groups, ages and occupations, including lawyers, policymakers, producers of technology, and citizens, as well as students of all ages and levels – from early childhood to secondary school to lifelong learners.

Although some of the interviewees gave specific competencies, examples and approaches that they thought would be helpful for addressing ethical issues, such as embedding ethical issues into the already existing curricula of human rights education or citizenship education, or enhancing digital/information literacy, the knowledge given was based more on their own personal expertise rather than empirical studies of what works best in addressing ethical issues of AI and Big Data through education. Therefore, more empirical research is necessary in the specific realm of the role of education in addressing ethical issues of SIS, in formal, informal and non-formal educational settings.

In addition to education, thirteen other recommendations emerged from the analysis of the interviews, including a focus on the 'public good', increasing transparency, embracing design thinking, adopting an inclusive transdisciplinary and global perspective, ensuring data is fair, free from bias and reliable, increasing accountability of companies and governments, as well as strengthening public engagement. Specific feedback on use of legislation, standards and regulators varied across the interviews, but the participants did offer certain 'conditions' of good practices that policymakers and legislators may want to consider when formulating new policies and adopting new legislation.

The *Exploratory* interview findings also revealed certain tensions and dilemmas, some of which partly explain why progress regarding ethical issues in the use of SIS has been relatively patchy and slow. Policymakers and other stakeholders may also want to keep these tensions in mind when deliberating over future policy reforms and carving out new recommendations or initiatives. Firstly, the tension between the interests of business/science innovation and the customer emerged from the data. Participants extensively discussed the need to put the customer first, to include more customer services that enable product liability, and the ability of the user to get adequate answers regarding how a specific product or service is affecting their data, their human rights, or their safety and security. The customer needs vs. scientific innovation dilemma ultimately means a problematic focus on a business model that prioritized innovation and left out, or left behind, what the customer might really require, as an individual with human rights. There was a call for a business process that starts from customer needs and then moves on to the ethical design and responsible innovation aspect. Essentially, at the core of this dilemma is a financial-ethical tension, which as discussed in Section 5.1.7., in practice often results in business corporations with strong profit motives unethically monetising data or using targeted advertising at the expense of citizens' privacy, security, liberty and autonomy.

A second tension is related to the globalised and interconnected nature of AI; it travels across country boundaries with vastly different legislations and regulations. There is on the need for some form of international governance, with a suggestion also to build on universal human rights frameworks, but also the reality that regional or country sovereignty translates to countries having a strong hold on national legislation within their borders. Even within the EU, it was argued that this is a potential issue, as each Member State has its own interpretations, preferences and adaptations of EU principles. Moreover, serious issues develop when technologies developed in one region of the world, enter another region with different legislation. Therefore, if there is to be international governance, it needs to take into account country differences, while at the same time mitigating the harm that occurs when technology enters countries that do not have the required structures in place to address ethical issues.

Along the lines of exploring opportunities for intentional governance, were the findings from the *Regulatory* interviews as well. Specifically, feedback from the regulatory interviews highlighted a number of challenges for a potential AI regulator, including, accountability issues, issues of democracy, fairness and non-discrimination, need for ensuring transparency, positive user-experience and the wellbeing of citizens. At the same time, the interviewees considered, in addition to the potential challenges of regulating AI, a number of opportunities for positive change, including positive

societal impact, more research and innovation, economic growth, better educational opportunities for professionals, more active involvement by the national authorities, etc.

Finally, the feedback on the Guidelines (*Guideline* interviews) can be summarised as positive overall. Even with the identified challenges and proposed updates to the two Guidelines documents, the overall reactions and comments are positive, especially on the practical nature of the Guidelines, which was one of the main goals of this task, with a number of suggestions that have been picked up for further improving the two sets of Guidelines.

Overall, the analysis of the thirty-five interviews provided insight to the ethical issues that come out of AI and big data. There was consensus among the different stakeholders across Europe that the rapid technological development might have a negative impact on fundamental human rights and raises several ethical issues, with consequences on individual and societal well-being. The interviewees acknowledged that SHERPA's outcomes – development of Guidelines, proposing Regulatory Options, and Terms of Reference for a new Regulator – are in the right direction of addressing those issues, but at the same time they underlined that much more needs to be done. Interviewees were not always very clear in providing specific suggestions on how to efficiently address the ethical issues, perhaps because there is currently no available data, such as evidence on how to efficiently support individuals' critical thinking when using AI. Regulation, politics – involving international cooperation– and education were identified as key players in future efforts. The question of *how* those players will address the ethical issues that come out of AI and big data and protect human rights, remains for future research to address.

7. References

Braun, Virginia and Victoria Clarke (2006) Using thematic analysis in psychology, *Qualitative Research in Psychology*, 3:2, 77-101

Bryman, Alan (2008) *Social Research Methods*, 3rd edition, Oxford: Oxford University Press

Charmaz, Kathy (2004) 'Grounded Theory', in M.S. Lewis-Beck, A. Bryman, and T.F.Liao (eds.), *The Sage Encyclopedia of Social Science Research Methods*, Thousand Oaks, California: Sage

Acknowledgements

The SHERPA project would like to thank those participating in the interviews, including Emil Eirola from Silo.AI; Kanta Dihal from Leverhulme Centre for the Future of Intelligence, University of Cambridge; Michalinos Zembylas from Open University Cyprus; John Basl from Northeastern University; Annika Wolff from LUT, Susan H. Perry from the American University of Paris, Winston Maxwell from Télécom Paris, Nathalie Smuha from KU Leuven, Joanna Bryson from Hertie School, Petros Mina from Fountech; Christiana Varda from UCLan Cyprus; Penny Duquenoy from Middlesex University, London; Iain G Mitchell QC; Jouni Seppänen, Mike Yates, Laura Crompton, Galit Wellner, Nikita Lukianets; Eduardo Magrani and other anonymous participants.

8. Appendices

REPUBLIC OF CYPRUS

CYPRUS NATIONAL BIOETHICS COMMITTEE

Ref.: ΕΕΒΚ ΕΠ 2018.01.108
Tel: 22809038/039
Fax: 22353878

June 28th, 2018

Dr Kalypso Iordanous
Associate Professor
UCLan Cyprus
University Avenue 12-14
Pyla
7080 Larnaca

Dear Dr Iordanous,

Application for bioethical review for the research entitled:
«Shaping the ethical dimension of information technologies – a European Perspective (SHERPA)»

The Cyprus National Bioethics Committee (CNBC) has reviewed your application for ethical approval for the project outlined above submitted on the 21st of June 2018. From the review of the documents you have submitted, your research proposal is deemed to meet the requirements of the Law Providing for the Establishment and Function of the National Bioethics Committee (No. 150 (I) / 2001 -2010) and does not necessitate a full bioethical review from the CNBC.

2. Kindly note that approval is granted provided that the following conditions apply:

a) conduct of the research is strictly in accordance with the proposal submitted and granted ethics approval, including any amendments made to the proposal required by the CNBC,
b) inform CNBC immediately of any complaints or other issues in relating to the project which may warrant review of the ethical approval of the project,
c) before implementing any amendments to the proposal as approved, request a new approval by CNBC,
d) provide a follow up report on the progress of the program every 6 months from the approval date,
e) provide a final report upon completion of the program,
f) inform us in writing in case the project is discontinued.

.../2

Appendix B

SHERPA - Shaping the ethical dimensions of smart information systems (SIS) – a European perspective

Task 4.2 – Stakeholder evaluation and validation

Information Sheet

Please take some time to read this information and ask questions if anything is unclear.

Contact details can be found at the end of this document.

What is the purpose of this study?

The SHERPA project investigates, analyses and synthesises our understanding of the ways in which smart information systems (SIS; the combination of artificial intelligence and big data analytics) impact ethics and human rights issues. The project aims to develop novel ways of understanding and addressing SIS challenges. The focus groups aim to explore stakeholders' views regarding the recommendations that have been developed in the project thus far, with the objective to improve those recommendations.

Who is organising this research?

The research for this study is being undertaken by the EU-funded SHERPA project (SHERPA is the acronym for 'Shaping the ethical dimensions of information technologies – a European perspective' (https://www.project-sherpa.eu)

A Research Ethics Committee has reviewed and approved this research.

Why have I been chosen?

The project aims to conduct 2 waves of 5 focus groups each with several stakeholder categories – e.g., representatives from policy, science, industry, civil society, politics, media, academia etc. The aim of the 1st wave of interviews is to explore initial reactions form stakeholders regarding the overall set of recommendations. Focus group participants will be asked to use an action plan to take the recommendations back to their constituents and collect broader feedback.

Do I have to take part?

Participation in this study is voluntary and you may ask any questions before agreeing to participate. If you agree to participate, you will be asked to sign a consent form. However, at any time, you are free to withdraw from the study and if you choose to withdraw, we will not ask you to give any reasons.

What will happen to me if I take part?

If you agree to take part in this study you will participate in a focus group in person, with other 9 stakeholders where you will discuss the recommendations that have been developed from the SHERPA consortium regarding the development and use of SIS[KI1] . The time and place that the focus groups will take place will be determined in coordination with all the participants in order to find the most convenient time and place for all attendees. You will be asked to use an action plan to take the recommendations back to your organization and collect broader feedback. We will be asked to participate in a second focus-group meeting, though participation in this is optional. During the second round of focus groups, you will be asked to provide specific suggestions concerning the formulation and implementation of the recommendations. [KI2]

What are the possible benefits of participating?

The study aims to develop proposals for the responsible use of SIS. In addition to helping the SHERPA project, advanced analysis will be carried out on the stakeholders' focus groups discussions to which you contribute, which may raise issues that your organisation would like to know about and take steps to remedy.

What are the possible risks of taking part?

There are no risks in taking part in this study. At any time during the interview you can choose to withdraw. You may also choose to withdraw your data from being used in the project at any time until 1st July 2020.

How will my interview/focus group data be used?

The focus group will combine quantitative and qualitative elements and will be designed and analysed by SHERPA project partners. The recording of the focus groups may be transcribed by parties outside of the consortium. If this happens, the transcription company will delete the recording and transcription after the transcription is approved. On the consent form we will ask you to confirm that you are happy for the SHERPA consortium to use and quote from your interview. Any such use will be anonymous unless you indicate otherwise on the consent form. Information which will identify your organisation will also be kept out of publications unless otherwise indicated on the consent form.

What will happen to the results of the project?

All the information that we collect about you during the course of the research will be kept strictly confidential. You will not be identified in any reports or publications and your name and other personal information will be anonymised unless you indicate otherwise on the consent form.

What happens to the focus group data collected during the study?

The focus groups discussion will be transcribed by the interviewers or a designated, approved third-party agency. If we use a third-party transcription service, we will ensure that there is a signed data processing agreement in place. The audio files will be deleted, once the analysis of the focus groups data is complete.

What happens at the end of the project?

You may request a summary of the research findings by contacting Kalypso Iordanou, University of Central Lancashire Cyprus (KIordanou@uclan.ac.uk).

What about use of the data in future research?

If you agree to participate in this project, the research may be used by other researchers and regulatory authorities for future research. The transcript will be kept for five years after the publication of the findings of the study.

Who is funding the research?

This research is funded by the European Commission under grant no. 786641.

What should I do if I have any concerns or complaints?

If you have any concerns about the project, please speak to the researcher, who should acknowledge your concerns within ten (10) working days and give you an indication of how your concern will be addressed. If you remain unhappy or wish to make a formal complaint, please contact Dr Kalypso Iordanou, KIordanou@uclan.ac.uk.

Fair Processing Statement

The information collected will be processed in accordance with the provisions of the EU *General Data Protection Regulation* (*GDPR*)

Appendix C

Project SHERPA – Consent form

Issue	**Respondent's initials**
I have read the information presented in the information letter	
I have had the opportunity to ask any questions related to this study, and received satisfactory answers to my questions, and any additional details I wanted.	
I am also aware that excerpts from the focus group meeting may be included in publications to come from this research. Quotations will be kept anonymous unless I give specific permission to the contrary (below).	
I give permission for my name to be associated with excerpts from the interview which may be included in publications to come from this research.	
I give permission for my organisation to be identified in any final publications produced by SHERPA.	
I give permission for the focus group to be recorded using audio recording equipment (if necessary).	
I understand that relevant sections of the data collected during the study may be looked at by individuals from or a project partner from SHERPA. I give permission for these individuals to have access to my responses.	
I understand that the audio recording may be given to a transcription service company to transcribe. I give permission for these organisations to have access to my audio files for transcription purposes.	

With full knowledge of all foregoing, I agree to participate in this study.

I agree to being contacted again by the researchers if my responses give rise to interesting findings or cross references.

□ No □ Yes

If yes, my preferred method of being contacted is:

□ Telephone: ……..

□ Email: ……

□ Other: ……..

Participant Name		Consent taken by	
Participant Signature		Signature	
Date		Date	

Appendix D

Questions for Guidelines Interviews

1. You have now read two guidelines, one for use and one for development. Although these guidelines often overlap (e.g., because we sometimes want to protect end-users by requiring developers to adapt their systems), they are supposed to provide different guidance when appropriate. Reflecting on that, do you see any reasons for revisions?

2. The guidelines are supposed to be easy for practitioners to read, understand and apply. Do you see any need for adjustments because of a risk of misunderstanding, conflations, or ambiguous language, either because the guidance is not clear enough or because it includes too much jargon?

3. The guidelines are supposed to be engaging, which is always a problem for a relatively long documents of instructions. How would you judge the guidelines with respect to engagement?

4. What is your impression of the use of graphics (tables, figures, pictures) in the document? Should any changes be made, if so, in what way and why?

5. If you have experience with many other similar documents, how do you compare these guidelines to other guidelines with respect to: understandability, engagement, and usefulness?

Questions on specific parts:

6. What is your evaluation of the "Introduction"?
    1. Does it cover what it needs to cover? Is anything missing?
    2. Does it give a good introduction to the guidelines?
    3. Is it engaging?
    4. Are the language and length appropriate?
    5. Does your impression vary between the two guidelines?

7. What is your evaluation of the "High-level requirement section"?
    1. Does the section make an important contribution to the rest of the guidelines?
    2. Is the language appropriate (understandable, no jargon, engaging)?
    3. Are the different high-level requirements and their sub-requirements sufficiently well explained/motivated?
    4. Are the language and length appropriate?
    5. Should something be removed or added?
    6. Does your impression vary between the two guidelines?

8. What is your evaluation of section 3 (i.e., models/methods for development/governance)?
    1. Is it well-adapted for practitioners?
    2. Is it suitable for your organization?
    3. Does it contribute to the overall guidelines?

4. Is the language appropriate (understandable, no jargon, engaging)?
5. Is it too long or too short?
6. Should something be removed or added?
7. Does your impression vary between the two guidelines?

9. What is your overall evaluation of section 4 (the ethical operational requirement)? For each sub-section:
    1. Is it well-adapted for practitioners?
    2. Is the language appropriate (understandable, no jargon, engaging)?
    3. Can it be properly applied?
    4. Is it too long or too short?
    5. Is there something that needs to be changed?
    6. Are there important issues not covered?
    7. Do the guidelines require something they should not require?
    8. Are the proposals linked to the correct phases of development/management and governance?
    9. Does your impression vary between the two guidelines?

10. What is your evaluation of section 5 (special topics)?
    1. Is it well-adapted for practitioners?
    2. Does it contribute to the overall guidelines?
    3. Is the language appropriate (understandable, no jargon, engaging)?
    4. Is it too long or too short?
    5. Should something be removed or added?
    6. Does your impression vary between the two guidelines?

Appendix E

Interview Questions on Regulatory options (T3.3.)

Goal: Solicit feedback on selected regulatory proposals for AI and big data

Target attendees: policy makers, civil society, legal scholars.

Input: D3.3/policy brief/exec summary of D3.3/extract of relevant info

Questions for discussion:

1. What are three high-risk, high-human rights impact AI/big data fields and/or applications that could benefit from stricter regulation?

2. Of the following international options, which three do you find most promising? Why?

    - Moratorium on lethal autonomous weapons systems
    - Binding Framework Convention for AI
    - Legislative framework for independent and effective oversight
    - Legal for human rights impact assessments on AI systems
    - Convention on human rights in the robot age
    - CEPEJ European Ethical Charter
    - International Artificial Intelligence Organization
    - Global legal AI and/or robotics observatory

3. Of the following EU-level options, which three do you find most promising? Why?
    - EU-level special list of robot rights
    - Adoption of common Union definitions
    - Creating electronic personhood status for autonomous systems
    - Establishment of a comprehensive Union system of registration of advanced robots
    - General fund for all smart autonomous robots

- Mandatory consumer protection impact assessment
- EU Taskforce of field specific regulators for AI/big data
- Algorithmic Impact Assessments under the GDPR
- Voluntary/mandatory certification of algorithmic decision systems

4. Of the following national options, which three do you find most promising? Why?
    - DEEP FAKES Accountability Act (US)
    - Algorithmic Accountability Act (US)
    - Canadian Directive on Automated Decision-Making
    - US Food and Drug Administration regulation of adaptive AI/ML technology
    - New statutory duty of care for online harms
    - Redress by design mechanisms for AI
    - Register of algorithms used in government
    - Digital Authority (UK)
    - Independent cross-sector advisory body (CDEI)
    - FDA for algorithms (US)
    - US Federal Trade Commission to regulate robotics

5. Of the following cross-over options, which one do you find most promising? Why?
    - Using anti-trust regulations to break up big tech and appoint regulators
    - Three-level obligatory impact assessments for new technologies
    - Regulatory sandboxes

6. What immediate regulatory actions are necessitated at the:
    1. International level

2. EU-level
3. National level

7. Should there be an international ban on the development/use of lethal autonomous weapons systems?

8. How can we strike a balance between enabling beneficial AI and risk mitigation? What will support this?

9. One key recommendation for AI and big data regulation emerging from SHERPA results is "smart mixing for good results" – is this feasible? If yes, how can this be achieved? Smart mixing refers to using a good combination of instruments, i.e., technical, standards, law and ethical that will offer complementarity, agility and flexibility needed to address the challenges of AI.

10. How can the law further support super-secure AI where it has high likelihood and high severity of risk and impact on rights and freedoms of individuals, especially vulnerable populations – children, minorities and the elderly?

11. What critical future developments need consideration in discussions/actions on the regulation of AI and big data?

Should we also consider given the current developments:

Should there be a ban on the use of facial recognition in public places? Why?

Appendix F

Interview questions on Terms of reference for new/bespoke regulator (T3.6)

Prepared by TRI

Objective: explore the feasibility of a bespoke new regulator for AI and big data at the EU and/or Member State levels and what its terms of reference should include
Target audience: experts including regulators, policy-makers and other relevant stakeholders. *Please ensure equal male/female ratios.*
Location: online/F2F
Dates: TBD

Background:
Should there be a new/bespoke regulator for AI and big data at the EU and/or national levels are questions the SHERPA project is currently deliberating. There are pulls and pushes to the creation of new regulators/regulatory bodies at the international, EU and national level. At the EU-level, the European Parliament request to the European Commission to consider the designation of a European Agency for robotics and artificial intelligence to provide the technical, ethical and regulatory expertise needed to support the relevant public actors, at Union and Member State level, in their efforts to ensure a timely, ethical and well-informed response to the new opportunities and challenges, in particular those of a cross-border nature, arising from technological developments in robotics, such as in the transport sector was not taken up by the European Commission. At the national level, new bodies have been created in countries such as the UK (e.g., the Centre for Data Ethics and Innovation which is tasked with connecting policymakers, industry, civil society, and the public to develop the right governance regime for data-driven technologies) are in the process of being set up (Regulatory Horizons Council to co-ordinate policy and regulation in areas of rapid technological advances in the UK) or proposed (e.g., an FDA for algorithms, calls in Netherlands for a national algorithm watchdog, Digital authority to co-ordinate regulators in the digital world).
SHERPA invites your inputs and feedback.

Questions for discussion (moderator to adapt and use)*:*

1. Do we need a new/bespoke regulator for AI and big data at the EU level? (yes, why; no, why; undecided)
2. Do we need a new regulator for AI and big data at the Member State level? (yes, why; no, why; undecided)
3. Why do you think a new regulator might be necessary? What gap would it address?
4. What type of regulator should this be ? (field-specific/general? Independent watchdog? Licensing body/authority? Inspectorate? Public sector/private sector/general? Professional regulator? Professional conduct authority? An EU regulators network? Supervisory agency? Statutory registration board; Commissioner; AI and big data standards agency; AI fundamental right protection agency? EU/national task force/Digital Authority)
5. What would/should it regulate? E.g., use of autonomous weapons? Human rights? Algorithms ? use/implementation?
6. What would be its legal basis?
7. What should its functions and tasks be?
8. What powers should it have?
9. What would be its role and responsibilities?
10. How should it be constituted? Who should its members be?
11. What should its conduct provisions be?

12. How would it operate? Discuss operation and procedural rules.
13. How would it be governed? How would it be funded? To whom would it report? e.g., the European Parliament?
14. How often should its terms of reference be reviewed?
15. What would be some challenges and barriers to its success? (creation and implementation – political will? regulatory creep/mission drift? Funding? Capacity, lack of independence, lack of teeth, competing priorities and conflicts; regulatory capture)
16. How could these be overcome?
17. Any other comments/considerations that need to be taken into account.

Appendix G

Visualisation of child codes of suggestions for addressing ethical Issues related to Big Data and AI (section 5.3.)