SHERPA

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

Deliverable No. 2.4: Delphi Study Report

30 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	D2.4 Delphi Study Report
WP/Task Related	WP2: Stakeholder analysis and consultation
Delivery Date	October 2020
Dissemination Level	PU
Lead Partner	TRI
Contributors	Nicole Santiago (Trilateral Research); Bernd Stahl (De Montfort University); Paul Keene (De Montfort University); Tally Hatzakis (Trilateral Research); Rowena Rodrigues (Trilateral Research); David Wright (Trilateral Research).
Reviewers	Basic QA by DMU, Trilateral and a professional editor
Abstract	This deliverable describes a three-stage Delphi study undertaken by the SHERPA consortium. It provides the justification for choosing a Delphi study and explains the process of developing the research instrument and the selection of participants. Each of the three stages of the Delphi study is described in detail. Data is presented and analysed and overall conclusions are drawn.
Key Words	Delphi study, artificial intelligence, smart information systems, ethics, human rights

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
0.1	23.10.2020	Nicole Santiago	Basic QA	First Draft
0.2	30.10.2020	Nicole Santiago		

Table of Contents

Executive Summary

Purpose

The SHERPA project undertook a Delphi study on ethical and human rights issues of smart information systems (SIS), i.e. systems drawing on and containing artificial intelligence (AI) and big data analytics.

The Delphi study ran between July 2019 to October 2020, in parallel or subsequent to a number of other activities with similar aims, including a set of case studies, scenario development, technical investigations into cybersecurity, and the exploration of possible mitigation options. The Delphi study aimed to develop insights into the following questions:

1. What are the most important ethical and human rights issues in AI and big data?
2. What are the approaches that are currently used to address these issues?
3. What are the problems with these current approaches?
4. Which suggestions exist that might be better suited to address these problems and whose responsibility is it to address them?
5. What would be an appropriate set of priorities to implement these approaches?

The purpose of the Delphi study was to validate findings and insights from across the project and provide input into the development of the project's recommendations.

The study was carried out in three rounds. Round 1 (R1) consisted of five open-ended brainstorming questions about the most important ethical or human rights issues raised by SIS and measures to address those issues. From those responses, a list of issues and measures was generated, supplemented by input from other SHERPA activities; the measures were categorised as either regulatory, technical, or 'other'. In Round 2 (R2), respondents were asked to begin narrowing down the issues and approaches by rating them against a set of three criteria. The top-scoring measures were presented to the respondents in Round 3 (R3), where they were asked to select the three most important measures for immediate action. By the end of the study, the top-scoring ethical and human rights concerns (from R2) and the top-ranking potential governance measures (from R3) were identified.

Key findings

In R1, the most prominent issues were lack of transparency, lack of human decision-making, lack of privacy, and discrimination. Regulation, specifically at the European level, was the most frequently discussed measure. There were also many mentions of 'other' measures, such as ethical frameworks, guidelines, and toolkits.

In R2, **lack of privacy** and **bias and discrimination** continued to be issues of high concern, along with **misuse of personal data**, lack of access to (and limitations on) **freedom of information**, and impacts on **democracy**. Other key concerns included **violation of human rights** for end-users, loss of freedom and **individual autonomy**, impacts on **power relations** (political and economic), **lack of transparency** and trust, **potential for criminal and malicious use**, and **disappearance of jobs**. While not cited in R1, the **environmental impact of SIS** was also among the key concerns. Responses about potential governance measures, however, were quite different from R1. Regulatory measures were the lowest scoring, with no regulatory measures making into the top fifteen measures overall. More promising were technical measures, which were the most desirable on average. Most promising overall were 'other' measures, which scored highest, and which were twelve of the top fifteen measures overall.

In R3, the top three potential governance measures were:

- **Methodologies for systematic and comprehensive testing of AI-based systems**
- **Framework, guidelines, and toolkits for project management and development**
- **Stakeholder dialogue and scrutiny**

In explaining the selections, many responses focused on the need for **public awareness**, **enhanced transparency into AI systems**, and **clarification about requirements**. Additionally, the need to **translate abstract norms into operationalised practice** was highlighted. In identifying who should be responsible for implementing the measures, a wide array of actors was cited, illustrating the **extensive ecosystem for AI governance**. Explicit references to **preventing/mitigating harms and accountability were notably absent**.

Key conclusions

The Delphi study demonstrated that SHERPA has a good overview of ethical and human rights issues and currently discussed mitigation strategies. The consistent themes over the three rounds are familiar, including concerns about lack of transparency, impact of bias and discrimination in AI systems, and a need for more public awareness. These findings are consistent with research and findings in other SHERPA activities, including stakeholder interviews, focus groups, online survey, and feedback from the stakeholder board. The top responses focused on well-known and well-documented concerns currently impacting end-users in Europe.

The study identified important issues and potential strategies, but was most useful as an illustration and mapping of the complexity of the concerns associated with SIS and the potential governance measures to address those concerns. The breadth of responses illustrates diverse opinions and knowledge about the most pressing concerns and possible solutions – even among experts. While there was no overwhelming consensus on which solutions to prioritise, it is clear that the complexity of the SIS ecosystem requires a 'smart mix' of measures and all stakeholders have roles to play.

The results of prioritising the most important potential governance measures for immediate action are consistent with SHERPA research, and reflect the difficulty of determining consensus on how to address concerns. The Delphi affirmed that the possible solutions are plentiful, but rarely clear. This study has highlighted how critically important it will be in implementing governance measures for SIS to carefully and clearly frame language and articulate precise recommendations for discrete audiences. This is a challenge not only for SHERPA, but for all stakeholders, and will inform the further development of SHERPA's final recommendations.

Like the Delphi results, the SHERPA project will prioritise stakeholder engagement, educational tools (tailored to different stakeholder groups), and concrete tools to translate principles into practice. However, unlike the Delphi panel, the SHERPA project (based on its research in its other activities) is recommending a stronger regulatory framework at the EU level, as well as an EU Agency for AI, impact assessments, standardisation on AI ethics, and the establishment of AI 'ethics' officers within organisations.

List of figures

Glossary of terms

Term	Explanation
Delphi Study	Well-established methodology to find solutions to complex and multi-faceted problem, carried out via multiple rounds of surveys sent to a panel of participants
Delphi team	SHERPA partners who contributed to administering and analysing the Delphi study
Panel	Experts invited to take part in the Delphi study
Respondents	Experts who provided responses to the Delphi study
SIS	The combination of Artificial Intelligence and big data analytics

1. Introduction

The Delphi study undertaken in the SHERPA project[1] forms part of Work Package 2 (WP2), Stakeholder analysis and consultation. The Delphi study follows the activities of Work Package 1 (WP1), which provided descriptions and visualisations of ethical and human rights issues of SIS via case studies, scenarios, technical, ethical and legal analysis. The Delphi study ran between July 2019 to October 2020, thus allowing it to contribute to Work Package 3 (WP3), Responsible Development of SIS, and Work Package 4 (WP4), Evaluation, validation and prioritisation.

In both timing and content, the Delphi study partly overlapped with the online survey in Task 2.3.[2] However, where the online survey collected broad input from a larger number of stakeholders, the Delphi study's aim was to provide more detailed insights from a smaller number of experts.

The purpose of the Delphi Study was described as follows in the Description of Action (DoA):

> "Delphi studies are a well-established methodology to find solutions to complex and multi-faceted problems (Adler and Ziglio, 1996; Dalkey et al., 1969; Linstone et al., 1975). Based on the successful deployment of the Delphi method in the Responsible-Industry project, **SHERPA will use a Delphi study to gather feedback on the prioritisation of options.**
>
> The Delphi will be based on the first draft of the SIS workbook (WP3) which will include both extant suggestions and novel contributions by the SHERPA consortium. A specific focus of the Delphi study will be on the forward-looking aspects of WP1, in particular, the scenarios (Task 1.2) which will be used as the basis of the design of the Delphi study. The Delphi method will allow the invited experts to work towards a mutual agreement by responding to a set of questions. Their responses will form the basis of a synthesis paper and the next round of questions. The experts' responses shift as rounds are completed based on the information brought forth by other experts participating in the analysis. Due to the methodology that responses of the participating experts are anonymous, the involved individuals do not need to have concerns about repercussions for their attitudes and convictions. Consensus can be reached over time as opinions are swayed. The results of the stakeholder interviews and online survey (T2.2, 2.3) will feed the development of the questionnaires used for the Delphi rounds.
>
> Our Delphi study will consist of three rounds of questions, starting with an open and more qualitative round that will be used to identify options that can then be narrowed down and quantified. The Delphi study will comprise about 60 experts from a representative range of stakeholders selected from Task 2.1. The outcome of the Delphi study will be reflected in the final version of the SIS workbook."

A Delphi study[3] is typically described as a future-oriented methodology, one example of future and foresight research[4] which includes numerous other methodologies, such as scenario development, citizen

[1] SHERPA Project: https://www.project-sherpa.eu/.
[2] Brooks, Laurence; Stahl, Bernd; Jiya, Tilimbe (2020): D2.3 Online survey report. De Montfort University. Online resource. https://doi.org/10.21253/DMU.11777478.
[3] Dalkey, N. C., B. B. Brown, and S. Cochran, *The Delphi Method: An Experimental Study of Group Opinion*, Rand Corporation Santa Monica, CA, 1969. http://192.5.14.43/content/dam/rand/pubs/research_memoranda/2005/RM5888.pdf.; Linstone, H. A., M. Turoff, and O. Helmer, *The Delphi Method: Techniques and Applications*, Addison-Wesley Publishing Company, Advanced Book Program, 1975. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=256068.
[4] Sardar, Ziauddin, "The Namesake: Futures; Futures Studies; Futurology; Futuristic; Foresight—What's in a Name?", *Futures*, Vol. 42, No. 3, April 2010, pp. 177–184.

panels or simulations.[5] Delphi studies have been identified as a particularly useful tool to support policy development.[6] The use of the Delphi study in the SHERPA project is aligned with this aim. The SHERPA Delphi constitutes a key activity in the process of identifying and shaping policy recommendations based on the insights produced by other activities of the project.

According to Ziglio, there are three key considerations for the application of Delphi studies to a policy problem:

1. "the problem does not lend itself to precise analytical techniques but can benefit from subjective judgements on a collective basis, [...];
2. the problem at hand has no monitored history nor adequate information on its present and future development [...];
3. addressing the problem requires the exploration and assessment of numerous issues connected with various policy options where the need for pooled judgement can be facilitated by judgmental techniques [...]"[7]

These considerations align well with the SHERPA Delphi, which focuses on the identification, evaluation and prioritisation of technological and regulatory options for the ethical and socially responsible development and deployment of AI and big data analytics which, despite significant research and policy efforts, remain unclear and contested.

More specifically, the SHERPA Delphi focuses on the following research question: **Which should be the priorities in addressing ethical and human rights issues in AI and big data?** This question will be addressed via the following sub-questions:

1. What are the most important ethical and human rights issues in AI and big data?
2. What are the approaches that are currently used to address these issues?
3. What are the problems with these current approaches?
4. Which suggestions exist that might be better suited to address these problems and whose responsibility is it to address them?
5. What would be an appropriate set of priorities to implement these approaches?

It is important to underline that the expectation of a Delphi study is not that it creates new knowledge in a traditional scientific sense, but that it aims to make best use of existing knowledge and the collective wisdom of the participants.[8] This aligns with the SHERPA work plan, which has previously created empirical and conceptual insights in WP1, and will use the collective wisdom of the experts to inform the options it develops and how they will be presented. The SHERPA Delphi aims towards the creation of an expert consensus that can then be used for further consultation with decision-makers.

This Deliverable provides an account of all stages and findings of the Delphi study. It starts with the study protocol, which contains the plan for the study. Delphi studies typically involve a number of experts, who are unaware of each other, to avoid undue influence and biases. Participant responses are anonymised and are communicated so that individuals are freed from concerns about repercussions for their attitudes and convictions. Consensus, or at least a clarification of the existing positions, can be reached over time

[5] Georghiou, Luke, Jennifer Cassingena Harper, Michael Keenan, Ian Miles, and Rafael Popper, *The Handbook of Technology Foresight: Concepts and Practice*, Edward Elgar Publishing Ltd, 2008.
[6] Adler, Michael, and Erio Ziglio, eds., *Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health*, Jessica Kingsley, London, 1996.
[7] Ziglio, Erio, "The Delphi Method and Its Contribution to Decision Making", in Michael Adler and Erio Ziglio (eds.), *Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health*, Jessica Kingsley, London, 1996, pp. 3–33.
[8] Sandrey, Michelle A., and Sean M. Bulger, "The Delphi Method: An Approach for Facilitating Evidence Based Practice in Athletic Training", *Athletic Training Education Journal*, Vol. 3, No. 4, October 1, 2008, pp. 135–142.

as opinions are swayed. The SHERPA Delphi study consists of three rounds of questions, starting with an open and more qualitative round that will be used to identify options that can then be narrowed down and quantified.

This Deliverable describes and justifies the research design, provides the analysis of all three rounds of analysis, and draws conclusions from the study.

2. Research design

This section outlines: a) the Delphi process in general, b) steps undertaken in the SHERPA Delphi study, and c) participant selection.

2.1 The Delphi process

Our Delphi study allowed experts to work towards a shared understanding of the technological and regulatory options that can be deployed to ensure the ethical and socially responsible development and deployment of AI technologies. Delphi studies are iterative in their approach. Participants are invited to respond to a set of questions. Their responses are then synthesised and inform the next round of questions. Okoli and Pawlowski describe the three phases of a Delphi study as:

1. Brainstorming
2. Narrowing Down
3. Ranking[9]

The SHERPA Delphi study follows this logic, with the aim of arriving at a reduced set of options to feed into the evaluation and prioritisation work undertaken in WP4 to inform the SHERPA recommendations.

The first step of brainstorming used open-ended qualitative questions to understand what the members of the Delphi panel perceived to be the ethical and human rights issues of AI and big data. This covers similar ground to the work undertaken in SHERPA's WP1, where such issues were investigated using case studies, scenarios, technical, ethical and human rights analysis. The first Delphi round (R1) also asked open-ended questions about possible mitigation strategies, some of which are the subject of investigation in SHERPA WP3. The second Delphi round (R2) of narrowing down issues has parallels to the SHERPA online survey, the interviews, and the focus groups of WP4. The final Delphi round (R3), ranking mitigation strategies, was of most immediate importance for the development of recommendations, and coincided with the Consortium discussion of the overall project recommendations.

2.2 Steps in the SHERPA Delphi Study

This section provides a more detailed account of the main stages of the research design and implementation.

2.2.1 Preparation and pilot testing and ethics approval (July – September 2019)

This phase comprised developing, pilot testing and agreeing the Delphi study protocol internally to ensure that questions are suitable, and understandable to external participants. In accordance with standard practice of social science, each of the Delphi rounds was pilot tested. This means that the survey instrument was checked for comprehensibility and usability. This was done by first circulating the survey in the Consortium and asking for feedback. Following this, the survey was tested by selected experts.

Ethics approval: Ethics approval was gained by the SHERPA coordinator from De Montfort University's Research Ethics Committee, Faculty of Computing, Engineering and Media. The ethics approval number 1920/519 was received on 03.10.2019 (see Appendix C).

[9] Okoli, Chitu, and Suzanne D. Pawlowski, "The Delphi Method as a Research Tool: An Example, Design Considerations and Applications", *Information & Management*, Vol. 42, No. 1, December 1, 2004, pp. 15–29.

2.2.2 Delphi Round 1 (October 2019 – January 2020)

The first round (R1) of the Delphi study asked respondents to brainstorm the issues raised by SIS, and ways to address these issues. The survey consisted of a set of five open questions designed to reflect and support the work undertaken in WP1 on the key issues arising from the uses of AI and Big Data:

1. What do you think are the three most important ethical or human rights issues raised by AI and / or big data?
2. Which current approaches, methods, or tools for addressing these issues are you aware of?
3. What do you think are the pros and cons of these current approaches, methods, or tools?
4. What would you propose to address such issues better?
5. Which should be the top 3 criteria for society to select and prioritise the most appropriate measures?

The survey was designed by the SHERPA Consortium, and following initial review by Consortium members, pilot tested by 27 Consortium members and 38 external advisers. Following ethics approval by De Montfort University, the panel was invited, via emails beginning November 13, 2019 , to take part in the study and begin the R1 survey (see Appendix A). Follow-up requests were sent to the panel on December 4, 2019, and January 10, 2020. Discounting as many duplicates as possible, 100 experts clicked to "Begin the Delphi Study" and "agree[d] with the use of my responses for research purposes of the SHERPA project as outlined above" on the redirected survey page. The R1 survey closed on January 15, 2020, after receiving 145 responses. Following review of the data and data cleansing, 41 responses contained sufficient information to warrant analysis.

The SHERPA team analysed all responses and synthesised them into a 14-page summary report,[10] presented here in Sections 3.1.2 R1 Summary of Responses and 3.1.3 Key Findings, and available on the SHERPA website. As part of the analysis, the potential governance measures identified by respondents were grouped into three categories: regulatory, technical, and 'other'. The general insights from R1, supplemented by other SHERPA activities, directly informed the creation of questions asked in R2; particular attention was paid to notable omissions. A link to the R1 summary report was sent to respondents in the email invitation for R2.

Additionally, a longer report of R1 responses was also prepared (Appendix D) and the raw data results are publicly available.[11]

2.2.3 Delphi Round 2 (March – August 2020)

The purpose of the second round (R2) of the Delphi study was to begin narrowing down the issues and approaches identified in the R1 brainstorming. R2 consisted of four sets of questions, asking participants to rate (on a scale of 1-5) issues and potential measures across three criteria:

1. Rate a list of **ethical and human rights issues** in terms of reach, significance, and attention.
2. Rate a list of **potential regulatory measures** in terms of desirability, feasibility, and probability.
3. Rate a list of **potential technical measures** in terms of desirability, feasibility, and probability.
4. Rate a list of **other potential measures** in terms of desirability, feasibility, and probability.

[10] Santiago, Nicole, *SHERPA Delphi Study - Round 1 Results*, Project Deliverable, SHERPA project, 2020. https://www.project-sherpa.eu/wp-content/uploads/2020/03/sherpa-delphi-study-round-1-summary-17.03.2020.docx.pdf.

[11] SHERPA T2.4 Delphi study raw data Round 1, available for download: https://dmu.figshare.com/articles/online_resource/T2_4_Delphi_study_raw_data_1/13128539.

The list of issues and approaches was based on the responses in R1, supplemented with additional issues and measures identified in SHERPA WP1 (D1.1 Case Studies[12]), WP2 (D2.3 Online Survey[13]), and WP3 (D3.3. Report on Regulatory Options[14] and D3.5 Technical options and interventions interim report[15]).

R2 was pilot tested by members of the SHERPA Consortium. On March 18, 2020, the survey was sent to the Delphi panel. A follow-up request was sent on May 12, 2020 (see Appendix A). The survey closed at the end of June 2020. Following review of the data and data cleansing, 26 responses contained sufficient information to warrant analysis.

The SHERPA team analysed all responses and synthesised them into a 6-page summary report,[16] presented here in Section 3.2.3 R2 Key Findings, and available on the SHERPA website. A longer summary of the results, broken down by question, is presented in Section 3.2.2 R2 Summary of Responses. A link to the R2 summary report was sent to respondents in the email invitation for R3. The highest scoring potential governance measures were isolated for prioritisation in R3.

The raw data results from R2 are publicly available.[17]

2.2.4 Delphi Round 3 (September – October 2020)

The final round of the Delphi study (R3) was designed to determine consensus on the prioritisation of potential governance measures. Respondents were asked to select the three most important potential governance measures for immediate action, from the list of fifteen highest scoring measures in R2. For each selection, respondents were prompted to explain: (a) why the measure is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure. The respondents did not rank their selections, therefore the order of selections was not relevant. Respondents were also given the option to identify any potential governance measures that should not be prioritised, as well as to provide any additional comments.

R3 was pilot tested by members of the SHERPA Consortium. The survey was sent to the panel on September 18, 2020. Follow-up requests were sent September 25 and October 1, 2020 (see Appendix A). The survey closed on October 7, 2020. Forty-three (43) respondents provided feedback, for a total of 117 discrete selections (not all respondents selected three options). Some respondents answered the follow-up and additional questions.

The results and analysis of R3 are presented only in this report. The raw data results from R3 are publicly available.[18]

[12] Macnish, Kevin; Ryan, Mark; Gregory, Anya; Jiya, Tilimbe; Antoniou, Josephina; Hatzakis, Tally; et al. (2019): D1.1 Case studies. De Montfort University. Online resource. https://doi.org/10.21253/DMU.7679690.v3

[13] Brooks, Laurence; Stahl, Bernd; Jiya, Tilimbe (2020): D2.3 Online survey report. De Montfort University. Online resource. https://doi.org/10.21253/DMU.11777478.

[14] Rodrigues, Rowena; Laulhe Shaelou, Stephanie; Lundgren, Björn (2020): D3.3 Report on regulatory options. De Montfort University. Online resource. https://doi.org/10.21253/DMU.11618211.v4.

[15] Kirichenko, Alexey; Marchal, Samuel (2020): D3.5 Technical Options and Interventions (Interim report). De Montfort University. Online resource. https://doi.org/10.21253/DMU.12973031.v1.

[16] Santiago, Nicole, *SHERPA Delphi Study - Round 2 Results*, Project Deliverable, SHERPA project, 2020. https://www.project-sherpa.eu/wp-content/uploads/2020/09/sherpa-d2.4-r2-summary-report-03.09.2020-final.pdf.

[17] SHERPA T2.4 Delphi study raw data Round 2, available for download: https://dmu.figshare.com/articles/online_resource/T2_4_Delphi_study_raw_data_2/13128572.

[18] SHERPA T2.4 Delphi study raw data Round 3, available for download: https://dmu.figshare.com/articles/online_resource/T2_4_Delphi_study_raw_data_3/13128581.

2.3 The selection of participants

In total, **231 experts were identified and invited to join the Delphi panel**. The experts were sourced from a) the SHERPA Stakeholder Board, b) the SHERPA contact list, and c) invitees from partners' networks. All participation was on a volunteer basis. Fifty percent of the experts on the panel were women. All 231 experts were contacted during R1 by email with an invitation to the R1 survey. Discounting as many duplicates as possible, **100 experts agreed to the use of their responses** (by clicking in the email to "Begin the Delphi Study" and clinking to "agree with the use of my responses for research purposes of the SHERPA project as outlined above" on the redirected survey page).

2.3.1 Criteria for selection

Delphi studies are not aimed at drawing conclusions about populations, so do not require sampling considerations with regards to questions of representativeness of the panel composition. The goal is to identify a sufficiently large number of options requiring diverse viewpoints, with the aim of ensuring that all relevant issues are taken into consideration.[19] While the number of participants is not crucial for the success of a Delphi study, the composition of the panel is of high importance. Hence, participants were selected on the basis of their expertise and ability to contribute to the topic while taking into account other criteria, such as gender balance, geographic diversity, and representation of different stakeholder groups (e.g., policymakers, technologists, business people, academics, civil society organisations). The list of invitees was constructed with a view to these aspects, and the Consortium explicitly aimed to ensure a gender balance among the respondents.

2.3.2 Recruitment process

Participants were recruited in several ways. First, all members of the SHERPA Stakeholder Board were invited to participate. The Stakeholder Board includes selected individuals who have expertise in some aspects of ethical and human rights issues of AI and big data. They are also long-term partners of the project and therefore likely to participate. In a second step, an open invitation was extended to the contact list asking for volunteers for the Delphi Study (see Appendix A). Based on the criteria, participants were selected from:

- SHERPA Stakeholder Board
- SHERPA Stakeholder list
- Volunteers who were asked to sign up using the SHERPA newsletter
- Other relevant sources including:
    - Participants of the 100+ brilliant women in ethics and AI event (Oxford, September 16, 2019)
    - High-Level Expert Group on AI
    - Relevant projects such as:
        - Humane AI[20]
        - Innovation Center for Artificial Intelligence[21]
        - Finnish Center for Artificial Intelligence[22]

[19] Goldschmidt, Peter, "A Comprehensive Study of the Ethical, Legal and Social Implications of Advances in Biomedical and Behavioural Research and Technology", in Michael Adler and Erio Ziglio (eds.), *Gazing into the Oracle: The Delphi Method and Its Application to Social Policy and Public Health*, Jessica Kingsley, London, 1996, pp. 89–130.
[20] Humane AI: https://www.humane-ai.eu/.
[21] Innovation Center for Artificial Intelligence: https://icai.ai/.
[22] ish Center for Artificial Intelligence: https://fcai.fi/

- Responsible Robotics[23]

Second, the Consortium targeted experts with complementary profiles to ensure that the final composition of the Delphi panel is balanced in terms of expertise, age, geographical distribution and gender.

[23] Responsible Robotics: https://responsiblerobotics.org.

3. Findings and Analysis

3.1 Round 1

The first round of the Delphi study (R1) consisted of a set of open questions that aimed to determine expert opinions on the following points:

1. What do you think are the three most important ethical or human rights issues raised by AI and / or big data?
2. Which current approaches, methods, or tools for addressing these issues are you aware of?
3. What do you think are the pros and cons of these current approaches, methods, or tools?
4. What would you propose to address such issues better?
5. Which should be the top 3 criteria for society to select and prioritise the most appropriate measures?

3.1.1 Analysis Method

Every response was read and analysed. Incomplete response sets (e.g. respondent only answered Q1-3) were reviewed and included. Duplicate response sets were only counted once.

Within each question, recurring key themes (or keywords) and sub-themes were identified, and responses were coded with the relevant themes or keywords. Responses that contained two distinct themes were counted twice. Responses that belonged to a different question were analysed within the question that was most logical (e.g. answer with *proposals* for regulatory measures was analysed under Q4, not where it was written in Q2).

The Round 1 Response Report (see Appendix D) is a more detailed summary of the responses, including outlier responses that did not fit within the categories. The number of relevant responses, as well as any particularly clear or insightful quotes, were integrated. When a response provided a specific example (e.g. law), that specific example has been included; absence of further explanation of an example means there was no additional information provided in the response. Aside from citations for specific examples provided in the responses, no information (explanatory or analysis) was added. To organise the analysis of potential measures, responses were categorised into regulatory, technical, and 'other' measures.

3.1.2 Summary of Responses

Question 1: What do you think are the three most important ethical or human rights issues raised by AI and / or big data?

There were 41 responses to Q1. One Q1 response was deemed more relevant to another question, and two responses to other questions were deemed more relevant to Q1. Therefore, a total of 42 responses were analysed under Q1. Lack of transparency, lack of privacy, bias and discrimination, and loss of human decision-making were the most frequently mentioned concerns. **Lack of transparency** was identified as a concern in that the average citizen does not understand how AI and data systems work, nor do they understand how decisions are made by SIS that affect them in their daily lives; a need for transparency (and explainability) about the sources of data and the decision-making processes was clearly expressed. When discussing **privacy concerns**, respondents tended to focus on the vast amounts of personal data collected, specifically raising concern about real-time surveillance. Regarding **bias and discrimination**, respondents were concerned with built-in and entrenched bias caused by AI systems that reproduce bias, both of which may produce unfair and/or unequal decisions that violate the right to equality. Concerns

about the impact of new technologies on humanity and the value of **human decision-making** were articulated in various ways, but most common was anxiety that humans were being taking "out of the loop" on critical decision-making as machine-intelligence is privileged, resulting in 'depersonalized' decisions and a perceived loss in human intellect.

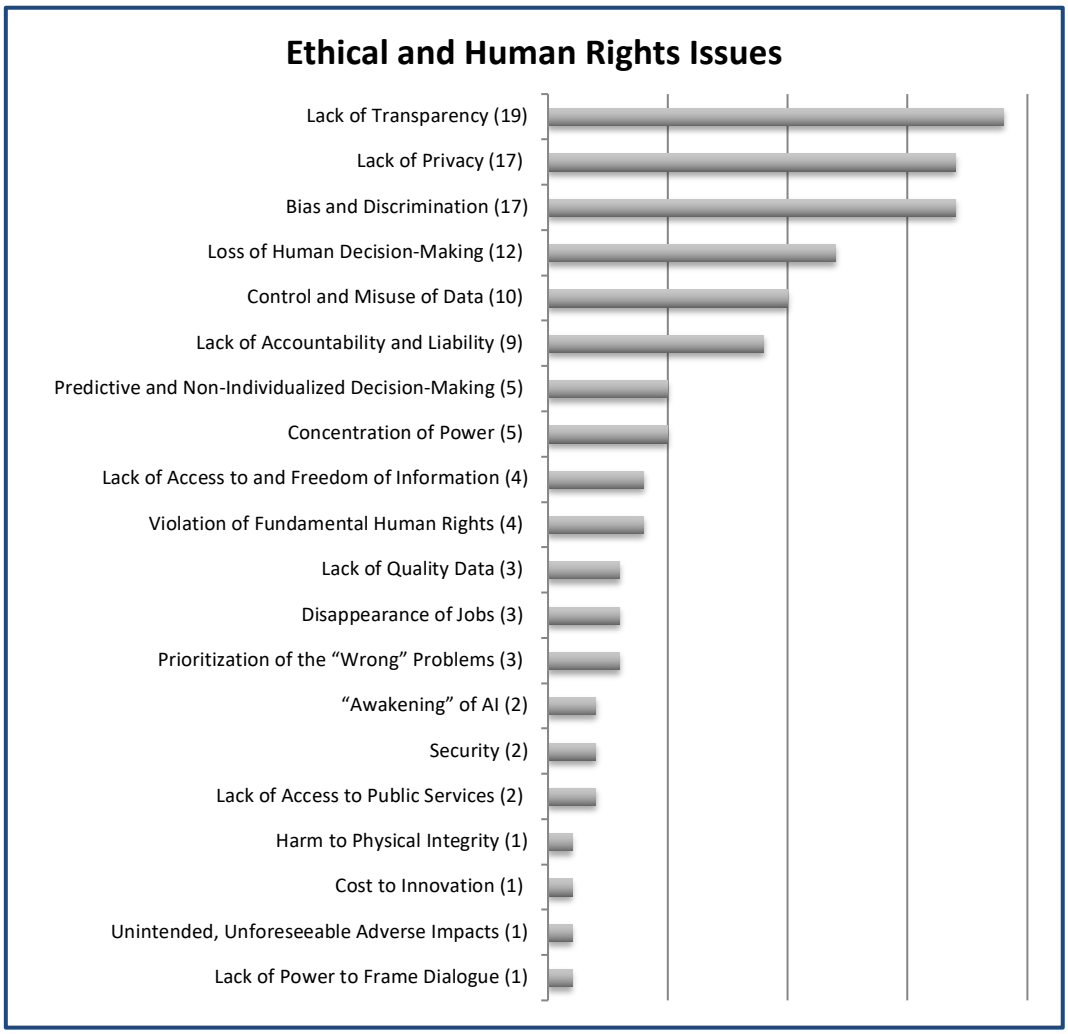

Figure 1: Ethical and human rights issues (R1)

Concerns about a **lack of accountability** and the **misuse of personal data** were also cited by a number of respondents. In regard to accountability, respondents called for a clear definition of legal responsibility for all actors, including AI systems. When discussing the misuse of personal data, respondents expressed specific concerns about abuse (e.g. mass surveillance), control, ownership, and commercialisation of data.

The remaining responses were each mentioned by only a couple respondents. Particularly notable was that harm to physical integrity was only mentioned once in relation to self-driving cars and autonomous weapons.

Question 2: Which current approaches, methods, or tools for addressing these issues are you aware of?

Category	Measures
Regulatory Measures	• Regulations (18)* • Public Register of Permissions to Use Data (1) • Reporting Guidelines (1) • Monitoring Mechanism (2)
Technical Measures	• Testing Algorithms on Diverse Subsets (1) • Using Analytics Systems to Judge Whether Decisions Are Equal/Fair (1) • Generative Adversarial Networks and Other Techniques for Deriving Explanations from Outcomes (1) • More Open Data (2)
Other Measures	• Codes of Conduct (3) • Education Campaigns (4) • Employing 'Fairness' Officer or Ethics Board (3) • Frameworks, Guidelines, and Toolkits (14) • Grievance Mechanism (1) • High-Level Expert Groups (6) • Individual Action (2) • International Framework (3) • Investigative Journalism (3) • NGO Coalitions (1) • Open Letters (1) • Public Policy Commitment (1) • Self-Regulation (1) • Stakeholder Dialogue and Scrutiny (3) • Standardisation (3) • Third-Party Testing and External Audits (2)
*Indicates number of respondents who explicitly referenced the measure	

Figure 2: Current governance measures (R1)

There were 36 responses to Q2. Two responses were deemed more relevant to another question. Therefore, a total of 34 responses were analysed under Q2.

There were very few responses identifying current approaches, methods, or tools at the international level. No respondent identified an international law instrument, and some noted the practical limitations of creating and implementing an international approach. When identifying approaches, methods and tools at the regional level, all examples cited referred to the European Union, and most frequently to the GDPR. At the national government level, the majority of responses referred to measures in Western Europe; only three responses concerned the United States and one concerned Hong Kong. National laws were the most frequently cited, but other specific examples cited included national policies and frameworks, and national education campaigns.

There was a greater variety of measures referenced that were developed by industry, NGOs, and civil society (including academia). A number of specific initiatives were included that had been created both by private-sector actors alone (e.g. Google), and in partnership with other stakeholders (e.g. Partnership for AI). It is worth noting in Q3 that there were no critiques of industry-driven initiatives like company codes of ethics or toolkits. From NGOs and civil society (which includes academia), specific measures

cited included educational tools, ethical guidance and frameworks, NGO coalitions, and an open letter signed by famous AI scientists and experts. One respondent cited a report by ETH Zurich[24] that found there are 84 projects and organisations working on AI issues, suggesting that there is a proliferation of frameworks, potentially leading to further confusion. Lastly, some respondents mentioned the role of journalists to investigate and highlight concerns, and the role that individuals assume to protect themselves (e.g. disabling ads on personal devices).

Question 3: What do you think are the pros and cons of these current approaches, methods, or tools?

There were 31 responses to Q3. Three responses to Q3 were deemed more relevant to another question. Therefore, a total of 28 responses were analysed under Q3. There were far more cons mentioned than pros.

Pros
• Dialogue means we **learn from each other** • Regulation has **power of enforcement** • Transparency measures means **building ethics into the design** • Education **enhances citizen/consumer power** • Ethical Impact Assessments provide **clear methodology & tools** • Standardisation has **objective set of criteria** • Oversight **addresses human rights violations**
Cons
• **Lack of understanding** about roles & responsibilities • **Risk of shifting burden of responsibility** to developers or consumers • Measures are **too abstract** • Creation & implementation is **resource intensive** • Non-binding measures have **no enforcement** • **No comprehensive approach** • **Too complicated** to implement new ways of thinking • Regulation has **limited application** • Technology **development outpaces rule-making process** • Measures **perceived as a hurdle** • Measures are **public-sector focused** • **Difficult to measure ethics objectively** • Educational campaigns ineffective because **don't reach people who need it most**

Figure 3: Pros and cons of current approaches (R1)

The 'pros' focused only on specific types of current measures; for example, one 'pro' of regulation cited was the power of enforcement. Other 'pros' mentioned referred to stakeholder dialogue, transparency efforts, ethical impact assessments, standardisation, and oversight mechanisms.

In contrast, nearly half of respondents identified at least one 'con' of existing measures; there were both general critiques and critiques specific to individual types of measures. A common general critique was that key players do not understand their responsibilities, and therefore do not appreciate the potential impact of their work. One respondent refused to put all the blame on developers, calling out an "apathetic set of consumers." Two other notable critiques were that current measures are too abstract to be

[24] Anna Jobin, 'Ethics guidelines galore for AI – so now what?', ETH Zürich, 17 January 2020, https://ethz.ch/en/news-and-events/eth-news/news/2020/01/ethics-guidelines-galore-for-ai.html.

effective and resource intensive to create and implement. This was one of the only three times that costs were mentioned by respondents.

In addition to general 'cons', respondents also evaluated the limitations of specific current measures. Regulations were the most frequently mentioned, with critiques ranging from their limited scope of application to the fear that they hamper innovation or contribute to compliance-only setting. Multiple respondents also noted that disconnect between the rapid development of new technologies and the slow speed of policy-making processes. Other 'cons' included long and overly complex guidance and hard to measure objectives.

Question 4: What would you propose to address such issues better?

There were 30 responses to Q4. Four responses to other questions were deemed more relevant to Q4. Therefore, a total of 34 responses were analysed under Q4.

Regulatory measures were the most frequently proposed, with regulations being the most common. There were no general themes for regulation, as each respondent proposed something unique (e.g. 'smart mix' of regulatory initiatives; legislation for transparent AI; and recognition of a right to work).

Proposed Measures	
Regulatory Measures	● Regulations (13)* ● Public Register of Permissions to Use Data (1) ● Reporting Guidelines (1) ● Monitoring Mechanism (2)
Technical Measures	● More Open Data (1) ● Use of AI to Protect Data (1) ● Improve Control of Data (1) ● Easily-Explained Algorithms (1) ● Comprehensive AI Example Sets (1) ● Retaining Possibility of Human Override (1)
Other Measures	● Citizen Juries (1) ● Codes of Conduct (1) ● Education Campaigns (11) ● Employing 'Fairness' Officer or Ethics Board (2) ● Ethical Mindset (1) ● Exchange of Best Practices (1) ● Frameworks, Guidelines, and Toolkits (2) ● Grievance Mechanism (1) ● High-Level Expert Groups (1) ● Individual Action (1) ● International Framework (3) ● More Open Source Tools (1) ● Retaining 'Unsmart' Products and Services (1) ● Stakeholder Dialogue and Scrutiny (5) ● Standardisation (1) ● Third-Party Testing and External Audits (2)
*Indicates number of respondents who explicitly referenced the measure	

Figure 4: Proposed measures (R1)

International or regional agreements were mentioned only a few times, which could be seen as either realistic given the difficulty of creating such agreements, or an unfortunate reflection that international agreement is extremely unlikely. These responses were, however, consistent with Q2 and Q3, which focused on regulatory measures at the regional and national level.

Additionally, respondents proposed a broad range of other measures, including technical measures, encouraging collaboration among stakeholders, developing differentiated toolkits, and implementing third-party auditing. One respondent proposed creating 'citizen juries,' which is a novel idea that could be a means to encourage stakeholder dialogue. Many respondents also proposed developing educational and awareness campaigns for all stakeholders at all levels, including children, students, developers and professionals, politicians and government officials, and members of the public generally.

Question 5: What should be the top 3 criteria for society to select and prioritise the most appropriate measures?

There were 31 responses to Q5. About half of respondents identified criteria that should guide the development of new technologies. While not a direct response to the question, the responses provide valuable insight into the types of issues and concerns that should be prioritised when developing and implementing appropriate measures. For example, measures could be developed in such a way that the two most frequently mentioned issues – societal impact and transparency – are addressed. These issues were consistent with the concerns raised in Q1 about lack of transparency and loss of human decision-making.

Figure 5: Prioritisation criteria for appropriate measures (R1)

More traditional criteria for evaluating measures – like costs, feasibility, and effectiveness – were mentioned only a few times. Monetary cost was only mentioned once, and the time required to develop and implement a measure was not mentioned at all. Additionally, the oft-cited concern that regulation would stifle innovation was only mentioned twice.

3.1.3 Key Findings

The most prominent issues running throughout responses to all questions were concerns about a lack of transparency and human decision-making. These concerns were articulated strongly in Q1 (most important ethical or human rights issues) and Q5 (criteria for developing new measures), and were common underlying themes in responses to other questions.

While privacy and discrimination were top-ranking concerns in Q1, they were not mentioned often in responses to other questions.

Regulation was the most frequently discussed 'approach, methods, or tool,' both in terms of criticisms (Q3) and potential solutions (Q4). Given that most respondents are Europe-based, most of the specific examples cited, including the GDPR, were in Europe. Most of the proposed regulatory solutions focused on the regional and national level; international solutions were rarely mentioned in a positive light, suggesting that the respondents do not view an international approach as the most effective.

There was also a lot of discussion of 'other' approaches, methods and tools. Ethical frameworks and toolkits in particular came up frequently, but there seemed to be some tension between those who found them useful, and those who believe they create a confusing hurdle. Perhaps this is due to the fact that a number of projects and organisations have put forward guidelines, frameworks, and toolkits.

There were a number of notable omissions in the R1 responses. In regard to ethical and human rights concerns, respondents were primarily focused on the immediate issues impacting end-users in Europe. For example, only one respondent discussed concerns related to physical integrity, despite the potential injury (or death) that could be caused by technologies like self-driving cars and autonomous weapons. Additionally, security, reliability, and trustworthiness scored very low. Furthermore, a number of related ethical and human rights issues were not referenced. For example, there was no discussion of the ethical and human rights abuses suffered by those extracting the resources and manufacturing the devices that enable SIS technologies to function; the long-term impact to physical and physiological health from using SIS devices and technologies; or the environmental impact of the manufacture, storage, and disposal of the devices that enable SIS.

In regard to current and proposed measures, there was a lack of reference to existing human rights law and mechanisms. There was also no specific mention of the creation of a new regulator or regulatory body. While respondents did mention the need for oversight and monitoring bodies, responses were generally vague about the structure and responsibilities of those bodies. Additionally, in general, the focus of responses was on existing measures in Europe.

Lastly, the cost of developing governing measures was rarely mentioned. A few respondents referenced the resources needed to develop governing measures, but only one respondent specifically cited the financial costs and no respondent discussed the time needed to develop and implement new measures.

3.2 Round 2

The second round of the Delphi study (R2) consisted of four sets of questions, asking participants to rate (on a scale of 1-5) issues and potential measures across three criteria:

- Rate a list of **ethical and human rights issues** in terms of reach, significance, and attention.
- Rate a list of **potential regulatory measures** in terms of desirability, feasibility, and probability.
- Rate a list of **potential technical measures** in terms of desirability, feasibility, and probability.
- Rate a list of **other potential measures** in terms of desirability, feasibility, and probability.

3.2.1 Analysis Method

Results were first analysed individually by question set: Q1 (ethical and human rights issues); Q2 (potential regulatory measures); Q3 (potential technical measures); and Q4 (other potential measures). Results from Q2-4 were then analysed together to compare all potential governance measures.

The criteria for Question 1 were: **Reach**; **Significance**; and **Attention**. The criteria for Questions 2-4 were: **Desirability**; **Feasibility**; and **Probability**.

To begin, the average scores for each issue or measures in the question set was calculated. For example, on 'Bias and Discrimination' in Question 1, there were 25 responses scoring the issue on the three criteria of reach, significance, and attention. The average of the 25 individual responses was used to calculate reach, significance, and attention scores. The average of the three criteria scores was then used to calculate an overall score for 'Bias and Discrimination'.

In order to better understand the relative ranking of issues and measures, the top- and bottom-scoring responses for each criterion and the overall score were isolated. This highlighted not only the top and bottom scores overall, but also the variances in scoring across the criteria. For example, an issue may have scored high on reach and significance, but low on attention – suggesting that the issue is important but not receiving enough attention. The top and bottom five measures were highlighted for Q1, Q2, and Q4. For Q3, only the top and bottom three measures were highlighted, as there were only eight options. When comparing all measures from Q2-4, the top and bottom ten measures were used.

To help understand the absolute value of the average scores, they were categorised by range:

Score ranges	
Very high	4.5-5
High	4-4.49
Mid-high	3.5-3.99
Mid-low	-3.49
Low	2-2.99

Figure 6: Score ranges (R2)

This made it possible to analyse how scores were distributed within each question set, which was particularly useful when comparing all the potential measures from Q2-4. Analysing these ranges also made it possible to capture higher- and lower-scoring measures (relative to others), even if they were not present in the top or bottom set.

3.2.2 Summary of Responses

Question 1: Ethical and Human Rights Issues

There was an average of 25 responses on each measure in Q1.

Q1 asked respondents to rate a list of ethical and human rights issues in terms of three criteria:

- **Reach** (number of people affected);
- **Significance** (impact on individuals); and
- **Attention** (likely to lead to public debate).

The comprehensive list of 39 issues was taken from Delphi R1 responses, and supplemented with issues identified in other activities of the SHERPA project, including analysis of case studies, stakeholder interviews, and an online survey. Respondents were asked to rate each issue on a scale of 1 to 5 (1 is low, 5 is high). For example, a low score (1) meant the issue affects few (or no) individuals, is trivial, and/or is not of serious concern. A high score (5) meant the issue affects individuals worldwide, has vital consequences, and /or is likely to generate robust public debate. At the end of the set of issues, respondents were given an opportunity to provide a free-text explanation of their ratings; two respondents provided an additional comment.

See Appendix E for a full list of the ethical and human rights issues raised, and the average score of each criterion.

Top Five Results

The top five overall scores were:

- Misuse of Personal Data (4.05)
- Lack of Privacy (3.96)
- Lack of Access to and Freedom of Information (3.85)
- Bias and Discrimination (3.80)
- Impact on Democracy (3.80)

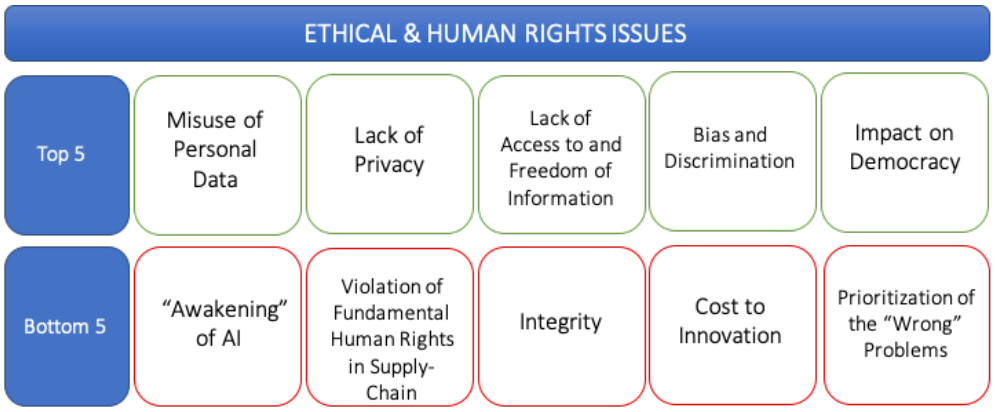

Figure 7: Highest and lowest rated ethical and human rights issues (R2)

There was very little correlation in the scores for ethical and human rights issues across the three criteria. Only one of the top issues was within the top five scores for all three of the individual criteria: *misuse of personal data.* Two issues were in the top five for two of three criteria: *lack of privacy* (22nd in significance) and *lack of access to and freedom of information* (7th in attention). The remaining two issues only scored

within the top five in one criterion: *bias and discrimination* (22nd in reach and 6th attention) and *impact on democracy* (6th in significance and 12th attention).

Other issues that scored in top five of only one criterion (and were not in the top five overall) were:

- Control and Use of Data and Systems (reach)
- Impact on health (significance)
- Potential for Criminal and Malicious Use (significance)
- Disappearance of Jobs (attention)
- Harm to Physical Integrity (attention)
- "Awakening" of AI (attention)

Bottom Five Results

The bottom five overall scores were:

- "Awakening" of AI (3.01)
- Violation of Fundamental Human Rights in Supply-Chain (3.00)
- Integrity (2.99)
- Cost to Innovation (2.89)
- Prioritization of the "Wrong" Problems (2.81)

There was a little more correlation in the bottom scores. Only one issue was within the bottom five scores for all three of the individual criteria: *prioritization of the "wrong" problems.* Three issues were in the bottom five for two of three criteria; two of those scored close to the bottom in all three: *violation of fundamental human rights in supply-chain* (30th in significance) and *cost to innovation* (26th in attention). *"Awakening" of AI* stood out as it was 4th in attention, despite being in the bottom for reach and significance. Lastly, while only falling into the bottom five for attention, *integrity* scored very near the bottom in both other criteria (29nd in reach and 34th in significance).

Other issues that scored in bottom five of the criteria were:

- Accuracy of Non-Individualized Recommendations (significance and attention)
- Potential for Military Use (reach)
- Human contact (significance)
- Lack of Quality Data (attention)

Overall Observations

The majority of the ethical and human rights issues measures (28 of 39) had higher significance scores, followed by reach, then attention scores. Six measures had higher reach scores, followed by attention, then significance scores. The remaining five measures scores fell in different orders.

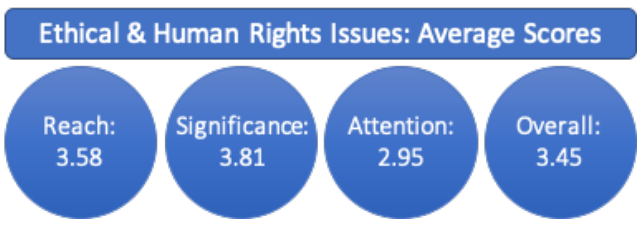

Figure 8: Average scores of the 39 ethical and human rights issues (R2)

For reach, four issues received a score in the high (4-4.49) range, and 22 measures received a score in the mid-high (3.5-3.99) range. No issue scored in the very high range.

For significance, 15 issues received a score in the high range, and 16 issues received a score in the mid-high range. No issue scored in the very high range.

For attention, 1 issue received a score in the high range, and 4 issues received a score in the mid-high range. No issue scored in the very high range. A complete list of the high and mid-high scoring measures for each criterion is in Appendix E.

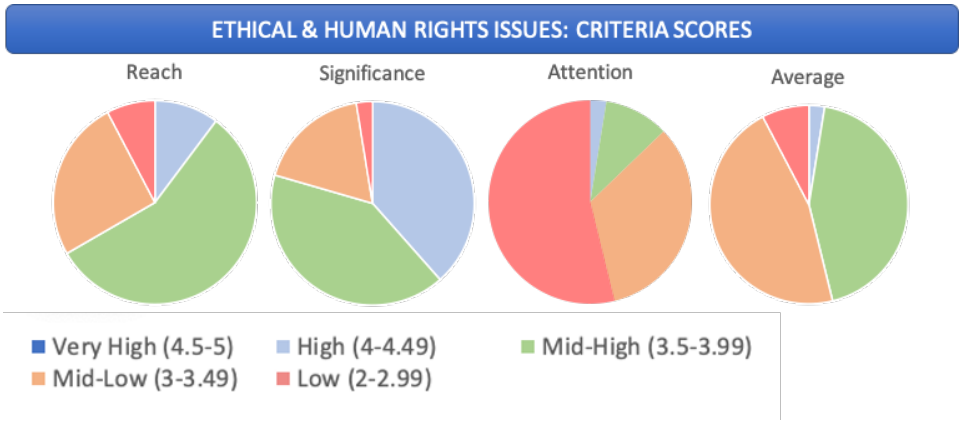

Figure 9: Distribution of scores of ethical and human rights issues (R2)

Overall, 1 issue received a score in the high range, and 17 issues received a score in the mid-high range. No issues scored in the very high range.

High Overall (4-4.49)	Mid-High Overall (3.5-3.99)	
• Misuse of Personal Data	• Lack of Privacy • Lack of Access to and Freedom of Information • Bias and Discrimination • Impact on Democracy • Impact on Health • Control and Use of Data and Systems • Concentration of Economic Power • Lack of Trust • Potential for Criminal and Malicious Use	• Power Asymmetries • Disappearance of Jobs • Violation of End-Users Fundamental Human Rights • Loss of Freedom and Individual Autonomy • Lack of Transparency • Power Relations • Accuracy of Data • Impact on Environment

Figure 10: Ethical and human rights issues scoring within the high and mid-high range (R2)

Question 2: Potential Regulatory Measures

There was an average of 21 responses on each measure in Q2.

Q2 asked respondents to rate a list of potential regulatory measures in terms of three criteria:

- **Desirability** (would you like to have this measure in place?)
- **Feasibility** (in theory, is it possible to have this measure in place?)
- **Probability** (in reality, is it likely that this measure would be put in place?)

The list of 18 potential regulatory measures originated from the Delphi R1 responses, and were refined and supplemented by analysis conducted in other deliverables of the SHERPA project, including D3.3. Report on Regulatory Options.[25]

Respondents were asked to rate each issue on a scale of 1 to 5 (1 is low, 5 is high). For example, a low score (1) means the measure will have a major negative effect, is very challenging to create, and/or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and/or is very likely to happen. At the end of the set of potential measures, respondents were given an opportunity to provide a free-text explanation of their ratings; however, there were no additional responses.

See Appendix E for a full list of the potential regulatory measures and the average score of each criterion.

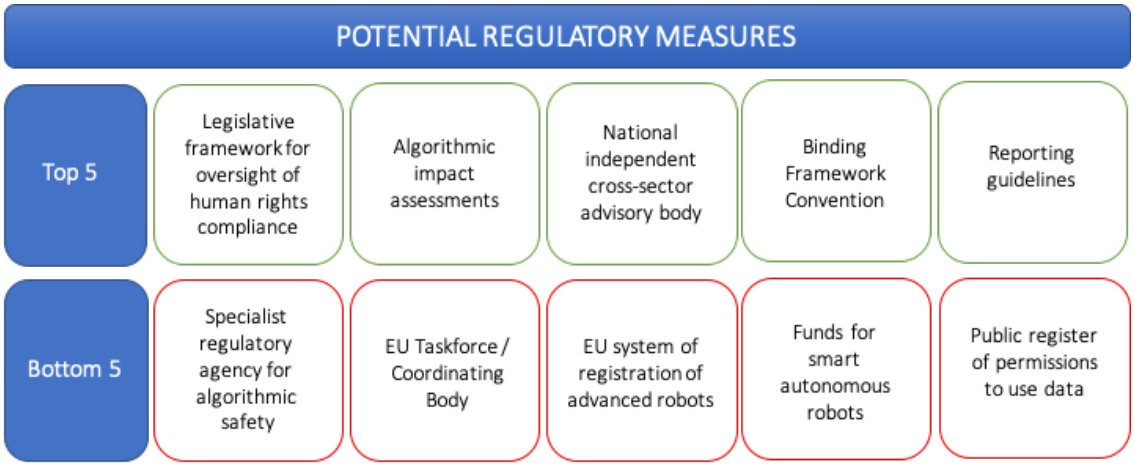

Figure 11: Highest and lowest rated potential regulatory measures (R2)

Top Five Results

The top five overall scores were:

- Legislative framework for independent and effective oversight of human rights compliance (3.70)
- Algorithmic impact assessments under the General Data Protection Regulation (3.65)
- New national independent cross-sector advisory body (3.59)
- Binding Framework Convention (3.51)
- Reporting Guidelines (3.50)

Two of the top regulatory measures were within the top five scores for all three of the individual criteria: *legislative framework for oversight of human rights compliance*, and *algorithmic impact assessments*.

The remaining three top measures displayed some divergences in scoring across the individual criteria. Creation of a *national independent cross-sector advisory body* received the highest relative score on both feasibility and probability, but was 13th on desirability. *Reporting guidelines* was also within the top five on feasibility and probability, and 8th on desirability. The sixth-highest ranked measure overall, *CEPEJ European Ethical Charter*, was also within the top five feasibility and probability (and received a high score for desirability).

[25] Rodrigues, Rowena; Laulhe Shaelou, Stephanie; Lundgren, Björn (2020): D3.3 Report on regulatory options. De Montfort University. Online resource. https://doi.org/10.21253/DMU.11618211.v4.

Conversely, creation of a *binding Framework Convention* was in the top five desirable options, but 7th in feasibility and probability. This illustrates how feasibility and probability were closely aligned; the same five measures received the highest scores, in nearly the same order, and three of those measures were in the top five overall. In contrast, the most desirable - *better enforcement of existing human rights law* - was in mid-range of feasibility (8th) and probability (9th). Creation of a *treaty for AI and Big Data* was also desirable, but less feasible (15th) or probable (16th).

Bottom Five Results

The bottom five overall were:

- Specialist regulatory agency to regulate algorithmic safety (3.07)
- EU Taskforce/Coordinating (3.06)
- EU system of registration of advanced robots (2.85)
- Funds for all smart autonomous robots (2.75)
- Public Register of Permission to Use Data (2.71)

There was even more consensus in the bottom five measures overall. Three of the bottom regulatory measures were within the bottom five scores for each of the individual criteria: *EU system of registration of advanced robots*; *funds for smart autonomous robots;* and *public register of permission to use data*. Creation of a *specialist regulatory agency for algorithmic safety* was not desirable or probable, and creation of *EU Taskforce/Coordinating Body* was not desirable or feasible.

Overall Observations

The majority of the potential regulatory measures (16 of 18) were rated more desirable than feasible or probable.

Figure 12: Average scores of the 18 potential regulatory options (R2)

However, on the whole, potential regulatory measures did not receive high ratings. No measure received a score in the very high (4.5-5) or high (4-4.49) range for feasibility or probability. Only four measures scored in the high range for desirability:

- Better enforcement of existing international human rights law
- Legislative framework for oversight of human rights compliance
- Algorithmic impact assessments
- Binding Framework Convention

Apart from the top-rated options, only one other measure rated in the mid-high range (3.5-3.99) by 2 of 3 criteria: *regulatory sandboxes for AI and big data* (in desirability and feasibility). Other measures (in addition to those discussed above) that scored in the mid-high range of desirability, but lower in feasibility and probability, were:

- Register of algorithms used in government
- Three-level obligatory impact assessments for new technologies
- Redress-by-design mechanisms for AI
- New laws regulating specific aspects

A complete list of the high and mid-high scoring potential regulatory measures for each criterion is in Appendix E.

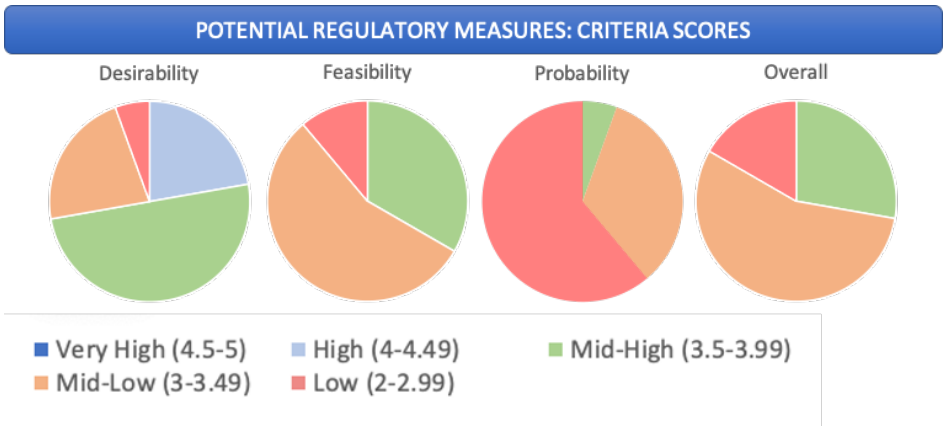

Figure 13: Distribution of scores from 'very high' to 'low' for potential regulatory measures (R2)

For the overall scores, no measures were scored overall in the very high or high range. Five measures were rated in the mid-high range (same as top five options). Ten measures were in the mid-low range (3-3.49), and three measures in the low range (2-2.99).

Question 3: Potential Technical Measures

There was an average of 20 responses on each measure in Q3.

Question 3 asked respondents to rate a list of potential technical measures in terms of three criteria (the same used in Q2):

- **Desirability** (would you like to have this measure in place?)
- **Feasibility** (in theory, is it possible to have this measure in place?)
- **Probability** (in reality, is it likely that this measure would be put in place?)

The list of 8 potential technical measures originated from the Delphi R1 responses, and were refined and supplemented by analysis conducted in other deliverables of the SHERPA project, including D3.5 Technical options and interventions interim report[26].

Respondents were asked to rate each issue on a scale of 1 to 5 (1 is low, 5 is high). For example, a low score (1) means the measure will have a major negative effect, is very challenging to create, and/or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and/or is very likely to happen. At the end of the set of potential measures, respondents were given an opportunity to provide a free-text explanation of their ratings; three respondents provided an additional comment.

See Appendix E for a full list of the potential technical measures and the average score of each criterion.

[26] Kirichenko, Alexey; Marchal, Samuel (2020): D3.5 Technical Options and Interventions (Interim report). De Montfort University. Online resource. https://doi.org/10.21253/DMU.12973031.v1.

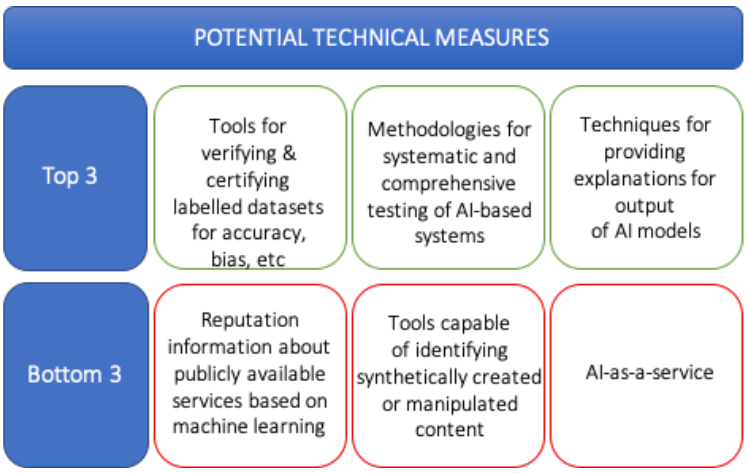

Figure 14: Highest and lowest rated potential technical measures (R2)

Top Five Results

The top three overall scores were:

- Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties (3.95)
- Methodologies for systematic and comprehensive testing of AI-based systems (3.90)
- Techniques for providing explanations for output of AI models (3.87)

Only one of the top technical measures was within the top five scores for all three of the individual criteria: *methodologies for systematic and comprehensive testing of AI-based systems*.

The other measures displayed some divergences in scoring across the individual criteria. *Tools for verifying and certifying labelled datasets* scored highest on feasibility and probability, and 5th on desirability. *Techniques for providing explanations for output of AI models* scored high on desirability and probability, but 5th on feasibility. Unlike Q2, there was not a close correlation between scores for feasibility and probability.

Two other measures scored highly on only one of the three criteria. *Tools capable of identifying synthetically created or manipulated content* was scored as desirable, but less feasible (8th) or probable (7th). *Reputation information about publicly available services based on machine learning models* was scored as feasible, but less desirable (6th) or probable (8th).

Bottom Five Results

The bottom three overall scores were:

- Reputation information about publicly available services based on machine learning models (3.63)
- Tools capable of identifying synthetically created or manipulated content (3.58)
- AI-as-a-service (3.52)

None of the technical measures were within the bottom three scores for all three of the individual criteria. However, three measures were ranked low in 2 of 3 criteria and overall. *AI-as-a-service* was rated less desirable (8th) or feasible (7th). As noted above, *tools capable of identifying synthetically created or manipulated content* was scored as desirable, but less feasible (8th) or probable (7th), and *reputation information about publicly available services based on machine learning models* was scored as feasible, but less desirable (6th) or probable (8th).

Overall Observations

All potential technical measures were rated more desirable, then feasible, then probable.

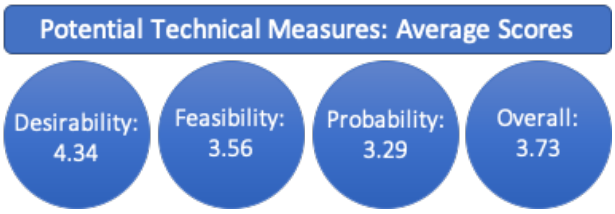

Figure 15: Average scores of the 8 potential technical options (R2)

Overall, potential technical measures received relatively high scores, due to the high scores for desirability; four measures were very high (4.5-5), and three were high (4-4.49) for desirability. The only measure to score lowest in the mid-high (3.5-3.99) range for desirability was *AI-as-a-service*. However, no measure received a score in the very high or high range for feasibility or probability, and no measure scored in the mid-high range for probability. Other measures (in addition to those in the top three overall) that scored in the mid-high range of feasibility were:

- Reputation information about publicly available services based on machine learning models
- Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models
- Techniques for providing explanations for output of AI models
- Tools for verifying and certifying publicly available services based on machine learning models.

A complete list of the high and mid-high scoring measures for each criterion is in Appendix E. For the overall scores, all measures scored in the in the mid-high range (3.5-3.99).

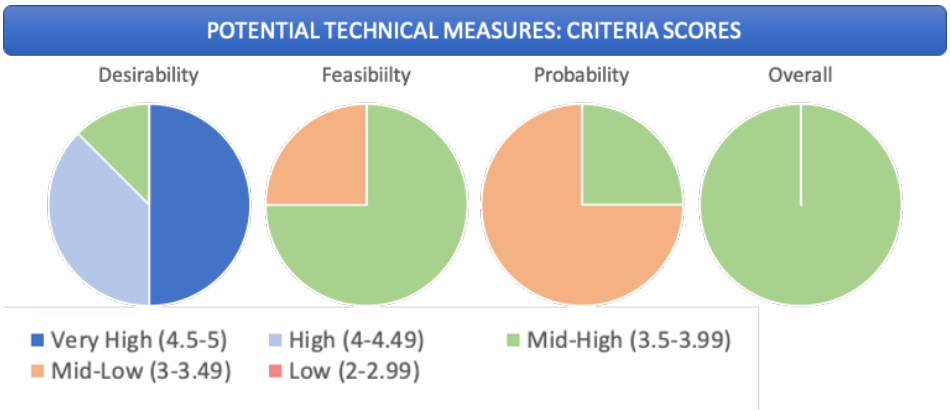

Figure 16: Distribution of scores from 'very high' to 'low' for potential technical measures (R2)

Question 4: Other Potential Measures

There was an average of 19 responses on each measure in Q4.

Question 4 asked respondents to rate a list of other potential measures in terms of three criteria (the same used in Q2 and Q3):

- **Desirability** (would you like to have this measure in place?)
- **Feasibility** (in theory, is it possible to have this measure in place?)
- **Probability** (in reality, is it likely that this measure would be put in place?)

The list of 26 other potential measures originated from the Delphi R1 responses, and were refined and supplemented by analysis conducted in other deliverables of the SHERPA project.

Respondents were asked to rate each issue on a scale of 1 to 5 (1 is low, 5 is high). For example, a low score (1) means the measure will have a major negative effect, is very challenging to create, and/or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and/or is very likely to happen. At the end of the set of other potential measures, respondents were given an opportunity to provide a free-text explanation of their ratings; two respondents provided an additional comment.

See Appendix E for a full list of the other potential measures and the average score of each criterion.

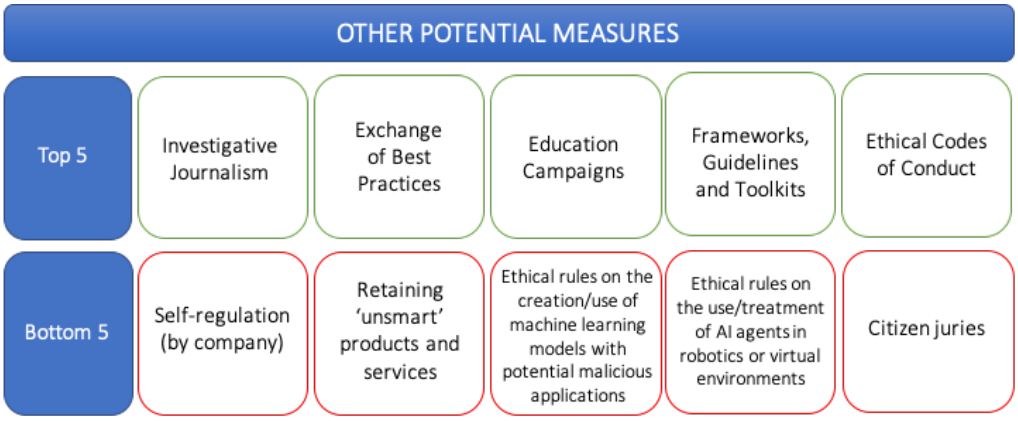

Figure 17: Highest and lowest rated other potential measures (R2)

Top Five Results

The top five overall scores were:

- Investigative Journalism about issues concerning SIS (4.48)
- Exchange of Best Practices (4.43)
- Education Campaigns (4.19)
- Framework, Guidelines, and Toolkits for project management and development (4.14)
- Ethical Codes of Conduct (4.11)

Two of the top other potential measures were within the top five scores for all three of the individual criteria: *investigative journalism*, and *exchange of best practices*.

Otherwise, there was not a discernible alignment between the scores across the individual criteria. The remaining three top measures scored within the top five in only one category. *Education campaigns* was in the top five for feasibility, and only marginally lower in desirability (6th) and probability (9th). *Frameworks, guidelines, and toolkits* was in the top five for probability, but lower on desirability (11th)

and feasibility (7th). *Ethical Codes of Conduct* displayed the most divergence; while in the top five for probability and 6th for feasibility, it was almost in the bottom five for desirability (22nd).

Two other measures scored in the top five in two individual criteria. *High-Level Expert Groups* were in the top five for feasibility and probability, but in the bottom five for desirability. *Grievance mechanisms for complaints on SIS* was in the top five for desirability and feasibility, but 18th for probability.

Two measures were only in the top five for desirability. *More open source tools that allow for transparency, explainability, and bias mitigation* was 16th for feasibility and 8th for probability. *Public "whistleblowing" mechanisms for the reporting of bias, inaccuracies, or ethical impacts of systems based on machine learning models* was 19th for feasibility and 21st for probability.

Bottom Five Results

The bottom five overall scores were:

- Self-Regulation by Company (3.58)
- Retaining 'Unsmart' Products and Services (3.54)
- Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases (3.40)
- Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments (3.22)
- Citizen Juries (2.82)

Two of the bottom other potential measures were within the bottom five scores for each of the individual criteria: *ethical rules pertaining to the use or treatment of AI agents in robotics or virtual environments* and *citizen juries*.

Some measures scored low on one criterion, but higher on the others. *Open Letters* was in the bottom five for desirability but was 10th for feasibility and probability. Conversely, *rules on how decisions in systems that have the capability to cause physical harm* should be made in difficult situations was in the bottom five for feasibility and 16th for probability, but was 7th for desirability. *Ethical Mindset adopted by companies* was in the bottom five for probability and 21th for feasibility, but was 8th for desirability.

The remaining bottom scoring measures also scored relatively low across the criteria; none had a score putting them higher than 15th place in any category. In addition to the five bottom-ranking overall, other measures scoring within the bottom five for desirability, feasibility and/or probability were:

- Self-Regulation by Company
- Retaining 'Unsmart' Products and Services not probable
- Ethical Codes of Conduct
- High-Level Expert Groups
- Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases.

Overall Observations

The majority of measures (23 or 26) rated more desirable than feasible or probable. *Ethical Codes* and *High-Level Expert Groups* were more probable and feasible than desirable; *Open Letters* were more feasible than desirable or probable.

Figure 18: Average scores of the 26 other potential measures (R2)

On the whole, 'other' potential measures received high scores. For desirability, five measures were in the very high (4.5-5) range, thirteen in the high (4-4.49) range, and seven were in the mid-high (3.5-3.99) range. The only measure to score lowest in the mid-low range for desirability was *Citizen Juries*. For feasibility, one measure was in the very high range, ten in the high range, and eleven were in the mid-high range. For probability, five measures were in the very high range, six in the high range, and ten were in the mid-high range.

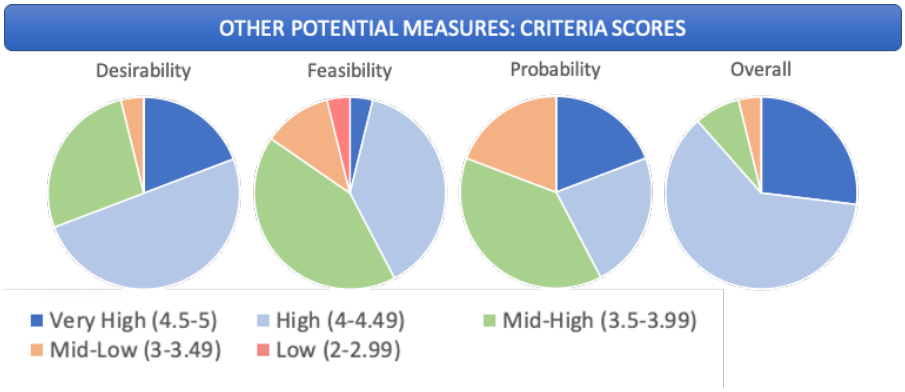

Figure 19: Distribution of scores from 'very high' to 'low' for other potential measures (R2)

For the overall scores, seven measures were in the very high (4.5-5) range, sixteen in the high (4-4.5) range, and two were in the mid-high range. The only measure to score lowest in the mid-low range overall was *Citizen Juries*.

High Overall (4-4.5)	Mid-High Overall (3.5-4)	
• Investigative journalism • Exchange of best practices • Education campaigns • Framework, guidelines, and toolkits • Ethical codes of conduct • High-level expert groups • More open source tools	• Grievance Mechanisms for complaints on SIS • NGO Coalitions on particular issues • Public Policy Commitment by company to be ethical • International Ethical Framework • Open Letters to governments and the public • Stakeholder Dialogue and Scrutiny • Third-party Testing and External Audits • Individual action • Public "Whistleblowing" Mechanisms	• Standardisation • Certification • Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations • 'Fairness' Officer or Ethics Board employed within companies using/developing SIS • Ethical Mindset adopted by companies • Self-Regulation by company • Retaining 'Unsmart' Products and Services by keeping them available to purchase and use

Figure 20: Other potential measures scoring within the high and mid-high range (R2)

3.2.3 Key Findings

Ethical and Human Rights Issues

In R1, the most prominent issues were a *lack of transparency*, *lack of human decision-making*, *lack of privacy*, and *bias and discrimination*. In R2, *lack of privacy* and *bias and discrimination* continue to be issues of high concern, along with *misuse of personal data*, *lack of access to (and limitations on) freedom of information*, and *impacts on democracy*. Other key concerns included *violation of human rights for end-users*, *loss of freedom and individual autonomy*, *impacts on power relations* (political and economic), *lack of transparency and trust*, *potential for criminal and malicious use*, and *disappearance of jobs*. While not cited in R1, the *environmental impact of SIS* was among the key concerns in R2.

Lack of human decision-making and *human contact* were not key concerns, scoring mid-low to low across the criteria. As in R1, *harm to physical integrity* was also a low-ranking concern; it scored much lower in reach and significance than attention. '*Awakening' of AI* also scored high in attention but was the lowest in both reach and significance. Two issues that scored lower than anticipated were *unintended, unforeseeable adverse impacts*, and *lack of accountability and liability*.

Potential Governance Measures

The governance of SIS requires a smart mix of instruments that will address ethical and human rights concerns. To better understand how the different types of measures (regulatory, technical, and other) should be prioritized, all 52 potential governance measures were compared. **Regulatory measures** were the lowest scoring in all criteria and overall, with no regulatory measures making into the top fifteen measures overall. More promising were **technical measures**, which were the most desirable on average. However, only three technical measures were in the top fifteen measures overall. Most promising were **other measures**, which scored highest on feasibility, probability and overall, and which were twelve of the top fifteen measures overall.

	Average Scores of Potential Measures			
	Desirability	Feasibility	Probability	Overall
Highest	Technical Measures (4.34)	Other Measures (3.85)	Other Measures (3.49)	Other Measures (3.82)
	Other Measures (4.12)	Technical Measures (3.56)	Technical Measures (3.29)	Technical Measures (3.73)
Lowest	Regulatory Measures (3.70)	Regulatory Measures (3.31)	Regulatory Measures (2.79)	Regulatory Measures (3.26)

Figure 21: Average scores of all potential measures (R2)

	Number of Potential Ratings Scored from 'Very High' to 'Low'								
	Desirability			Feasibility			Probability		
	Regulatory	Technical	Other	Regulatory	Technical	Other	Regulatory	Technical	Other
Very High	-	4	5	-	-	1	-	-	5
High	4	3	13	-	-	10	-	-	6
Mid-High	9	1	7	6	6	11	1	2	10
	72%	100%	96%	33.3%	75%	85%	5%	25%	81%
Mid-Low	4	-	1	10	2	3	6	6	5
Low	1	-	-	2	-	1	11	-	-
	28%	-	4%	66.6%	25%	15%	95%	75%	19%

Figure 22: Potential measures from 'very high' to 'low' (R2)

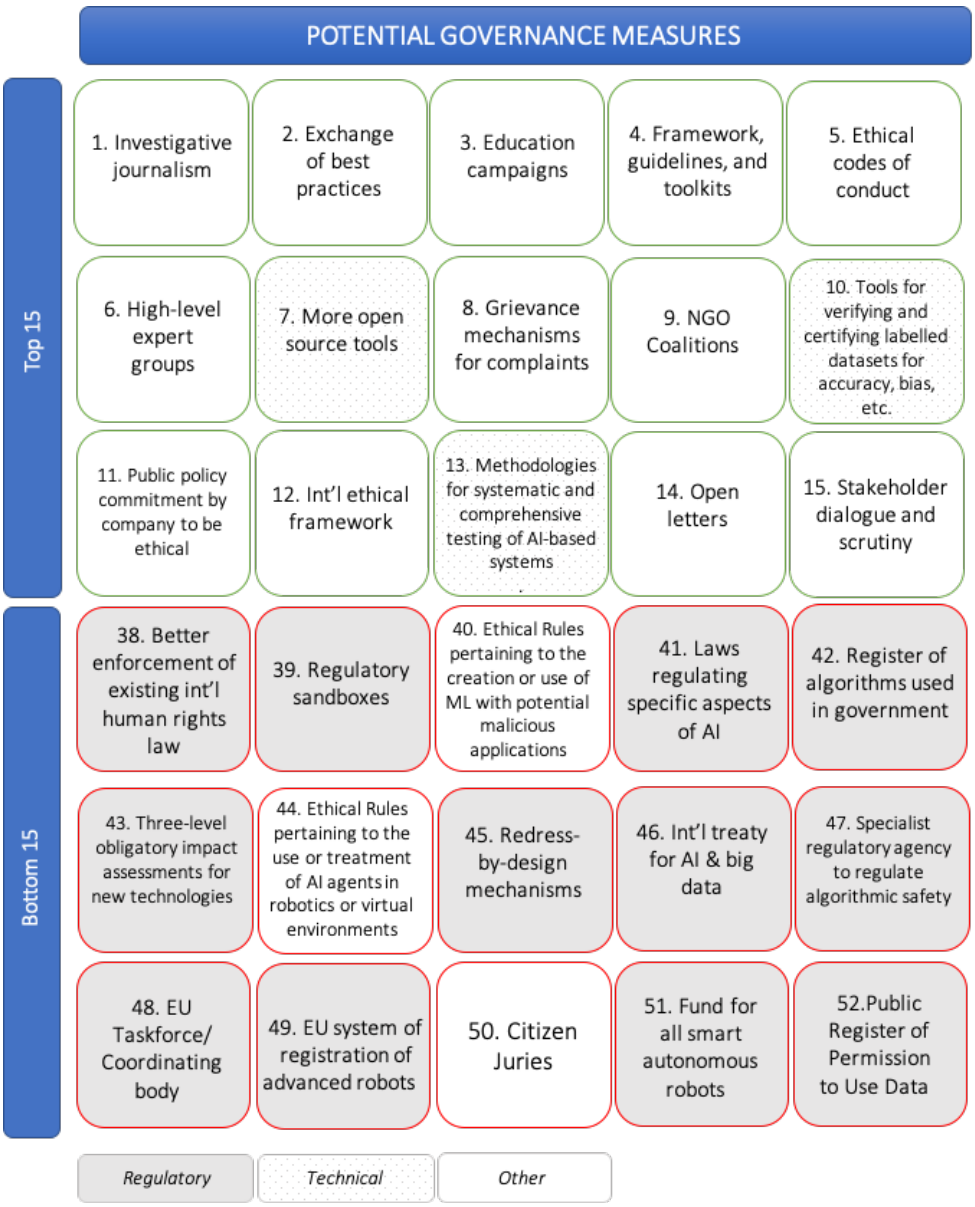

Figure 23: Top and bottom fifteen potential governance overall (R2)

Regulatory Measures

In R1, regulation was the most frequently cited example of a possible 'approach, method, or tool' to address the ethical and human rights concerns associated with SIS.

However, in R2, most potential regulatory measures scored low, both in absolute terms and relative to other types of potential measures. No regulatory measure was in the top fifteen measures, and twelve regulatory measures were in the bottom fifteen potential measures. This was because potential regulatory measures received the lowest average scores in all three criteria and overall. For the overall scores, no regulatory measures scored in the very high (4.5-5) or high (4-4.49) range. All of the top five regulatory measures (below) scored in the mid-high (3.5-3.99) range, which was lower than the top scoring technical and other measures. More significantly, potential regulatory measures had the highest percentage of measures scoring in the mid-low (3-3.49) to low (2-2.99) range for all three criteria. This was particularly true of probability, where 95% of measures scored low. Within potential regulatory measures, the majority (16 of 18) were rated more desirable than feasible or probable.

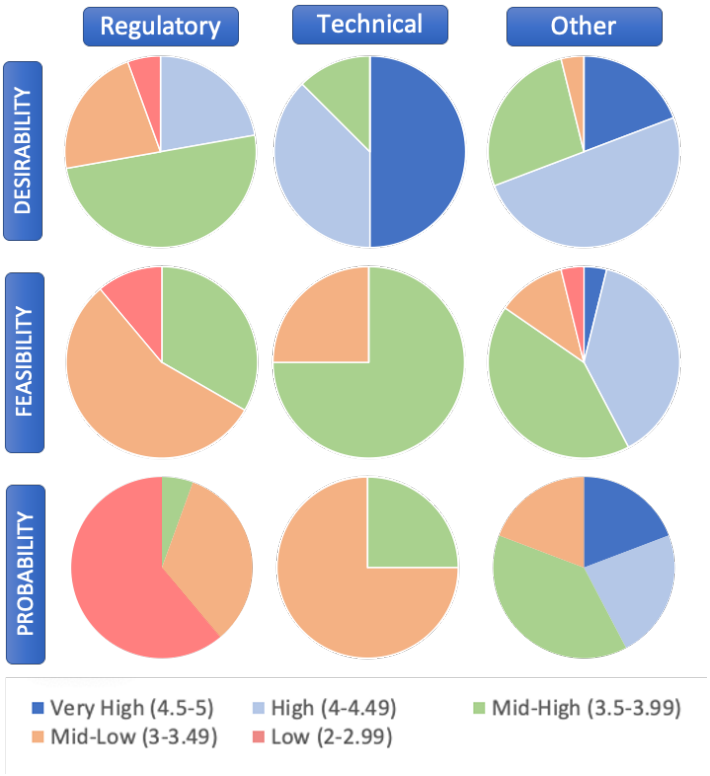

Figure 24: Distribution of scores from 'very high' to 'low' for each category of measures (R2)

Technical Measures

In R1, technical measures were rarely mentioned. However, in R2, technical measures scored relatively high, particularly in regard to desirability; all technical measures were very high (4.5-5) or high (4-4.49) for desirability. However, with lower average scores in feasibility and probability, only three technical measures were in the top fifteen measures. For the overall scores, all technical measures scored in the mid-high range (3.5-3.99). All potential technical measures were rated more desirable, then feasible, then probable.

Other Measures

In R1, respondents cited a broad range of other measures. In R2, these other potential measures scored high, both in absolute terms and relative to the other two categories of measures. Twelve of the top fifteen measures were other measures. This was because other measures received the highest average scores in feasibility, probability, and overall. For the overall scores, seven measures were in the very high (4.5-5) range, sixteen in the high (4-4.49) range, and two in the mid-high (3.5-3.99) range. The only measure to score in the mid-low (3-3.49) range overall was *Citizen Juries*. The majority of measures (23 of 26) scored more desirable than feasible or probable.

3.3 Round 3

The third, final round of the Delphi asked respondents to select, from the list of fifteen highest scoring measures in R2, the three most important potential governance measures for immediate action. For each selection, respondents were prompted to explain, (a) why the measure is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure. The respondents did not rank their selection, therefore the order of selections was not relevant. Respondents were also given the option to also identify any potential governance measures that should not be prioritised, as well as any additional comments.

The fifteen top scoring measures from R2, in no particular order, were:

- Investigative journalism about issues concerning SIS
- Exchange of best practices
- Education campaigns
- Framework, guidelines, and toolkits for project management and development
- Ethical codes of conduct
- High-level expert groups
- More open source tools that allow for transparency, explainability, and bias mitigation
- Grievance mechanisms for complaints on SIS
- NGO coalitions on particular issues
- Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties
- Public policy commitment by company to be ethical
- International ethical framework
- Methodologies for systematic and comprehensive testing of AI-based systems
- Open letters to governments and the public
- Stakeholder dialogue and scrutiny.

3.3.1 Analysis Method

As respondents were not asked to rank their selections, and each individual selection had equal weight, all 117 discrete selections were pooled. From the complete list, the number of selections for each potential governance option was tallied. Responses to the follow-up and additional questions were grouped together by question; all responses were taken into consideration, including potential duplicate responses (where written text was identical or nearly identical).

Analysis of the grouped selection by each respondent was not conducted (i.e. which three choices were selected together), therefore no conclusions were drawn about how individual respondents would group the three most immediate measures for prioritisation.

3.3.2 Summary of Responses

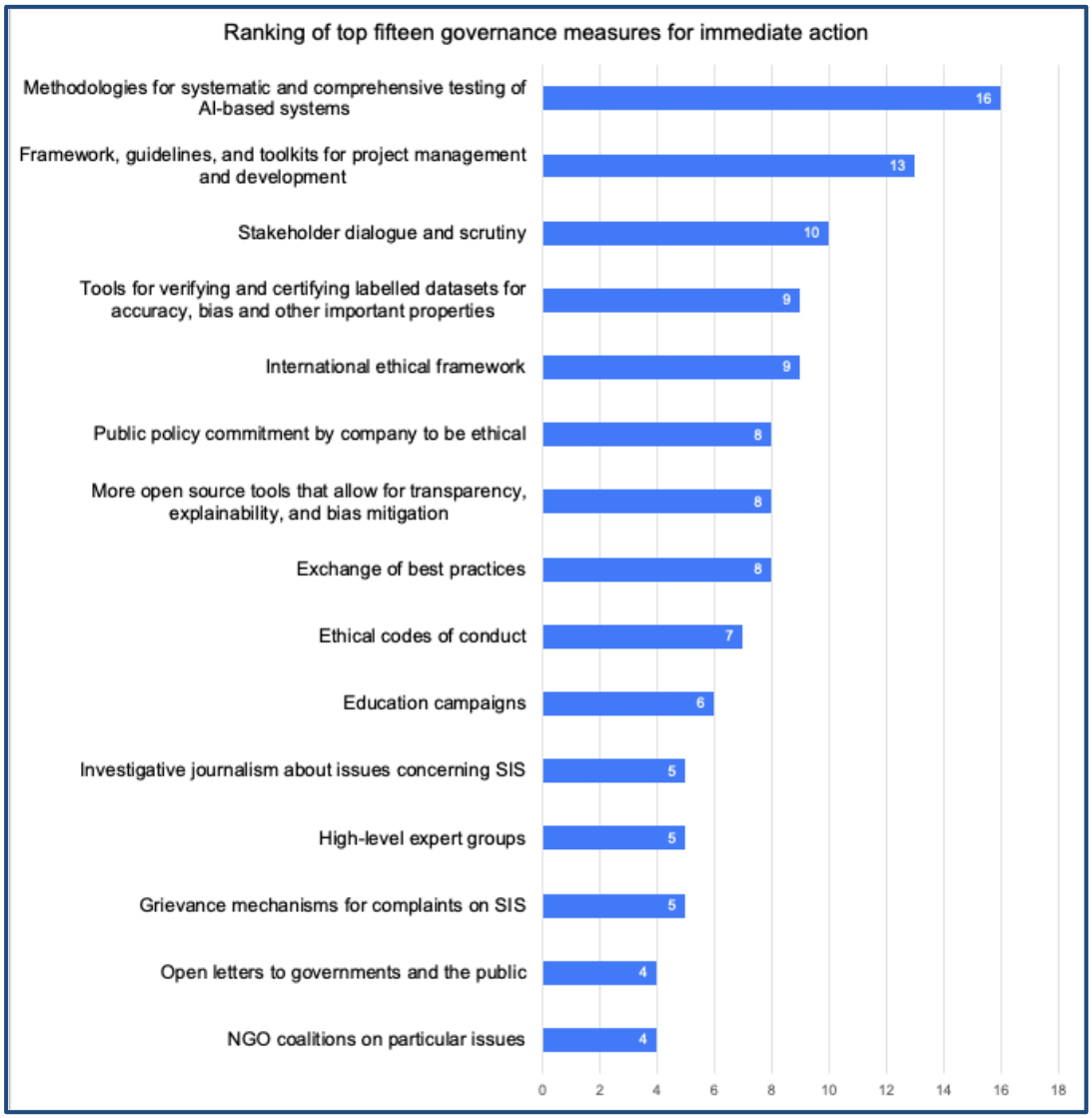

Figure 25: Ranking of op 15 measures for immediate action (R3)

Investigative journalism (2 responses)
Why is it important? • To raise public awareness and understanding of urgent issues (e.g. privacy and personal data protection issues) • Shift attention away from theoretical concerns (e.g. 'Awakening' of AI)
Who should implement it? And how? • **Journalists, bloggers and authors**
Indicators to measure success? • Number of published pieces • Attention received online (e.g. number of views, shares, likes)

Exchange of best practices (3 responses)
Why is it important?

- Need exchange of best practices to have development/evolution
- Understanding ethics and human rights in context of AI must be based on broad exchange of experiences with technology and the impact on humans
- Important to build collective knowledge in moving landscape

Who should implement it? And how?

- **Industry**
- **All actors engaged** (from users to developers)
- Should not be done by governments

Indicators to measure success?

- Well-being of humans affected by AI & Big Data (could be measured through surveys)
- Increased awareness
- Implementation of policies

Education campaigns (3 responses)

Why is it important?

- Helps public be aware of opportunities and threats
- Will stimulate search for knowledge
- Need to make knowledge more accessible to public (from early age and for public at large)

Who should implement it? And how?

- **Governments**: Propose as part of compulsory curricula; create general public awareness campaigns
- **Universities**: Teach at bachelor and master level
- **Specialists**: Develop easily digestible instructional videos for dissemination online (e.g. Youtube, Linkedin and Instagram)
- (unspecified): Create toolkit for developing and conducting education campaigns

Indicators to measure success?

- Number of people who are aware of new technologies and understand what new technology bring
- Increase in knowledge and skills; at university, can be measured with exams that test primary and secondary knowledge gained. Secondary knowledge fields may include: knowledge about programming language; knowledge about systems from the past and what impact that has had on the systems now; and, Culture differences locally and internationally

Framework, guidelines, and toolkits for project management and development (7 responses)

Why is it important?

- Tools for ethical consideration not currently part of normal project management
- Ethics is often only perceived in narrow sense of 'research ethics'
- One of the most effective ways to influence the development of SIS
- Is the necessary "operationalizing" of ethical codes
- Should contribute to development of principles in the technology design
- Need coordinated framework to avoid development of incompatible approaches
- Must have understanding, learning, and sharing
- Need to translate abstract norms into concrete proposals

Who should implement it? And how?

- **UN organisations**: Publication of a code (publicly-available)
- **National governments** (e.g. data authorities, PPP, standardization authorities): Develop standardised norms and practices
- **Companies**: Create toolkits that allow for the implementation and evaluation of ethical compliance throughout different phases of product development (need tools during conceptual phase, when realising and evaluating proofs of concept, and when evaluating production versions); Create templates for evaluation of AI products in relation to ethical dimensions; Embed compliance with ethical standards throughout the development and life cycle of AI related products
- **Stakeholders**
- **'Users of the systems'**
- **EU projects (& partners)**

Indicators to measure success?

- Use of frameworks
- Reported use of tools in companies' annual reporting
- Widespread creation and adoption of corporate guidance
- Adoption of necessary regulation by governments
- Use of tools in project managements (i.e. reflection after project completion)
- Number of countries signing and applying the code
- Widespread use and adoption of commonly accepted ethical developments process templates (practices) that include the tools to evaluate medical compliance within the different phases of development and life cycle of AI products
- Adoption of certification based on standards and practises

Ethical codes of conduct (2 responses)
Why is it important? • There is risk of wrong utilization • All professions should have ethical codes
Who should implement it? And how? • **UN organisations**: Produce a code, signed by members • **Governments** • **Professional bodies**
Indicators to measure success? • Number of complaints • Percentage of non-compliant publications • Changed ethical views

High-level expert groups (no responses)
Why is it important? • No response
Who should implement it? And how? • No response
Indicators to measure success? • No response

More open source tools that allow for transparency, explainability, and bias mitigation (5 responses)
Why is it important? • Vital for perception and reality of 'fair AI' and to empower individuals / organisations to access rights • Transparency and explainability are essential features of AI solutions • Open source tools are sustainable • Could help both in solving the technical aspects of bias/transparency/explainability handling and raise developer awareness of these issues • Programmers must aware that everything they develop is transparent and users or ethical groups should be able to check where the data comes from, what is the algorithm, etc. to ensure that bias are avoided or inform users about possible bias
Who should implement it? And how? • **Open source repositories**: Check transparency of code • **Global coalition of experts**: Reach consensus • **Global companies**: Adopt through soft law measures; publish their own frameworks • **Governments**: Fund R&D of prototypes of frameworks • (unspecified): Make Open Data Directive a global standard (i.e. data produced by public authorities and by the people should be open) • **Developers community**
Indicators to measure success? • Number of open data sets • Good practice of tools that allow transparency, explainability, etc. • Adoption and popularity of the developed frameworks (e.g., number of installs & dependent packages) • Analysis of outcomes

Grievance mechanisms for complaints on SIS (2 responses)
Why is it important? • Could increase transparency (if rights of users regarding automatic decision-making under GDPR are developed further) • Is an important fall-back mechanism that could make use of existing voice mechanisms such as consumer protection
Who should implement it? And how? • **EU**: Create/expand legal liability within EU consumer protection law through application of a precautionary principle (already applied in environmental law) • **EU & national governments**: Enshrine in administrative law and redress processes; Create and strengthen redress mechanisms (at national level, could be mandated to consumer or telecom authorities); Fund development of technical prototypes / model mechanisms • **Companies**: Develop technical prototypes / model mechanisms • (unspecified): Preclude or limit use of private arbitration by Big Tech companies
Indicators to measure success? • Prevalence and quality of working grievance mechanisms in popular online services • Amount of complaints resolved

NGO coalition on particular issues (no responses)
Why is it important?

• No response
Who should implement it? And how?
• No response
Indicators to measure success?
• No response

Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties (4 responses)
Why is it important? • Enables algorithmic training on appropriate data sets • Reduces possibility of tainted or biased data being utilized • Enables compliance assessment with governance guidance and regulation • Mitigate risks and unwanted consequences (when data is of fine quality) • Enable stakeholders (developers, project managers, users, etc.) to check that code and data used are safe / ethical
Who should implement it? And how? • **Governments**: Adopt tools • **International experts**: Develop tools by consensus • **Software development companies**: Develop labels and seek certification • **Standardisation bodies**: Develop labels and certify • **Public-private partnerships** • **ICO/CDEI/ATI** • **Global consultancies** (e.g. Deloitte, EY, KPMG) • **Global law firms** • **Blue chip companies** • **Influential think tanks** (e.g. Ada Lovelace, Nesta)
Indicators to measure success? • Independent checking and monitoring of usage of data sets and outcomes • Development and widespread introduction of these tools • Data set quality assessment scheme and certification • Increase in applications certified

Public policy commitment by company to be ethical (1 response)
Why is it important? • Is a strong message • Ensures compliance
Who should implement it? And how? • **Companies**: Include statement in annual report
Indicators to measure success? • Controls of the annual reports and percentage of deviations

International ethical framework (5 response)
Why is it important? • Companies need a clear standard to test compliance

- The digital economy is international and requires an international ethical basis to operate to best effect
- International standard has more effect than local measures (e.g. effectiveness of GDPR because European legislation, not national)
- Needs to be aligned internationally; not a unilateral issue, because everything will be driven by the use of totalitarian regimes.

Who should implement it? And how?

- **EU**: adopt 'ruling'
- **Democratic states**: find their own way to reply to protect society
- [unspecified]: Implement framework internationally, with effects felt by European entrepreneurs and third parties from other countries; violation of the framework must lead to a public debate among residents and politicians
- [unspecified]: Need to specify (1) overarching principles, (2) factors that should be taken into account in risk assessment, (3) governance and/or regulation that is appropriate for different levels of risk, and (4) tools that should be considered in assessing compliance with ethical codes, governance guidelines and regulation

Indicators to measure success?

- Cannot be measured (is a precondition)
- World peace
- Confrontation (within the framework) of entrepreneurs by residents and accountability
- Take-up by transnational body (e.g. UN, OECD, Council of Europe, G20)

Methodologies for systematic and comprehensive testing of AI-based systems (4 responses)

Why is it important?

- Should be common best practice
- Once a broad set of tested methodologies are available and companies have more experience, the costs of ethical compliance will come down and it will be much easier to require ethical compliance
- Crucial that methods and tools are available to ensure that this is done properly, given the complexity and unfamiliarity of the subject
- Enables transparency and explainability and public trust in systems
- Necessary to make it possible to translate the ethical framework

Who should implement it? And how?

- **Group of global experts**: Reach consensus; Draw up a user-friendly, flexible method that can be applied in custom environment (this is fastest and most efficient way)
- **Governments**: Adopt method
- **EU**: Fund academic institutions to develop and test methodologies
- **Academia**: Integrate method; Provide oversight
- **Press**: Provide oversight
- **Industry**
- **AI coalitions** (including PPPs)

Indicators to measure success?

- Broad adoption of standardised and academically tested methodology when developing AI products (i.e. comprehensive testing is common practice)
- Testing is seen as an important standard in business processes (like the "9-plane" model of Prof. Rick Maes, which is always referred to when a standard setup is required for an information management component in an organization in the IM domain)
- Number of applications

• Reports that methodologies have been used (included in annual reporting of companies and organisations)

Open letters to governments and the public (no response)
Why is it important? • No response
Who should implement it? And how? • No response
Indicators to measure success? • No response

Stakeholder dialogue and scrutiny (1 response)
Why is it important? • Essential to exchange of best practices • Must have regular and systemic dialogue to achieve successful outcome
Who should implement it? And how? • **All actors engaged**
Indicators to measure success? • Set of guidelines and best practices

3.3.3 Key Findings

The top three potential governance measures with the most selections by respondents were:

- *Methodologies for systematic and comprehensive testing of AI-based systems* was selected 16 times (37% of respondents).
- *Framework, guidelines, and toolkits for project management and development* was selected 13 times (30% of respondents)
- *Stakeholder dialogue and scrutiny* was selected 10 times (23% of respondents).

Closely behind, sharing the 4th and 5th slots, were *Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties*, and *International ethical framework*, with 9 selections each (20.9% of respondents).

The popularity of potential governance measures in R3 did not closely reflect the order of how the fifteen measures scored in R2. For example, *investigative journalism* was the top scoring option in R2, but was tied for the 11th-13th places in R3, with only 5 selections.

By comparison, *Stakeholder dialogue and scrutiny* was the lowest scoring of the top 15 in R2, but was in 3rd place in R3.

However, one discernible trend was that the three technical measures all ranked in R3 on par or higher than the ranking in R2: *Methodologies for systematic and comprehensive testing of AI-based systems* went from #13 to #1, *Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties* from #10 to #4, and *More open source tools that allow for transparency, explainability, and bias mitigation* stayed at #7.

The level of detail in the answers to the supplemental explanatory questions was varied. This is partly because the average number of respondents who answered the question for a particular measure was only 2.6. Therefore, it is difficult to generalize any consensus opinion from the responses in regard to specific options. However, some themes emerge across the options. When asked why a particular option is important, many responses focused on the need for public awareness, enhanced transparency into AI systems, and clarification about requirements. Additionally, the need to translate abstract norms into operationalized practice was also identified as important to ensuring compliance. Nevertheless, preventing and rectifying harms, protecting individuals and society, and access to justice were not mentioned as a reason why governance measures are important.

When asked how measures should be implemented and by whom, one notable theme was the breadth of actors identified. A few specific actors were mentioned more than once in the context of multiple potential governance measures: industry/companies, government, groups of experts, academia, and the EU. Many others were only cited once, including national organisations, EU projects, professional bodies, open source repositories, standardisation bodies, global consultancies, think tanks, and global law firms. This range of actors helps illustrate how widespread the ecosystem for AI governance could be. In fact, some respondents simply referred to some variation of 'all stakeholders', suggesting there is a role and responsibility for all.

Responses to the question of indicators to measure success also varied greatly in specificity. While some respondents answered generally with ambitions like 'world peace' and 'well-being of human affected', other responses were more concrete (e.g. number of articles published, reported use of ethical tools in corporate annual reporting, number of open data sets). However, once again, there was no explicit mention of stopping harms or holding accountable those responsible.

There were only three responses to the question 'Which potential measures should not be prioritised?', and none were from the list of 15 measures. One respondent cautioned against more "heavy in-depth research" as "there is already a lot of pioneering in the market and this kind of research is delaying the response time frame [for] Europe." Another respondent listed "reporting of current activities"; however, without further clarification, it is not clear which activities and whose reporting is meant. The third respondent had a very specific and time-bound recommendation against "making changes to the GDPR post-Brexit until we fully understand the implications."

Lastly, the final additional comment identified two issues not reflected among the top 15 potential governance measures (and which never emerged in R1 or R2): data trusts and digital identity. For this respondent, both deserve attention as a matter of priority.

4. Conclusions

4.1 Limitations

A Delphi study is a well-established methodology to find solutions to complex and multi-faceted problems. Nevertheless, there were unexpected challenges that impacted the results and the conclusions that can be drawn from them.

One unexpected limitation was a lack of engagement. In the DoA, it was anticipated there would be feedback from 60 experts. Knowing that it was unlikely all experts contacted would respond, a total of 231 experts received invitations to complete all three rounds of the study, and 100 experts began the survey in R1. Unfortunately, the maximum number of responses received was 43 (in Round 3). This concern was identified early in Round 1, and multiple methods were attempted to increase participation. In all three rounds, follow-up emails were sent to the panel, including tailored emails sent directly from the SHERPA Project Coordinator, and personal messages from our Stakeholder Board coordinator to members of the SHERPA Stakeholder Board on the panel. After a lower than expected response rate in Round 1, consideration was given to expanding the participant list to include more experts, but that was not done in order to preserve the integrity of the Delphi methodology.

Additionally, in Round 2, the SHERPA project offered token vouchers worth €10 to compensate respondents for their time completing the survey. This did not increase the number of responses, and no respondent contacted the project to claim the voucher. There are many possible explanations for the low response rate. The panel may have been experiencing consultation fatigue; in 2019-2020, there have been numerous opportunities for experts in AI and related fields to provide feedback and expertise on the future of AI. Additionally, the survey questions may have seemed too long or time-consuming; Round 2 in particular was a lengthy questionnaire. While the intention was to help the respondents by providing as much information as possible, the reality may have been that the options were too overwhelming. Lastly, the second and third rounds were both conducted in the midst of the COVID-19 global pandemic. For many reasons, the pressures on experts' time were great, and time for external surveys more challenging to capture. While disappointing, the lower than expected response rate does not jeopardise the validity of the Delphi study exercise. Delphi studies, as outlined earlier, are not meant to be statistically representative of larger populations, but they are a method for understanding complex future-oriented questions. The insights gathered from this Delphi study confirm that this aim was achieved.

Another unexpected challenge was the vague language and lack of specificity in many responses. As the responses were anonymous and the surveys administered online, it was not feasible to ask respondents for clarification. As a result, very often the meaning of a response was not clear. This particularly impacted how respondents and the SHERPA Delphi team categorised and interpreted potential governance measures. For example, 'stakeholder dialogue and scrutiny' is a broad concept and can have many different meanings. As a result, it is not clear whether the SHERPA understanding of this concept is the same as the respondents' understanding, or even if all the respondents understood the same meaning. This is, however, a typical problem of qualitative research, where the meaning of terms by participants may differ from the meaning of the same terms as interpreted by the researchers. It points to possible future research where the detailed implications of key concepts should be defined more clearly.

A final limitation of the study related to the breadth of issues addressed and the level of multiple expertise required to provide robust, informed responses. Particularly in Rounds 2 and 3, many questions required an in-depth knowledge of the wide-range of measures across the AI-ecosystem. In retrospect, perhaps it was too ambitious to expect all experts to be able to comment meaningfully on areas outside of their

expertise. This challenge was exacerbated by the fact that the survey was administered online and respondents could not readily ask for clarification.

4.2 Lessons Learned

While the SHERPA Delphi study did identify solutions, it was more useful as an illustration and mapping of the complexity of the concerns associated with AI and Big Data and the potential governance measures to address those concerns. There were many consistent themes over the three rounds that are familiar, including concerns about lack of transparency, impact of bias and discrimination in AI systems, and a need for more public awareness. Yet there were also many responses that stood alone and did not fit easily within emerging categorisations. And there were some notable omissions in the responses, which were identified in other SHERPA activities. The breadth of responses and their varying degrees of specificity illustrate that – even among experts – opinions and knowledge about the most pressing concerns and possible solutions are diverse. The impacts of AI and Big Data are felt in very different ways by different stakeholder groups, and that was reflected in the differing, and often opposing, responses received. Therefore, while there was no overwhelming consensus on which solutions to prioritise, it is clear that the complexity of the AI and Big Data ecosystem requires a 'smart mix' of measures and all stakeholders have roles to play.

4.2.1 Ethical and Human Rights Issues

In general, the results of the Delphi study in identifying the 'most important ethical and human rights issues' was consistent with research and findings in other SHERPA activities, including stakeholder interviews, focus groups, online survey,[27] and feedback from the stakeholder board. The top responses focused on well-known and well-documented concerns currently impacting end-users in Europe.

From R1 and R2, the highest scoring issues were **lack of privacy** and **bias and discrimination**, along with **misuse of personal data**, **lack of access to and freedom of information**, and **impacts on democracy**. **Lack of transparency** in particular, though not a highest scoring issue in R2, was a key feature of the justification for many potential governance measures. Other top concerns related to power distributions (concentration of economic power, power asymmetries), lack of trust, potential for criminal/malicious use, disappearance of jobs, violation of end-users' rights, and loss of autonomy.

There was also consistency in the lower scoring concerns. Three were suggested directly by respondents in R1: 'Awakening' of AI, cost to innovation, and prioritisation of the 'wrong' problems. 'Awakening' of AI is particularly notable because it scored among the top issues for the amount of attention it receives, but was last in regard to the impact of its reach and significance on individuals. Given the debate around the likelihood and immediacy of artificial general intelligence, it is perhaps not surprising to see respondents reject this concern in favour of what many perceive to be current issues. This also illustrates that respondents understood 'most important' to mean most relevant to Western societies currently and in the immediate future, and not necessarily potential far-future concerns. This distinction could also help explain why other more 'distant' issues were also not prioritised as important, including the potential for physical harm, impact on the environment, and violations of human rights within the AI-systems supply chain.

4.2.2 Potential Governance Measures

The results of the Delphi study in prioritising the most important potential governance measures for immediate action are also consistent with other SHERPA research, and generally reflect the difficulty of determining consensus on how to address concerns. Furthermore, the Delphi did identify a broad range

[27] Brooks, Laurence; Stahl, Bernd; Jiya, Tilimbe (2020): D2.3 Online survey report. De Montfort University. Online resource. https://doi.org/10.21253/DMU.11777478.

of potential governance measures, many of which have been the subject of further analysis in other SHERPA activities. However, the ultimate prioritisation of potential governance measures by the respondents does not wholly reflect the general recommendations emerging from other SHERPA activities.

The top-ranked option was a technical measure: **methodologies for systematic and comprehensive testing of AI-based systems**. Also high-ranking were two other technical measures: **tools for verifying and certifying labelled datasets for accuracy, bias and other important properties** and **more open source tools that allow for transparency, explainability, and bias mitigation**. However, the majority of the top-scoring potential measures in R2, prioritised in R3, were soft-law initiatives and generalised strategies for stakeholder engagement. The measures include **guidelines and toolkits**, **stakeholder dialogue**, **ethical frameworks**, **public policy commitments**, **exchange of best practices**, **codes of conduct**, and **education campaigns**. The fact that many of these initiatives directly involve the general public echoes other calls for increased public awareness of (and engagement with) the impacts of AI and Big Data. Additionally, the prevalence of these types of initiatives has been noted in many other SHERPA activities, and reflects the need for concrete tools to help translate ethical principles and human rights norms into practice.

However, most of the highest-ranked options are not precisely defined (in either the context of this Delphi and in broader conversations), and none specify any binding implementation methods. Furthermore, there was no consensus in the responses on the entity/entities responsible for implementation and the target audience for these options; thus, there is role for everyone, but no one is accountable. By comparison, none of the potential regulatory options scored high enough in R2 to be listed in R3. Despite the fact that regulation was the most frequently cited potential measure in the R1 brainstorm, specific regulatory proposals made up the majority of bottom-scoring options in R2. In some sense, these results reflect consensus toward the least common denominator; most of the top-ranking potential options are relatively easy to implement, inexpensive, and feasible in a short timeframe. For example, when faced with the choice between binding regulation and stakeholder dialogue, it is understandable that many would select the latter, as the former is time-consuming and politically challenging. Therefore, even if stakeholder dialogue is potentially less effective, it is preferable to no action. For this reason, it was not surprising to see the softest options emerge on top.

The prioritisation of potential governance measures does not wholly reflect the general recommendations emerging in parallel from other SHERPA activities. Like the Delphi study, the SHERPA project will prioritise stakeholder engagement, educational tools (tailored to different stakeholder groups), and concrete tools to translate principles into practice. However, unlike the Delphi panel, the SHERPA project (based on its research in its other activities) is recommending a stronger regulatory framework at the EU level, as well as an EU Agency for AI, impact assessments, standardisation on AI ethics, and the establishment of AI 'ethics' officers within organisations. All of these were presented to the Delphi panel in R2, but did not make it through to prioritisation in R3.

4.2.3 Final Observations

Ultimately, the primary value of this Delphi study was not in the final results, but in the exercise of mapping the concerns and potential solutions. Even though the panel was comprised of 'experts', the expertise represented was diverse. As a result, there was a broad range of responses and little consensus. Frequently, two responses were directly contradictory, but both were equally valuable. Furthermore, the method of carrying out this study online via email often made it very difficult to discern the actual meaning of a respondent's answer and impossible to know for certain whether two similar responses were in fact the same. As a result, the possible solutions are plentiful, but rarely clear. Conducting this Delphi study has highlighted how critically important it will be in implementing governance measures for AI and Big Data to carefully and clearly frame language and articulate precise recommendations for discrete audiences. This is a challenge not only for SHERPA, but for all stakeholders, and will inform the further development of SHERPA's final recommendations.

Appendix A: Participant Interaction

Appendix A1: Round 1 Invitation email (Nov. 13, 2019)

Dear [participant name],

On behalf of the SHERPA project consortium, I would like to invite you to participate in the SHERPA Delphi study. The study has the aim to identify and prioritise ways in which ethical and human rights impacts of artificial intelligence and big data should be addressed. The results of the Delphi will shape the outcomes of the SHERPA project which provides policy advice to the European Commission.

You have been selected as a leading expert in the field and because we believe that your insights can help us ensure that the coverage of the topic area is comprehensive. We are aiming to enrol 60 experts with a range of backgrounds to provide a broad understanding of these issues.

We now will ask you to contribute to the first of what will be three rounds of the Delphi study. We estimate that responding to this round of questions should take no more than 40 minutes, but it will of course depend on how much detail you wish to provide in your responses. We hope that you will then also respond to the subsequent two rounds of the study.

The SHERPA consortium will analyse and synthesise the responses from panel members into a brief paper (e.g., 10 pages) which we will send to panel members for review and comment together with a second round of questions flowing from the synthesis.

We hope you will participate in this important study by clicking on this link:

[Link to Begin the Delphi Study]

Thank you very much

on behalf of the SHERPA consortium

Bernd Stahl

Professor Bernd Carsten STAHL
Director, Centre for Computing and Social Responsibility
School of Computer Science and Informatics
Faculty of Computing, Engineering and Media

DE MONTFORT UNIVERSITY
T: +44 116 207 8252
E: bstahl@dmu.ac.uk
W: http://dmu.ac.uk/berndstahl

Appendix A2: Welcome message displayed on website

Dear Participant

We would like to ask you to participate in the data collection for the European Research Project SHERPA (Shaping the ethical dimensions of smart information systems– a European perspective (SHERPA). SHERPA is a project funded by the European Commission (Grant no. 786681; www.project-sherpa.eu). It is led by De Montfort University in Leicester, UK and includes 11 partner organisations from across Europe.

In collaboration with stakeholders, the SHERPA project will investigate, analyse and synthesise 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. It will develop novel ways of understanding and addressing SIS challenges, evaluate with stakeholders, and advocate the most desirable and sustainable solutions.

The **SHERPA consortium is undertaking a Delphi study to gather feedback on the nature of these issues, ways of addressing them and how to prioritise them.** The Delphi method will allow the invited experts to work towards a mutual agreement by responding to a set of questions. The Delphi study comprises 3 rounds of questioning. The first round comprises 5 open questions. Filling in these study will take no more than 40 minutes. We hope that all participants in the first round will agree to participate in the subsequent rounds as well.

You are invited to participate in this Delphi Study because of your expertise in the area. Your participation in this study is entirely voluntary. If at any point you want to withdraw from the study and would like you data to be deleted, you can do so by contacting the SHERPA project coordinator Prof. B. Stahl at bstahl@dmu.ac.uk or +44 116 207 8252.

There are no known or anticipated risks arising from your participation in this study. We will pseudonymise any personal information we collect from you. We will use your email address to generate a unique identifier for their entries across the three rounds of questioning. This identifier will look and function in the same way as a username but will be unrecognisable to anyone analysing the information without pre-approved access to the original mailing list. Each identifier will be a unique combination of numbers and letters carried by a user throughout the three sections of the study, allowing for the collection of data at the end of the study and anonymous entries too.

The surveys and subsequent data collection we will created using the SHERPA website and its inbuilt forms system. This allows us to retain full control of what data we store, how we store it and who receives access to it. All entries will be stored within a password protected area of the website that is also encrypted by 256-bit SSL, 4-hourly security sweeps, version tracking and user access logs. Data can be exported via CSV files by those with the correct account permissions ensuring that even the anonymised data is secure.

All individual responses will be stored on a secure server. Responses will be anonymised and made available to other participants to generate discussion. Data analysis will be undertaken by consortium partners using well-established collaborative software tools (NVivo Server). For each round of the Delphi Study there will be a report that will be made public as part of the project's obligations to provide open access to data.

The Delphi Study data will be used to inform the SHERPA project, be the basis of a public deliverable and of academic publications. Following the conclusion of the analysis and SHERPA work, any link to individuals will be deleted. The data will be retained on DMU's data management platform (www.figshare.com) for future use in relevant research.

Your name or any other personal identifying information will not appear in any publications resulting from this study.

If you have any questions regarding this study or would like additional information please contact the SHERPA consortium: https://www.project-sherpa.eu/contact/.

By filling in this survey you indicate that you understand its purpose and consent to the use of the data as indicated above.

Thank you for your cooperation.

Professor Bernd Stahl

on behalf of the SHERPA Project Consortium

I agree with the use of my responses for research purposes of the SHERPA project as outlined above.

Yes

No

Appendix A3: Round 1 Follow-up email (Dec. 4, 2019)

Dear <<First Name>>,

This is just a brief reminder that you have been invited to the SHERPA Delphi Study.

The study has the aim to identify and prioritise ways in which ethical and human rights impacts of artificial intelligence and big data should be addressed. The results of the Delphi will shape the outcomes of the SHERPA project which provides policy advice to the European Commission.

You have been selected as a leading expert in the field and because we believe that your insights can help us ensure that the coverage of the topic area is comprehensive. We are aiming to enrol 60 experts with a range of backgrounds to provide a broad understanding of these issues.

We now will ask you to contribute to the first of what will be three rounds of the Delphi study. We estimate that responding to this round of questions should take no more than 40 minutes, but it will of course depend on how much detail you wish to provide in your responses. We hope that you will then also respond to the subsequent two rounds of the study.

The SHERPA consortium will analyse and synthesise the responses from panel members into a brief paper (e.g., 10 pages) which we will send to panel members for review and comment together with a second round of questions flowing from the synthesis.

We hope you will participate in this important study by clicking on the button below...

[Link to Begin the Delphi Study]

Thank you very much, on behalf of the SHERPA consortium

Bernd Stahl

Appendix A4: Round 1 Follow-up email (Jan. 10, 2020)

Dear <<First Name>>,

Our very best wishes for the New Year!

If you are still interested in participating in the study, please do so by **Wednesday 15 January, 2020**. The study takes only 20 to 30 minutes to complete; we know everyone is busy these days.

On our side, we are committed to finalising the synthesis of findings by the end of the month and share with all participants some of the very interesting findings that already transpire from the analysis of findings.

We look forward to your views.

[Link to Begin the Delphi Study]

Thank you very much,

on behalf of the SHERPA consortium

Bernd Stahl

Appendix A5: Round 2 Invitation email (March 13, 2020)

[also sent March 18 and April 7, 2020]

Dear <<First Name>>,

On behalf of the SHERPA project consortium, I would like to thank you for participating in the first round of our Delphi Study that explores ethical and human rights aspects of artificial intelligence and big data. We have analysed the first round of responses. You can find a report summarising key findings here:

[Link to Download the Report]

The first round of the Delphi Study, together with other work undertaken in the SHERPA project, has provided rich insights into the range of issues and possible ways of addressing them. We would now like to use the second round of the study to rate and prioritise these issues and mitigation measures.

We therefore ask you to help us by providing your expert insights to prioritise these items. This round of the Delphi Study contains mostly structured and closed questions, asking you to rate issues and measures against three criteria. You will have the option to provide additional written comments, but that is not required.

We estimate that responding to this round of questions should take no more than 40 minutes.

The results of this second round will again be summarised and used as the basis of a final round, to be launched prior to the summer. The results of the entire Delphi Study will shape the outcomes of the SHERPA project which provides policy advice to the European Commission.

We hope you will participate in this important study by clicking on the button below...

[Link to Participate in the Delphi Study]

Thank you very much, on behalf of the SHERPA consortium
Bernd Stahl

Appendix A6: Round 2 Follow-up email (May 12, 2020)

Subject: AI for a better future - request for your support

Dear [first name]

I hope this email finds you well in this difficult time.

The reason why I am sending this email is that I would like to ask you for your help concerning the SHERPA project (www.project-sherpa.eu). SHERPA explores ethics and human rights in artificial intelligence and big data and will contribute to the debate about how AI can be used to make a better future.

Based on your expertise we invited you to contribute to the first round of our Delphi study. The findings from this first round are available here: https://www.project-sherpa.eu/workbook/

By now, you should have received an email requesting your input into the second round. This round consists of a number of questions that will help us rank the various ethical and human rights issues and ways of addressing them.

I realise that this is a difficult time for most of us. But I hope that you appreciate that this work is important and needs your input. There are four sets of questions: ethical and human rights issues; potential regulatory measures; potential technical measures; and other potential measures. We estimate that responding to all questions will take about 40 minutes, but your response is still extremely valuable even if you are only able to comment on issues most relevant to your expertise or areas of interest.

In order to show our appreciation, we offer you a shopping voucher worth £10, if you complete the survey.

We would be grateful, if you could consider this request. If you are willing to support us, then please fill in the Delphi survey here:

https://www.project-sherpa.eu/delphi-study/part-two/?delphiid=[delphiid]

Thank you very much in advance,

Bernd

Appendix A7: Round 3 Invitation email (Sept. 18, 2020)

Dear <<First Name>>

We would like to thank you for your continued willingness to contribute to SHERPA's Delphi study on the ethical and human rights issues of AI and big data (smart information systems). This third and final round of the study is based on input from the first and second rounds (see here for summaries) and other activities of the SHERPA project. Having identified and rated ethical and human rights issues and

potential governance measures, the final step is to prioritise these options. Therefore, in this final survey, we need you to select the most important measures for immediate action.

The survey asks you to select three options from among the fifteen highest scoring options from the second round. These fifteen options received the highest average scores based on desirability, probability, and feasibility. To help us formulate final recommendations, we also ask you to explain (a) why the measure you selected is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure.

There is also an option for you to identify any potential governance measures that should not be included in our final recommendations, as well as space to provide any additional comments.

The survey should take no more than 10 minutes to complete.

Thank you again for your contributions to the SHERPA project.

[link to survey]

Thank you very much, on behalf of the SHERPA consortium

Bernd Stahl

Appendix A8: Round 3 Follow-up email (Sept. 25, 2020)

Dear << Test First Name >>

We would like to thank you for your continued willingness to contribute to SHERPA's Delphi study on the ethical and human rights issues of AI and big data (smart information systems). This is just a quick reminder about this study as this third and final round of the study is based on input from the first and second rounds (see here for summaries) and other activities of the SHERPA project. Having identified and rated ethical and human rights issues and potential governance measures, the final step is to prioritise these options. Therefore, in this final survey, we need you to select the most important measures for immediate action.

The survey asks you to select three options from among the fifteen highest scoring options from the second round. These fifteen options received the highest average scores based on desirability, probability, and feasibility. To help us formulate final recommendations, we also ask you to explain (a) why the measure you selected is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure.

There is also an option for you to identify any potential governance measures that should not be included in our final recommendations, as well as space to provide any additional comments
The survey should take no more than 10 minutes to complete.

Thank you again for your contributions to the SHERPA project.

[Link to Participate in the Delphi Study]

Thank you very much, on behalf of the SHERPA consortium
Bernd Stahl

Appendix A9: Round 3 Follow-up email (Oct. 1, 2020)

Dear XXX

I want to thank you personally for your contributions to the first two rounds of the SHERPA Delphi study. Your valuable input, drawn from your expertise and experience, will help us better understand how to address ethical and human rights concerns with AI and big data.

As the SHERPA project moves toward developing final recommendations, we need your help in completing the third and final round of the study. I know that time is precious, so we have carefully designed the questions to take less than 10 minutes to complete.

The survey will close next Wednesday, October 7th. I would really appreciate having your feedback.

https://www.project-sherpa.eu/delphi-study/part-three/?delphiid=xxx

Thank you very much in advance,

Bernd

Appendix B: Delphi Survey Questions

Appendix B1: Round 1 Questions

1. What do you think are the three most important ethical or human rights issues raised by AI and / or big data?
2. Which current approaches methods, or tools for addressing these issues are you aware of? These may be organisational, regulatory, technical or other.
3. What do you think are the pros and cons of these current approaches, methods, or tools?
4. What would you propose to address such issues better?
5. Which should be the top 3 criteria for society to select and prioritise the most appropriate measures?

As part of the SHERPA Delphi Study we would like to understand whether there are particular issues or questions that are of particular interest to certain groups. We would therefore like to ask you to provide us the following information about yourself

1. Gender
    a. female
    b. male
    c. other / not disclosed
2. country of residence []
3. age in years
4. Would you consider yourself to represent the views of any of the following stakeholder groups (multiple options are possible)
    a. researcher
    b. policymaker
    c. industry
    d. civil society
    e. other (please name)

Thank you very much for taking part in this first round of the SHERPA Delphi Study. We will analyse the responses of this first round to come to an understanding of the shared view of our panel members. Once this analysis has been completed, we will share these findings with you and ask you for further feedback in the second round of the study.

Please indicate whether you would like to receive a copy of the report containing the analysis of this first round of the Delphi Study

yes/no

Please indicate whether you are happy to be invited to participate in the second round of the Delphi Study

yes / no

Appendix B2: Round 2 Questions

Question 1: The following ethical and human rights issues and possible mitigation measures (question 2,3,4) were taken from the DELPHI Round 1 responses. The issues were supplemented with issues identified in other activities of the SHERPA project, including analysis of case studies, stakeholder interviews, and an online survey.

Please rate the ethical and human rights issues in terms of:

- **Reach** (number of people affected)
- **Significance** (impact on individuals)
- **Attention** (likely to lead to public debate)

Issues should be rated on a score of 1 to 5. A low score (1) means the issue affects few (or no) individuals, is trivial, / or is not of serious concern. A high score (5) means the issue affects individuals worldwide, has vital consequences, / or is likely to generate robust public debate. In the last column, please provide a brief explanation of why you hold this opinion.

- Example : Lack of Privacy
    - **Reach:** You may think this is an issue that affects all individuals who have access to the Internet, but not all individuals worldwide. Reach rating: 4
    - **Significance:** You may think this is an issue that is critically important to every individual using Internet-based services because there are potential consequences to many facets of life. Significance rating: 5
    - **Attention:** You may think this issue has already led to robust public debate internationally. Attention rating: 5
- However,
    - **Reach:** You may think this will affect only a very limited number of individuals on a local level. Reach rating: 2
    - **Significance:** You may think this is an issue that is not important to every individual using Internet-based services because there are not potential consequences to many facets of life. Significance rating: 1
    - **Attention:** You may think this issue will not lead to any debate. Attention rating: 1

Ethical and Human Rights Issues	**Reach**	**Significance**	**Attention**	**Average**
Lack of Privacy Related to which type of data and how much data is collected, where from, and how it is used				
Misuse of Personal Data Related to concerns over how SIS might use personal data (e.g. commercialization, mass surveillance)				
Lack of Transparency Related to the public's need to know, understand, and inspect the mechanisms through which SIS make decisions and how those decisions affect individuals				
Bias and Discrimination Related primarily to how sample sets are collected/chosen/involved in generating data and how data features are produced for AI models; and how decisions are made (e.g. resource distribution) according to the guidance arising out of the data				

Unfairness Related to how data is collected and manipulated (ie. how it is used), also who has access to the data and what they might do with it as well as how resources (eg. Energy) might be distributed according to the guidance arising out of the data				
Impact on Justice Systems Related to use of SIS within judicial systems (e.g. AI used to 'inform' judicial reviews in areas such as probation)				
Impact on Democracy Related to the degree to which all involved feel they have an equal say in the outcomes, compared with the SIS				
Loss of Freedom and Individual Autonomy Related to how SIS affects how people perceive they are in control of decisions, how they analyse the world, how they make decisions (e.g. impact of manipulative power of algorithms to nudge toward preferred behaviours), how they interact with one another, and how they modify their perception of themselves and their social and political environment				
Human Contact Related to the potential for SIS to reduce the contact between people, as they take on more of the functions within a society				
Loss of Human Decision-Making Related to how SIS affects how people analyse the world, make decisions, interact with one another, and modify their perception of themselves and their social and political environment				
Control and Use of Data and Systems Related to how data is used and commercialised, including malicious use (e.g. mass surveillance); how data is collected, owned, stored, and destroyed; and how consent is given				
Potential for Military Use Related to the use of SIS in future possible military scenarios (e.g. autonomous weapons), including the potential for dual-use applications (military and non-military)				
Potential for Criminal and Malicious Use Related to the use of SIS in criminal and malicious scenarios (e.g. cyber-attacks and cyber espionage)				

Ownership of Data Related to who owns data, and how transparent that is (e.g. when you give details to an organisation, who then 'owns' the data, you or that organization?)				
Lack of Informed Consent Related to informed consent being difficult to uphold in SIS when the value and consequences of the information that is collected is not immediately known by users and other stakeholders, thus lowering the possibility of upfront notice				
Lack of Accountability and Liability Related to the rights and legal responsibilities (e.g. duty of care) for all actors (including SIS) from planning to implementation of SIS, including responsibility to identify errors or unexpected results				
Accuracy of Predictive Recommendations Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS interprets an individual's personal data				
Accuracy of Non-Individualized Recommendations Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS makes a decision based on data not specific to an individual				
Power Relations Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to 'act above the law'				
Concentration of Economic Power Related to growing economic wealth of companies controlling SIS (e.g. big technology companies) and individuals, and unequal distribution of resources				
Power Asymmetries Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to 'act above the law'				
Lack of Access to and Freedom of Information				

Related to quality and trustworthiness of information available to the public (e.g. fake news, deepfakes) and the way information is disseminated and accessed				
Accuracy of Data Related to using misrepresentative data or misrepresenting information (ie. predictions are only as good as the underlying data) and how that affects end user views on what decisions are made (ie. whether they trust the SIS and outcomes arising from it)				
Integrity Related to the internal integrity of the data used as well as the integrity of how the data is used by a SIS				
Impact on Health Related to the the use of SIS to monitor an individual's health and how much control one can have over that				
Impact on Vulnerable Groups Related to how SIS creates or reinforces inequality and discrimination (e.g. impacting on the dignity and care for older people, for example how much a care robot might exert over an older person's life and 'tell them what to do'				
Violation of End-Users Fundamental Human Rights Related to how human rights are impacted for end-users (e.g. monitoring and control of health data impacting right to health; manipulative power of algorithms nudging towards some preferred behaviours, impacting rights to dignity and freedom				
Violation of Fundamental Human Rights in Supply-Chain Related to how human rights are impacted for those further down the supply-chain extracting resources and manufacturing devices (e.g. impacts on health, labour violations, lack of free, prior and informed consent for extractives				
Lack of Quality Data Related to using misrepresentative data or misrepresenting information in building AI models				
Disappearance of Jobs Related to concerns that use of SIS will lead to significant drop in the need to employ people				

Prioritization of the "Wrong" Problems Related to the problems SIS is developed to 'solve' and who determines what the immediate problems are				
"Awakening" of AI Related to concerns about singularity, machine consciousness, super-intelligence etc. and the future relationship of humanity vis-a-vis technology				
Security Related to the vulnerabilities of SIS and their ability to function correctly under attacks or timely notify human operators about the need of response and recovery operations				
Lack of Trust Related to using misrepresentative data or misrepresenting information (ie. predictions are only as good as the underlying data) and how that affects end user views on what decisions are made (ie. whether they trust the SIS and outcomes arising from it); also related to informed consent and that helps with trust				
Access to Public Services Related to how SIS could change the delivery and accessibility of public services for all (e.g. through privatisation of services)				
Harm to Physical Integrity Related to the potential impacts on our physical bodies (e.g. from self-driving cars, autonomous weapons)				
Cost to Innovation Related to balancing the protection of rights and future technological innovation				
Unintended, Unforeseeable Adverse Impacts Related to future challenges and impacts that are yet known				
Impact on Environment Related to concern about the environmental consequences of infrastructures and devices needed to run SIS (e.g. demand for physical resources and energy)				
Do you have any further comments regarding Ethical and Human Rights Issues?				

Question 2: The following potential regulatory measures originated from the DELPHI Round 1 responses. The examples given were refined and supplemented by analysis conducted in other deliverables of the SHERPA project, including a report on regulatory options.

Please rate the following potential regulatory measures in terms of:

- **Desirability** (would you like to have this measure in place?)
- **Feasibility** (in theory, is it possible to have this measure in place?)
- **Probability** (in reality, is it likely that this measure would be put in place?)

Issues should be rated on a score of 1 to 5. A low score (1) means the measure will have a major negative effect, is very challenging to create, and / or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and / or is very likely to happen.

- Example : New International Treaty
    - **Desirability:** You may think measure is very appealing because it would create legal obligations and has the potential to have a positive impact globally. Desirability rating: 5
    - **Feasibility:** You may think this measure is theoretically feasible because the international community, through the United Nations, has processes in place for negotiating and adopting an international agreement. However, you recognize that international treaties generally take a significant amount of time to finalize. Feasibility rating: 4
    - **Probability:** You may think that, given the current international geopolitical context, it is very unlikely that States could cooperate and agree on an international binding treaty. Probability rating: 1
- However,
    - **Desirability:** You may think measure is not appealing because it is unlikely to have any impact. Desirability rating: 1
    - **Feasibility:** You may think this measure is not feasible because it does not fall within the mandate of any international regulatory bodies. Feasibility rating: 1
    - **Probability:** You may think that, given the current international geopolitical context, it is very likely that States could cooperate and agree on an international binding treaty. Probability rating: 5

Potential Regulatory Measures	**Desirability**	**Feasibility**	**Probability**
Creation of new international treaty for AI and Big Data (open for adoption by all countries)			
Better enforcement of existing international human rights law			
Binding Framework Convention to ensure that AI is designed, developed and applied in line with European standards on human rights, democracy and the rule of law (Council of Europe) including through a new ad hoc committee on AI (CAHAI)			
CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment			

Legislative framework for independent and effective oversight over the human rights compliance of the development, deployment and use of AI systems by public authorities and private entities (Council of Europe)			
General fund for all smart autonomous robots or individual fund for each and every robot category (EU Parliament)			
Establishment of a comprehensive Union system of registration of advanced robots within the Union's internal market where relevant and necessary for specific categories of robots and establishment of criteria for the classification of robots			
Algorithmic impact assessments under the General Data Protection Regulation (GDPR)			
Creation of new body: **EU Taskforce/Coordinating body of field-specific regulators for AI/big data**			
Redress-by-design mechanisms for AI (High-Level Expert Group on Artificial Intelligence (AI HLEG)			
New laws regulating specific aspects, e.g., deepfakes, algorithmic accountability.			
Register of algorithms used in government			
New national independent cross-sector advisory body (e.g. UK Centre for Data Ethics and Innovation)			
New specialist regulatory agency to regulate algorithmic safety			
Public Register of Permission to Use Data (individuals provide affirmative permission in a public register for companies to use their data)			
Reporting Guidelines (for publicly registered or traded companies based on corporate social responsibility reporting as described by GRI)			
Regulatory sandboxes for AI and big data			
Three-level obligatory impact assessments for new technologies			
Do you have any further comments regarding Potential Regulatory Measures?			

Question 3: The following potential technical measures originated from the DELPHI Round 1 responses. The examples given were refined and supplemented by analysis conducted in other deliverables of the SHERPA project, including a report on security in SIS.

As in Question #2, please rate the following potential technical measures in terms of:

- **Desirability** (would you like to have this measure in place)
- **Feasibility** (is it possible to have this measure in place)
- **Probability** (is it likely that this measure would be put in place)

Issues should be rated on a score of 1 to 5. A low score (1) means the measure will have a major negative effect, is very challenging to create, and / or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and / or is very likely to happen.

Potential Technical Measures	**Desirability**	**Feasibility**	**Probability**
Methodologies for systematic and comprehensive testing of AI-based systems (incldluding fairness of decisions)			
Techniques for providing explanations for output of AI models (e.g., Layerwise relevance propagation for neural networks)			
Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models			
AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model)			
Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties			
Tools for verifying and certifying publicly available services based on machine learning models			
Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models)			
Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public)			
Do you have any further comments regarding Potential Technical Measures?			

Question 4: The following other potential measures originated from the DELPHI Round 1 responses. The examples given were refined and supplemented by analysis conducted in other deliverables of the SHERPA project.

As in Question #2-3, please rate the following potential measures in terms of:

- **Desirability** (would you like to have this measure in place)
- **Feasibility** (is it possible to have this measure in place)
- **Probability** (is it likely that this measure would be put in place)

Issues should be rated on a score of 1 to 5. A low score (1) means the measure will have a major negative effect, is very challenging to create, and / or is impossible to achieve. A high score (5) means the measure will have a very positive effect, is not difficult to create, and / or is very likely to happen.

Other Potential Measures	**Desirability**	**Feasibility**	**Probability**
Certification (e.g. initiative for IEEE Ethics Certification Program for Autonomous and Intelligent Systems)			
Citizen Juries to evaluate risk of various AI technologies and propose appropriate tools			
Education Campaigns (e.g. Finnish Element of AI course; Dutch Nationale AI Cursus)			
Ethical Codes of Conduct (e.g. EU High Level Expert Group Guidelines for Trustworthy AI, SHERPA guidelines)			
Ethical Mindset adopted by companies			
Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases (e.g. rules governing recommendation systems: how they should work, what they should not be used for, how they should be properly hardened against attacks, etc.)			
Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments (e.g., AI robots resembling dogs, sex robots)			
Exchange of Best Practices			
'Fairness' Officer or Ethics Board employed within companies using/developing SIS			
Framework, Guidelines, and Toolkits for project management and development (e.g. UK Data Ethics Framework; IBM AI Fairness 360 Open Source Toolkit; Dutch Data Ethics Decision Aid (DEDA) tool)			

Grievance Mechanisms for complaints on SIS			
High-level Expert Groups (e.g. UN AI for Good Global Summit)			
Individual Action (e.g. participating in conferences to raise awareness; protecting oneself by refusing cookies online)			
International Ethical Framework (e.g. OECD Principles on AI)			
Investigative Journalism about issues concerning SIS			
More Open Source Tools that allow for transparency, explainability, and bias mitigation			
NGO Coalitions on particular issues (e.g. Campaign to Stop Killer Robots)			
Open Letters to governments and the public (e.g. 2015 Open Letter on AI)			
Public Policy Commitment by company to be ethical			
Public "Whistleblowing" Mechanisms for the reporting of bias, inaccuracies, or ethical impacts of systems based on machine learning models			
Retaining 'Unsmart' Products and Services by keeping them available to purchase and use			
Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations (e.g. self-driving vehicles and other systems)			
Self-Regulation by Company (e.g. Twitter's self-imposed ban on political ads)			
Stakeholder Dialogue and Scrutiny with scientists, programmers, developers, decision makers, politicians, and the public at large			
Standardisation (e.g. IEEE P7000 series of standards for addressing ethical concerns during system design).			
Third-party Testing and External Audits (e.g. of data used for training for quality, bias, and transparency)			
Do you have any further comments regarding Other Potential Measures?			

Appendix B3: Round 3 Questions

Welcome

We would like to thank you for your continued willingness to contribute to SHERPA's Delphi study on the ethical and human rights issues of AI and big data (smart information systems). This third and final round of the study is based on input from the first and second round (see here for summaries) and other activities of the SHERPA project. Having identified and rated ethical and human rights issues and potential governance measures, the final step is to prioritise these options. Therefore, in this final survey, we need you to select the most important measures for immediate action.

The survey asks you to select three options from among the fifteen highest scoring options from the second round. These fifteen options received the highest average scores based on desirability, probability, and feasibility. To help us formulate final recommendations, we also ask you to explain (a) why the measure you selected is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure. There is also an option for you to identify any potential governance measures that should not be included in our final recommendations, as well as space to provide any additional comments.

In this survey we do not collect any personal data. By ticking the following box, you indicate that you are happy to contribute to the study and provide your insights for the benefits of the SHERPA data analysis.
*I am happy to contribute to the study

Potential Measures

Please select the **three most important measures for immediate action**. It does not matter the order of your selection. For each measure, please explain (a) why the measure you selected is important, (b) how the measure should be implemented and by whom, and (c) what indicators would show the successful implementation of the measure.

Please select three choices from the list below.

- Investigative journalism about issues concerning SIS
- Exchange of best practices
- Education campaigns
- Framework, guidelines, and toolkits for project management and development
- Ethical codes of conduct
- High-level expert groups
- More open source tools that allow for transparency, explainability, and bias mitigation
- Grievance mechanisms for complaints on SIS
- NGO coalitions on particular issues
- Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties
- Public policy commitment by company to be ethical
- International ethical framework
- Methodologies for systematic and comprehensive testing of AI-based systems
- Open letters to governments and the public
- Stakeholder dialogue and scrutiny

Explanations and Discussions

[for each option selected]

Why do you think [selected measure] is important?

How do you think this measure should be implemented and by whom?

What indicators would show that this measure has been successful?

Additional Questions

Are there any potential governance measures that should not be prioritized for immediate action and why?

Any additional comments?

Appendix C: Ethics Approval

DE MONTFORT UNIVERSITY LEICESTER

03 October 2019

Dr Tilimbe Jiya
Centre for Computing & Social Responsibility (CCSR)
School of Computer Science & Informatics
Faculty of Computing, Engineering and Media

Dear Tilimbe

Research Ethics Application Approval: 1920/519 - SHERPA project Delphi Study

Your application to gain ethical approval for research activities has been considered and approved by Professor Kathleen Richardson

Your approval is valid for three years from the date of this communication. Should you wish to continue this research after this period, please note that you must resubmit your application using the normal process.

Please be aware that changes to the project plan or unforeseen circumstances may raise ethical issues. If this is the case it is the researcher's duty to repeat the ethics approval process.

Yours sincerely

Professor Kathleen Richardson
Chair, Faculty Research Ethics Committee

Appendix D: Round 1 Response Report

Q1: What do you think are the three most important ethical or human rights issues raised by AI and / or big data?

There were 41 responses to Q1. One Q1 response was deemed more relevant to another question, and two responses to other questions were deemed more relevant to Q1. Therefore, a total of 42 responses were analysed under Q1.

Most respondents identified exactly three issues. A few identified less than three, and a few identified more than three. All responses have been incorporated fully.

Lack of privacy, bias and discrimination, lack of transparency, and loss of human decision-making were the most frequently mentioned concerns. One respondent noted generally that concerns emerge due to a "lack of appropriate or adequate regulation" in a context where "innovation is being led by private actors."

Lack of Transparency

Nineteen respondents included it among their top three concerns, with eleven of those specifically using the word "transparency" in their response.

All of these responses conveyed general concern about the lack of transparency (or 'explainability') for AI systems and decision-making processes, with one respondent specifically citing the "fake transparency that companies offers in regard to personal data". As one response noted, "for the vast majority of people, these issues are unknown and difficult to understand" and therefore the general population does not understand how decisions using AI and data systems are made. As a result, there is no opportunity to appeal these decisions and it compromises individuals' ability to "exercise their rights to challenge decisions".

Furthermore, this lack of understanding "may leave errors undiscovered and lead to degeneration of the software code and related services in the long term", and it "makes true accountability impossible".

Respondents indicated that transparency should not only concern decision-making (i.e. how the data is used); transparency is also needed on the provenience and sources of data.

One respondent noted potential reasons for a lack of transparency may include "technical reasons (the functioning of complex AI systems cannot always be understood by humans) or due to structural and legal reasons (business secrets by companies building the systems, finger-pointing and lack of clear responsibilities)."

Greater transparency would enable understanding, scrutiny, and audits of the AI and data systems. One respondent suggested "AI research should aim to produce 'glass boxes' ". Another framed transparency as a "right to an explanation of how a decision affecting an individual was reached", while another proposed developing "interfaces that ordinary users can understand that explain why and how a system has made as decision".

Lack of Privacy

Seventeen respondents included it among their top three concerns, with fifteen of those specifically using the word “privacy” in their response.

Given GDPR, it is not surprising that privacy concerns rank among the top in experts and users mind; one respondent simply said that “the case of right to privacy is quite obvious”.

Respondents noted that the monitoring of individuals and collection of vast amounts of data, some potentially sensitive, is constant. Specific collection means cited were: web-tracking; IoT devices; geolocation; and real-time surveillance, including facial recognition. Those collecting data include state and private actors; one respondent noted the U.S. National Security Agency’s PRISM program and the joint NSO-Google X-Keyscore system as examples. Yet, as one respondent said, this monitoring and collection happens “in in ways that people are not aware”.

While one respondent acknowledged potential commercial justification for collecting bulk data (i.e. “for service improvement and company needs”, including advertisement purposes), all responses expressed concern about negative consequences. The loss of privacy was characterized as being an undue interference, intrusion, or invasion that impacts individuals in unprecedented ways. Specific concerns included: impact on human rights; the use of data for nefarious activities (citing the PRISM and X-Keyscore programs); “exploitation of the data for surveillance/monitoring and prediction of individual behavior”; concentration of “new power over the data subjects” exerted by those deploying technologies; and potential impacts on already marginalized groups.

Multiple respondents suggested that existing legislative measures are inadequate.

One respondent proposed that individuals should have “the right to have their data fully anonymized before being stored and/or processed by any company” and that “no one should be re-identifiable from the data.” Another respondent proposed that individuals should have “the right to participate in society to roam around in public and private space without being tracked and traced, have your personal data stored and possible be scored and rated for various means.”

Bias and Discrimination

Seventeen respondents identified discrimination and bias as an important ethical or human rights issue. Specific concerns included built-in bias and discrimination, discriminatory and inaccurate decision-making, and the right to equality and fairness.

Respondents expressed concern about existing built-in bias and discrimination, as well as entrenching bias and discrimination that reproduced by the algorithm. In both cases, “The model will be less accurate in making predictions” even if it appears as though the AI systems are producing accurate results.

This raises concerns about the “right to equal treatment regardless of race, gender, religion, etc” and questions about how the AI systems are “limiting opportunities or penalising people based upon data (e.g. health, ethnicity, gender)”. According to one respondent, “this is particularly alarming when the systems are being used to make decisions that have fundamental impacts on people's lives (e.g. in areas such as education, security or the health sector)”.

Loss of Human Decision-Making

Twelve respondents identified protecting human decision-making as an important ethical or human rights issue. These respondents expressed concern about the impact of AI systems on humanity and the value of human decision-making.

Multiple respondents articulated the concern that humans will be taken "out of the loop" on critical decision-making, resulting in "the privileging of machine logic over human intelligence" and "unpersonal" decisions being made. Furthermore, one respondent suggested that human would "stop thinking about what is reasonable" because trust is put into the AI systems, which would "depersonalize" humans, while another argued that "the impression is generated mankind is less 'smart' then machine and machines are superior in decision-making".

Four respondents articulated the concern as a violation of the right to autonomy. One respondent referenced micro-targeting in commercial and political advertisements, where many people "are generally not aware of the extent to which they are profiled" and feel a sense of "powerlessness" because they did not realize the extent to which their autonomy was limited. Another respondent proposed ensuring that "sufficient safeguards [are in place] to ensure people to maintain desired levels of autonomy", including protecting "space for people to make mistakes or even collectively undesirable choices".

Control and Misuse of Data

Ten respondents identified the control and use of data an important ethical or human rights issue. Specific concerns included the misuse, control, ownership, and commericialisation of data.

Six respondents identified misuse of data as an important issue. Three respondents specifically cited mass surveillance; one respondent also cited profiling based on data not shared for that purpose. Two respondents specifically cited exploitation and manipulation of individuals. One respondent referenced Cambridge Analytics as an example.

Two respondents specifically identified the right to control personal data an important issue.
Two respondents specifically identified the ownership of data as an important issue, with one respondent arguing that "some important personal data has to stay in my possession to prevent Identity Fraud and other negative consequences". One respondent specifically identified the commercialization of private information as an important issue.

Lack of Accountability and Liability

Nine respondents identified the accountability and liability of AI as an important ethical or human rights issue, specifically the current lack of accountability. Respondents called for a clear definition of legal responsibilities from planning to implementation for all actors.

Predictive and Non-Individualized Decision-Making

Five respondents identified predictive non-individualized decision as an important ethical or human rights issue. One respondent framed this concern as a "the right to be judged as an individual, while another made a connection to the presumption of innocence when individuals are unable to disprove a certain label. One respondent raised concerns about the allocation of resources or roles to individuals based.

Concentration of Power

Five respondents identified concentration of power as an important ethical or human rights issue. One respondent described the concentration of power by digital platforms as "rapid and accelerating", while another responded expressed concern about the potential for "exploitation" by companies "legally acting as above the law".

Lack of Access to and Freedom of Information

Four respondents identified the quality and freedom of information as an important ethical or human rights issue. One respondent called for the "right to be justly informed" to prevent the **"**dissemination of overtly fake information to influence opinions". All three other respondents expressed concern about the manipulation of information, fake news, and propaganda.

Violation of Fundamental Human Rights

Four respondents identified fundamental human rights as an important ethical or human rights issue. Expressing concern about limits to human freedom, one respondent cited to Article 1 (born free and equal in dignity and rights), Article 3 (right to life, liberty and security), and Article 4 (prohibition on slavery). Another respondent cited freedom of expression and association.

Lack of Quality Data

Three respondents identified the quality of data as an important ethical or human rights issue. As one respondent noted, unreliable data leads to wrong conclusions.

Disappearance of Jobs

Three respondents identified the disappearance of jobs as an important ethical or human rights issue. Two respondents framed this concern within the "right to work", calling for urgent "focus on raising the education of the population globally"; otherwise, "an enormous number of people will be replaceable by machines and will be replaced, making poverty and scarcity commonplace."

Prioritization of the "Wrong" Problems

Three respondents identified whether we are prioritizing the "wrong" problems as an important ethical or human rights issue. One respondent noted that "very few ethical concerns" are taking into account, and another asked whether technology is solving the "big problems" as identified by the SDGs. Another respondent noted with concern technology development is "overwhelmingly technology-driven."

"Awakening" of AI

Two respondents identified 'awakening' of AI as an important ethical or human rights issue. In the immediate future, one respondent asked if we are "able to control AI, especially in the case of machine/deep learning?". Looking into the long-term future, another respondent raised concern about the potential "intelligence explosion" that presents "new unsolved questions of e.g. machine intention and human-machine-societies and the coexistens or symbiosis/of fusion of both actors".

Security

Two respondents identified security as an important ethical or human rights issues. However, neither specified or defined what type of security is at risk.

Lack of Access to Public Services

Two respondents identified lack of access to public services as an important ethical or human rights issue. Both respondents suggested that AI is a "public (strategic) good" that should be developed by government or regional organization.

Harm to Physical Integrity

One respondent identified risks to physical integrity as an important ethical or human rights issue, citing autonomous weapons systems, autonomous cars and robots providing care for humans, and biotech developments.

Cost to Innovation

One respondent identified balancing innovation as an important ethical or human rights issue, particularly balancing the costs to innovation with safeguarding of rights.

Unintended, Unforeseeable Adverse Impacts

One respondent raised the concern of unpredictable and "potential dangerous AI outcomes," described by Nick Bostrom as "perverse instantiations," as an important ethical or human rights issue

Lack of Power to Frame Dialogue

One respondent raised concern about framing the dialogue around ethical and human rights issues, asking who decides "what constitutes an issue." Using an example of an issue related to climate change being framed for a "poor country" by outsiders, this respondent identified a very specific type of power-imbalance.

Q2: Which current approaches methods, or tools for addressing these issues are you aware of? These may be organisational, regulatory, technical or other.

There were 36 responses to Q2. Two responses were deemed more relevant to another question. Therefore, a total of 34 responses were analysed under Q2.

While many respondents provided many specific examples, one respondent expressed concern that they "hardly see any approaches / methods / tools addressing" ethical and human rights issues. Three respondents noted that the issues are becoming more well-know, with one respondent citing the Cambridge Analytica "abuse scandal." One respondent noted that some companies have an "increasing knowledge" that "data quality is a major factor in decision making." However, because of "lack of joint principles and coordination", the fact that much information is "not know and open to the public", and the "large distance between practice and theory" improved measures are needed.

Regulatory Measures

Regulations
Eighteen responses referred to regulations. In presenting specific examples, many responses reflect on the positives and negatives of individual measures, presented below.

Without providing specifics, one respondent critiqued existing regulatory provisions as being "intrusive, far-reaching and not equipped to address the challenges posed by the newest technologies," "against human rights," and having "potential and actual impacts which are not ethical".

Regulation of facial recognition and the 'platform economy' were mentioned, but no specific laws cited. One respondent noted that current legislation only partly address ADM by default, but cited no specific laws. Another respondent referenced procurement process and the probation of the use of technology in certain area, but also cited no specific laws.

Two responses referred to regulation at the international level. One respondent asserted that "addressing these issues is the obligation of [regional] or worldwide legislation". However, no respondent identified a binding international law instrument; as one respondent explained: "More and bigger international approaches aren't known to me, for the international community was not able to agree on deeper regulations".

Thirteen responses referred to regulations at the regional level, all of which are in the European Union (E.U.). One respondent asserted that "addressing these issues is the obligation of an EU or worldwide legislation", while another noted "a clear need for a European enactment" to address AI. One respondent expressed concern that "much of this legislation has yet to be tested in the courts".

Unsurprisingly, the E.U. General Data Protection Regulation (GDPR) was the most frequently cited regulatory measure, with ten specific references. Positively, the GDPR was described as "the main regulatory instrument in this area", "a step forward in the right direction", and "one of the best tools".

Specific strengths mentioned were its foundation in "fundamental rights and its provisions for ensuring fairness, transparency, lack of discrimination, and the fact that it is "suffient [*sic*] enough to protect personal data, on this stage at least". At the same time, respondents generally noted limitations in GDPR's scope and implementation.

Other E.U. measures mentioned were: European Parliament's study for the creation of European Civil Law Rules in Robotics[28], and European Commission's proposed Regulation on Privacy and Electronic Communications[29]. Five responses referred to regulations at the national level. Only examples in Europe and the US were referenced. Without specifying where, one respondent noted "some governments have adopted or proposed regulation on government use of SIS."

In the United Kingdom, the following laws were referenced:

- 2000 **Regulation of Investigatory Powers Act (RIPA)**[30];
- 2014 **Data Retention and Investigatory Powers Act** (DRIPA)[31];
- 2015 **Serious Crime Act (SCA)**[32], *amending* 1990 **Computer Misuse Act (CMA)**[33]**:**
- 2016 **Investigatory Powers Act (IPA)**[34] the so-called 'Snooper's Charter'; and
- 2018 **Data Protection Act**,[35] which the respondent believed "does not go far enough and … there is current a lack of legal safeguards, especially those which address most recent technological innovations".

In the Netherlands, two examples were cited:

[28] http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf
[29]https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications
[30] http://www.legislation.gov.uk/ukpga/2000/23/contents
[31] http://www.legislation.gov.uk/ukpga/2014/27/contents/enacted/data.htm
[32] http://www.legislation.gov.uk/ukpga/2015/9/contents/enacted
[33] http://www.legislation.gov.uk/ukpga/1990/18/contents
[34] http://www.legislation.gov.uk/ukpga/2016/25/contents/enacted
[35] http://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga_20180012_en.pdf

- **Dutch Open Government Law**, requiring the government to publish open data[36], and
- *possible* ethical guidelines for the use of AI in the financial sector, which could be implemented by the national bank in addition to existing banking regulations[37].

In the United States, three laws were referenced (one state and two federal):

- 2018 **California Consumer Privacy Act (CCPA)**[38];
- 1986 **Computer Fraud and Abuse Act (CFAA)**[39]; and
- *proposed* **Active Cyber Defense Certainty Act (ACDC)**[40].

Public Register of Permissions to Use Data

One respondent referenced "public registers of all ADM systems used in highly sensitive areas," but did not provide a specific reference to an existing public register.

Reporting Guidelines

One respondent cited a proposed law in the Netherlands "requiring government bodies to report on the use of SIS in a dedicated register".

Monitoring Mechanism

Two respondents referenced monitoring mechanisms and "oversight bodies." One respondent cited discussions in the Netherlands to set up an "algorithm authority that would be charged with maintaining oversight over the use of AI."

Technical Measures

Six respondents cited a range technical measures that can be adopted internally to address concerns, particularly transparency and bias. One respondent cited the ACM FAT Conference,[41] where technical approaches to fairness, accountability, and transparency are explored. One respondent noted that these technical measures could be combined with (a.k.a required by) law.

Testing Algorithms on Diverse Subsets

One respondent cited testing algorithms on diverse subsets.

Using Analytics Systems to Judge Whether Decisions Are Equal/Fair

One respondent cited using analytics systems to judge whether decisions are equal.

Generative Adversarial Networks and Other Techniques for Deriving Explanations from Outcomes

One respondent cited generative adversarial networks and other techniques for deriving explanations from outcomes.

More Open Data

Two respondents cited the availability of open data. One respondent argued that it is important because "lack of data strangles innovation and can push organisations into opaque practices to gather data." The

[36] https://business.gov.nl/regulation/freedom-of-information/

[37] https://www.dnb.nl/en/news/news-and-archive/DNBulletin2019/dnb385020.jsp

[38] http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4.&chapter=&article=

[39] https://www.law.cornell.edu/uscode/text/18/1030

[40] https://www.congress.gov/bill/116th-congress/house-bill/3270

[41] https://fatconference.org

other respondent also argued for "making data available for social innovation" and "increasing the accessibility of datasets representing the diversity in society.

Other Measures

International Principles and Frameworks
Three respondents cited international frameworks developed by international organisations. The specific frameworks cited were:

- OECD Principles on Artificial Intelligence[42] ; and the
- G20 Artificial Intelligence Principles[43].

High-Level Expert Groups
Six respondents cited expert groups and initiatives created by or within existing international organizations. The following examples were given:
Centre for Artificial Intelligence and Robotics[44] at the United Nations Interregional Crime and Justice Research Institute (UNICRI);

- United Nations Educational, Scientific and Cultural Organization (UNESCO) work on 'artificial intelligence with human values for sustainable development'[45];
- Global AI Council and Centre for the Fourth Industrial Revolution at the World Economic Forum (24)[46] to facilitate sharing of best practices between States;
- United Nations' AI for Good Global Summit;[47] and
- EU High-Level Expert Group on AI[48], which has produced *Ethics Guidelines for Trustworthy Artificial Intelligence* [49] and *Policy and Investment Recommendations for Trustworthy Artificial Intelligence*[50].

Grievance Mechanism
One respondent cited complaint mechanisms related to "transparency over the functioning of ADM systems."

Frameworks, Guidelines, and Toolkits
Fourteen respondents cited frameworks, guidelines, and toolkits. The specific examples provided have been developed by national governments, industry organisations and private companies, and NGOs and civil society organisations (including academia). Many were developed through partnerships between entities. These tools include "sets of ethical questions that developers, policy makers and AI specialists should ask themselves constantly during the development, building, implementation and every relation processes in order to keep an ethical focus". The tools may take the form of ethical codes, guidelines, audits, risk management strategies, and impact assessments. One respondent cited a report by ETH

[42] https://www.oecd.org/going-digital/ai/principles/
[43] https://www.mofa.go.jp/files/000486596.pdf
[44] http://www.unicri.it/topics/ai_robotics/
[45] https://en.unesco.org/artificial-intelligence; Beijing Consensus on Artificial Intelligence and Education (https://unesdoc.unesco.org/ark:/48223/pf0000368303; Preliminary Study on the Ethics of Artificial Intelligence (https://unesdoc.unesco.org/ark:/48223/pf0000367823).
[46] https://www.weforum.org/centre-for-the-fourth-industrial-revolution/; Framework for Developing a National Artificial Intelligence Strategy https://www.weforum.org/whitepapers/a-framework-for-developing-a-national-artificial-intelligence-strategy
[47] https://aiforgood.itu.int/about-us/
[48] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
[49] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top
[50] https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Zurich[51] that found there are 84 projects and organisations working on AI issues. One respondent noted generally "the various code and principle guidelines developed by the EU," but did not cite any specific examples. Two respondents specifically referenced impact assessments. One respondent specifically referenced risk management strategies "to identify and address human rights ethical concerns."

Frameworks at the national government level include:

- UK Data Ethics Framework,[52] which consists of seven principles and an "accompanying workbook to be used when starting new projects";
- *draft* UK Guidelines for AI Procurement[53];
- Ethical guidelines for research developed by the Norwegian National Research Ethics Committees[54];
- Ethical Accountability Framework for Hong Kong, China,[55] prepared for the Office of the Privacy Commissioner for Personal Data and including a model ethical impact assessment[56]; and
- U.S. Menlo Report: Ethical Principles Guiding Information and Communication Technology Research[57] and companion guide[58] for the Department of Homeland Security.

Guidelines and toolkits developed by or in partnership with private companies include:

- One respondent cited the AI Fairness 360 Open Source Toolkit,[59] developed by IBM, as a tool to address data bias.
- One respondent cited Watson Openscale,[60] developed by IBM, as a transparency tool.
- One respondent cited the People + AI Guidebook,[61] developed at Google.
- Two respondents cited Explainable AI,[62] developed by Google.
- One respondent cited Machine Learning Fairness work at Google[63]
- One respondent cited the work at The Partnership for AI[64]
- One respondent cited the work at The Institute of Business Ethics[65]

Three respondents cited the Data Ethics Decision Aid (DEDA),[66] a tool develop at the Utrecht University in The Netherlands. One respondent described it as "a game that makes you think about the concequences [*sic*] of the AI project you're want to start" and "a practical tool that helps you ask questions and become aware of ethical aspects of a data project." One respondent cited the Understanding Artificial Intelligence Ethics and Safety[67] tool, developed in the U.K. by the Alan Turing Institute to provide "a detailed approach to incorporating an ethical approach into the design of AI systems in the public sector". One respondent

[51] Anna Jobin, 'Ethics guidelines galore for AI – so now what?', ETH Zürich, 17 January 2020, https://ethz.ch/en/news-and-events/eth-news/news/2020/01/ethics-guidelines-galore-for-ai.html
[52] https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework
[53] https://www.gov.uk/government/publications/draft-guidelines-for-ai-procurement/draft-guidelines-for-ai-procurement
[54] https://www.etikkom.no/en/ethical-guidelines-for-research/
[55] https://www.pcpd.org.hk/misc/files/Ethical_Accountability_Framework.pdf
[56] http://informationaccountability.org/wp-content/uploads/Model-Ethical-Data-Impact-Assessment-January-2019-002.pdf
[57] https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf
[58] https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCOMPANION-20120103-r731_1.pdf
[59] https://aif360.mybluemix.net
[60] https://www.ibm.com/cloud/watson-openscale
[61] https://pair.withgoogle.com/about/
[62] https://cloud.google.com/explainable-ai/
[63] https://developers.google.com/machine-learning/fairness-overview
[64] https://www.partnershiponai.org
[65] Institute of Business Ethics and Investment Managers
[66] https://dataschool.nl/deda/?lang=en
[67] https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety

cited the AI4People's Ethical Framework for a Good AI Society,[68] developed by the Atonium-European Institute for Science, Media and Democracy (EISMD). One respondent cited the Framework for Practical Ethics,[69] by Daniël Tijink and Peter Paul Verbeek. One respondent cited the work of Luciano Floridi and Josh Cowls to develop five principles for AI in Society[70] (building on Beauchamp and Childress' four core principles of bioethics). One respondent cited the work of Virginia Dignum at Delft University of Technology, advocating for the application of accountability, responsibility and transparency (ART) design principles to AI.[71]

Codes of Conduct
Three respondents referenced codes of conduct. Two respondents cited EU ethical codes, but did not provide a specific example. Another respondent noted they observe "a lot of activity on developing codes of ethics and principles for responsible use of analytics and artificial intelligence … However translating these often more abstract principles to more concrete modus operandi is where the challenge lies." The third respondent referenced "professional ethics codes for people developing and using ADM systems," but did not provide any specific examples.

Education Campaigns
Four respondents reference educational campaigns for actors at all ages and across fields. One respondent referenced "programs for data literacy." Two specific programs were cited that educate the pubic on "the power of AI":

- Finland Elements of AI free online course[72];
- Dutch Nationale AI Cursus (National AI Course)[73], a free online course.

Stakeholder Dialogue and Scrutiny
Three respondents referenced measures to increase diversity and active participation from stakeholders, which helps "fight discriminatory structures in tech organizations" and makes "people from underrepresented groups in the field more visible".

Employing 'Fairness' Officer or Ethics Board
Three respondents cited "employing a 'fairness' officer or ethics board" or the "intervention of ethics committees and DPOs". One respondent noted "they are not perfect tools," though they are "nice." One respondent argued that "human revision should be included at several stages," so long as the "constant checking and monitoring" avoid bottle-necks in the development process.

Policy Commitment
One respondent referred public self-commitments by organisations developing and employing relevant technologies.

Self-Regulation
One respondent cited one example of self-regulation: Twitter's self-imposed ban on political ads.
Third-Party Testing and External Audits Two respondents cited the use of various audits "of data used for training, bias, explainability." One respondent specifically cited 'bias testing' of data sets and ADM by neutral oversight bodies.

[68] https://www.eismd.eu/ai4people-ethical-framework/
[69] https://ecp.nl/wp-content/uploads/2019/11/060-001-Boek-Aanpak-begeleidingsethiek-240165-binnenwerk-digitaal.pdf
[70] https://hdsr.mitpress.mit.edu/pub/l0jsh9d1
[71] https://www.itu.int/en/journal/001/Documents/itu2017-1.pdf
[72] https://www.elementsofai.com
[73] https://www.ai-cursus.nl

Standardisation
Three respondents cited standardisation, specially the ISO[74] and IEEE (e.g. Handbook on Ethically Aligned Design[75] and the P7000 series[76] of standards for addressing ethical concerns during system design).

NGO Coalitions
One respondent cited th**e** Campaign to Stop Killer Robots,[77] an NGO coalition "working to ban fully autonomous weapons and thereby retain meaningful human control over the use of force."

Open Letters
One respondent cited the 2015 Open Letter on Artificial Intelligence,[78] signed today by over 8,000 people (including Stephen Hawkins and Elon Musk), which "urges the international community to take regulatory actions".

Investigative Journalism
Three respondents mentioned the role of media and investigative journalism, noting that publication of stories can "unveil examples of violations" and "hold governments and companies to account". One respondent specifically noted that writing articles was a individual initiative they could personal undertake.

Individual Action
Two respondents identified measures that can be undertaken at the individual level. One included writing article and participating in conferences to raise awareness about "the ethical dilemmas". The other respondent referenced the use of software and tools that prevent tracking (e.g. disabling ads or using alternative search engines) as an approach to avoid being individually monitored.

Q3: What do you think are the pros and cons of these current approaches, methods, or tools?

There were 31 responses to Q3. Three responses to Q3 were deemed more relevant to another question. Therefore, a total of 28 responses were analysed under Q3.

There were far more cons mentioned than pros. Only thirteen respondents identified a 'pro', focusing on specific types of current measures. In contrast, nearly half of respondents identified at least one 'con' of existing measures, giving general critiques and critiques specific to individul types of measures.

Pros of Current Measures

Only 13 respondents identified a 'pro' of current approaches, methods, and tools. Almost all of the responses mentioned a particular type of measure, but in the general sense; only two mentioned a specific example of a current approach (the GDPR and CCPA). As one respondent succinctly put it, the pros were "various depending on approach", while another summarized the pros of these measures with "it should mitigate the risks".

Benefits of Dialogue
In a general sense, two respondents believed that any current dialogue about these issues is positive: "the fact that [frameworks] exist and are publicly available is to be welcomed" (4) and "it is good to talk with

[74] https://www.iso.org/committee/6794475.html
[75] https://ethicsinaction.ieee.org
[76] https://standards.ieee.org/project/7000.html
[77] https://www.stopkillerrobots.org/about/
[78] https://futureoflife.org/ai-open-letter/?cn-reloaded=1

each other and to learn from each other". Within the scientific community, one response described the dialogue as "agile".

Benefits of Legislation

Four responses addressed legislation. One responded referred to legislation as "the most powerful instrument for changing practice across the board," noting that it is the law "forcing controllers to make DPIAs**"**, not "a recognition that it is good practice". Another respondent noted "legislative approaches have the power of the state behind them, and a 'machine' to enforce them". Two respondents cited the GDPR and one the CCPA specifically, describing them as "effective in the regions [where] they apply" and "useful in theory".

Benefits of Internal Efforts for Transparency

Four responses dealt generally with approaches, methods and tools utilized by developers to increase transparency, claiming they would be "be effective" and "have few downsides". Another respondent stated organisational and technical measures, if done well, would "build ethical quality into the design". From the user perspective, one response argued that transparency "is necessary … to give those affected the possibility to complain against mistakes made and to denounce human rights violations". The same respondent noted "public self-commitments can help raise awareness on the importance of ethical questions" if they are include specific commitments and lead to organizational change.

Benefits of Education and Awareness Campaigns

Two responses mentioned education and awareness. One respondent wrote that "societal measures" are "probably one of the few ways in which to enhance citizen/consumer power". Focused more on the awareness within industry, another respondent noted "the increasing knowledge in companies that data quality is a major factor in decision making", and that the use and quality of data should undergo thorough analysis.

Benefits of Ethical Impact Assessments

Writing about ethical impact assessments (EIA), one respondent referenced their "place in operationalising the general principles," emphasizing that EIAs give "practitioners a clear methodology and tools".

Benefits of Standardisation

One response addressed standardisation specifically, characterizing it as "potentially a powerful tool" because "it provides an objective set of criteria and an established mechanism of assurance and certification".

Benefits of Oversight Mechanisms

In reference to oversight mechanisms, specifically in regards to ADM, one respondent wrote they can be "indispensable in certain areas and overall very effective for addressing potential human rights violations".

Potential Benefit of Ethical Approach: Competitive Advantage

One respondent identified a particular long-term benefit for those developing and/or employing new technologies with an ethical perspective: a competitive advantage. By "involving a diversity of stakeholders from the outset and educating them regarding all aspects is a guarantee for the take up of technologies," developers can help prevent against "products or services' failure due to unethical aspects".

Potential Benefit of Ethical Approach: 'Fresh Eye' to Solve Problems & Better Health

A couple responses focused on the positive benefits of specific technologies or applications, rather than the current approaches, methods, and tools to address identified issues. One respondent wrote that evidence-based analytics systems "can bring a fresh eye" to solving problems. Another respondent "welcomed" applications in healthcare that "will help society for a better health".

Cons of Current Measures

Twenty-five respondents identified a 'con' of current approaches, methods, and tools.

Some responses referred to 'cons' in the general sense, without specifying which type of measure was being critiqued. Therefore, the first section below contains a list of common 'cons'.

Other responses addressed particular types of measures, sometimes also providing a specific example.

Con: Lack of Understanding
Four responses dealt with a lack of understanding about the issues and the role of various actors. As one respondent noted "All of the approaches require that the important parties feel ethical responsibilities". On the developer side, respondents wrote that "not all companies have it firmly on the radar", and that there is "still a lack of recognition from companies … of the far-reaching effects of their conduct on democracies, economies, beyond the short-term economic goals". Furthermore, "how to implement of ethics is not always clear to companies", and there is a "lack of using the already available tools". On the other end, one respondent pointed out that "an apathetic set of consumers is not helpful in moving controls forward quickly".

Con: Risk of Shifting Burden of Responsibility
Two responses discussed the risks associated with ill-conceived measures to assign responsibility. One respondent expressed concern that governments will shift the burden of "discovering, prosecuting and penalizing unethical AI to the enterprises that offer the technology", while another respondent argued that the burden should not be shifted "towards those affected by the systems".

Con: Too Abstract
Four responses critiqued many existing approaches, methods, and tools for being "too abstract", "often theoretical", or not "tangible". More specifically, one respondent claimed that, while there may be many initiatives, there are no wrap-ups or joint conclusions.

Con: Resource Intensive
Four responses addresses the resources needed to develop and actualize new approaches, methods, and tools. One respondent wrote "funding and support … could be larger, given the massive impact of this problem". Three respondents focused on the costs, noting "these approaches are all time [and] money consuming" and characterizing the process as "sluggish".

Con: No Enforcement
Three respondents claimed, in the general sense, that existing measures "lack enforcing mechanisms", are "optional [and] not legally binding" and not "mandatory".

Con: No Comprehensive Approach
Two respondents critiqued the existing system of measures as lacking a "systemic approach", as they are "too often deployed in isolation". Calling for a 'smart mix' of instruments, this respondent argued "human rights and ethical issues cannot be addressed by legislation or risk management or technical measures by themselves".

Con: Too Complicated
One respondent seemed to dismiss the effectiveness of any approach, writing that its "too complicated" and "almost impossible" to implement "a new way of thinking and production in a conservative bureaucracy that's risk avoiding orientated".

Con: No Requirement to Justify Data Collection
One respondent critiqued the fact that "principle of minimization of the data collected is hardly respected", meaning that companies do not have to justify their data collection purposes.

Cons of Regulation
In a general sense, one respondent claimed that there are presently "insufficient incentives and regulations".

Five respondents focused on the limited scope of application of specific regulations, which "may be either too broad or too specific". Three illustrative examples were cited. The first example is the fact that domestic or regional regulation only applies within its jurisdiction, and therefore does not apply "outside the EU in countries which may not appreciate ethical or human rights as highly as the EU does" or in "authoritarian regimes such as China … [which can] export it to other authoritarian countries without any such limitations". The second example was exceptions granted to intelligence services like the U.S. National Security Council, U.K. Government Communications Headquarters, and the 5/9/14 Eyes. One respondent claimed that privacy legislation is "undermined" by these exceptions because intelligence services' access is "virtually unlimited and unchecked"; another respondent noted generally "the profound implications for privacy and human rights". The third example is the Dutch Open Government Law, which only applies to "data that was supposed to be public in the first place" and thus does necessarily bind private companies. Closely related, one respondent noted that large business with "more power" are "good at evading jurisdictions".

Three respondents noted how technology development outpaces the rule-making process, resulting in policy that "falls behind the cutting edge developments", "probably be out of date in a very few years" and "likely behind the developments as measures tend to follow after a problem has occurred". One respondent also referred to the fact that "legislation is slow to create and change, [and] relatively inflexible" as part of the problem.

Two respondents claimed "regulatory actions lack an enforcement by the political leaders and institutions". Writing specifically about the GDPR, one respondent claimed "the legislation is being overtaken by reality and it is not actually being enforced by the personal data authority", and as a result "the law is taken less seriously and economic interests of data is once again the main objective".

One respondent noted that legislation "requires case law to bring out its real application", which limits immediate value.

One respondent expressed concern that legislation can "generate a compliance-only setting, where the letter but not the spirit of the law is observed". One respondent argued that the weakness of regulatory measures, in particular the GDPR, is that they are "based on the idea that people are aware of the importance of data", which is not true. Therefore, users "just 'tick the box' to gain access to a service, without knowing what is written on the terms and conditions". One respondent expressed concern that existing legislation does not address the disappearance middle-class jobs as a result of new technologies. One respondent expressed concern that "new legislation may also hamper innovation".

Cons of Ethical Guidance, Frameworks, and Impact Assessments

Three respondents provided numerous critiques of guidance and assessment measures that could be used during the development process.

One, these measures "may presuppose a level of prior knowledge" that some organisations, including small-medium sized enterprise may lack. Referring specifically to DEDA, one respondent expressed concern that we "need more fundamental tools...people need to get educated so they will understand it" before the tools can be effective. Two, the respondent claimed the processes are "overly long and feel a bit unwieldly for practical use". Third is the risk that organisations will come to view the process as "yet another" assessment or "more as a hurdle that has to be tackled", and as a result the organization will "only pay lip service to them and they will fall out of use". "This could be a problem from both the committees, as well as the people submitting to committees". Lastly, the respondent noted that existing measures are "generally quite public-sector focused", meaning that private and third sector organisations **"**may be underserved despite potentially processing quite large amounts of personal data in novel ways while quite young" at a time when "they are not subject to public law remedies".

Cons of Standardisation
Two respondents discussed the cons of developing and implementing standardisation. Both noted the complexity of agreeing on standards given the subject matter; the "ethical impact of a system is by its nature intangible and difficult to measure objectively**"**. The specific example of assessing "risks posed by third generation deep learning neural networks" was cited. Two other limitations of standardisation were that it "takes time to agree and put in place" standards and "certification and technical measures are costly".

Cons of Transparency Measures
One respondent wrote that transparency measures, particularly techniques for deriving explanation, are limited in application because they "may only apply in a restricted set of cases". Another respondent noted that these transparency measures "need legislative changes that would make them mandatory in certain areas".

Cons of Awareness and Education Campaigns
One respondent claimed that the efficacy "is often low and may have issues reaching the people who need it most".

Cons of Oversight Mechanisms
One respondent noted that oversight mechanism are subject to criticism "by those developing and deploying the systems because they fear breaches of business secrets and public backlash".

Cons of Public Statements
One respondent addressed the risk that public statements "remain vague and do not lead to any organizational change while drawing away attention from existing problems (ethics washing)".

Cons of Particular Applications
Three responses focused on the cons of specific technologies or applications, rather than the current approaches, methods, and tools to address identified issues. The first noted that analytics systems can "bring their own bias" and it is not clear who defines the framework (e.g. who defines 'equal' treatment?). The second raised concern about bias in applications in education and industry, and claimed that "applications in Defense are very dangerous for all people". The third asserted that "automatization will happen", which will have an impact on human decision-making. For example, in health care, "if an algorithm recommends a concrete treatment, a physician will hardly contradict".

Q4: What would you propose to address such issues better?

Thirty respondents provided an answer in Question 4. An additional four responses to other questions were considered more relevant to Question 4. Therefore, there were a total of 34 responses analyzed under Question 4.

Formal legal measures were the most frequently proposed, with regulations being the most common. Additionally, many respondents proposed technical measures, as well as other measures such as frameworks and guidelines, educational and awareness campaigns, and indivudal action.

In a general sense, one respondent called for concerted thinking about "public value management" and "vision about the society you want to create." Another respondent cautioned against "the assumption that things like the conceptual list of 'issues' can be known in advance."

Regulatory Measures

Regulations
Thirteen respondents proposed some form of regulation, though one respondent added that regulation alone is not enough. Generally speaking, one respondent proposed "no hasty ill-thought legislation", while another called for "getting FAST legislation in place". One respondent called for a 'smart mix' of regulatory initiatives; "for example, legislation may require identification of (most salient) risks not only in the operations of the market actor itself but also in value chains, a requirement to address these risks and to remedy if things have gone wrong". Two respondents referenced the GDPR; one called for applying the "principles on big data and its secondary use and informed consent, while the other proposed the adoption of GDPR-like privacy in new jurisdictions. In addressing privacy, this respondent emphasized the need for more attention on "the data available to ad-tech, e.g., browsing and location histories". One respondent called for regulation on the European Commission-level. One respondent suggesting classifying technologies "based on risk clusters" and regulating requirements accordingly. One respondent suggested modeling regulation after law for food and drug safety (like the FDA in the US) to "ensure that the AI algorithms that we use in our daily lives do not impact our fundamental human rights". One respondent called for legislation for 'transparent AI". One respondent called for legislation requiring companies "to prove that collection [of data] is necessary for the service provided, not for any other reason," and this obligation should be monitored by the State. One respondent called for "removing the right to store nominative data." Two respondents called for requiring "anonymization of the data for storing/processing." Two respondents called for legislation to determine the legal liability of AI, and "how the liability is shared/displaced to humans interacting with it (trainer, creators, users,...)". One respondent called for legal recognition of "the right to work (even imperfectly), earn a living, and have meaning by working, as a basic human need that cannot be alienated by AI". One respondent emphasized that regulation should allow "human beings to make their own decisions and by all means avoid putting human beings under the tutelage of machines".

Public Register of Permissions to Use Data
One respondent proposed setting up a register to manage permissions to use individuals' data. "If your permission is not on the register, the company that uses your data is in violation with the GDPR."

Reporting Guidelines
One respondent proposed imposing "reporting guidelines for all listed tech companies", modeled on "current best practices for corporate responsibility reporting as described by GRI."

Monitoring Mechanism
Two respondents proposed monitoring of technology developments. One respondent called for monitoring "to prevent global acting companies from hopping on to more libertarian policy environments

of different states", and the other respondent argued that "strict and continuous monitoring" is needed when "big data is used in joint public-private efforts, e.g. medical research and big pharma or opinion analyses".

Technical Measures

Six responses proposed measures to be implemented within industry. As one respondent noted, measures to address issues of concern should be made "current practice" and there should be open channels for "tracking of discussions".

More Open Data
One respondent proposed making more open data available. After establishing a "consensus on what should be open data to everyone", that data can be used freely, while the rest of the data can only be used on a permission basis by a selective list of companies that have received permission in the register."

Use AI to Protect Data
One respondent proposed using AI itself "as a tool to help customers make decisions on access to their data."

Improve Control of Data
One respondent proposed improving data protection within organization, in part by ensuring that there is an entity mandated to deal with data protection. This respondent argued "at the moment there is way too much data to invest control at a single body."

Employ Algorithms That Can Be Explained
One respondent proposed not employing "algorithms that are too complex for meaningful explanations."

Create Comprehensive AI Example Sets
One respondent proposed creating more comprehensive examples sets of AI use for recommender systems.

Retaining Possibility of Human Override
One respondent proposed ensuring more opportunities for human override of AI decisions.

Other Measures

International or Regional Framework
Three respondents proposed developing an international agreement, "ethical development standards", or a "worldwide acceptable ethical framework". Two respondents noted the difficulties associated with this suggestion, as "sadly we are in an area where cooperation between countries is declining" and "it is difficult to be accepted from all nations, as the ethical approaches are different in each area". However, one respondent noted that a global framework "should be concentrated in common ethical values for the society".

High-Level Expert Group
One respondent proposed establishing "independent, multidisciplinary, multicultural bodies to provide the technology industry with independent definitions and evaluation criteria."

Grievance Mechanism

One respondent proposed developing grievance mechanisms, specifically to enable complaints on SIS. The respondent believes these mechanisms "may trigger enhanced (industry specific) risk management and certification in order to show compliance," as well as spurn the development of "technical measures to enable these risk assessments." It was not clear whether the respondent was referring to a government-based grievance mechanism or individualized mechanisms within companies, but the respondent did characterize the proposal as a "type of regulation."

Citizen Juries

One respondent proposed creating citizen juries that could "evaluate risk of various AI technologies and propose appropriate tools" by means of "a broad range of stakeholders including AI researchers and developers risk experts and policy specialists."

Frameworks, Guidelines, and Toolkits

Two respondents proposed integrating ethical guidelines and toolkits into current standard methodologies for project management and IT development. One respondent specifically suggested creating differentiated toolkits that are "layered i.e. the user can drill down into more detail at each stage as desirable". The other respondent proposed developing guidelines that take "into account current gender policies" and reflect "the need for diversity and inclusivity in the groups".

Codes of Conduct

One respondent proposed binding ethical codes of conduct that cover "planning to implementation." This respondent argued that "legislation is not enough" because big companies "apologise [and] pay the fine which they can afford and go on as before" when they violate the law.

Education Campaigns

Eleven responses proposed educational training and awareness campaigns. Educational training was recommendation for all levels, including children, students, developers and professional, politicians and government officials, and members of the public generally. One respondent argued that education must begin with "people understanding how they can be manipulated or deprived from privacy, with all negative consequences". Benefits of educational training cited included: citizens being able to "make informed decisions and demands to their governments", and able to "avoid bad practices…and reject unethical services"; politicians and decision-makers "understand[ing] what they need to legislate/regulate about" and being "aware of the possible dangers of AI and of their role to protect citizens from this"; and professionals "incorporate[ing] ethical considerations in their designs". One respondent wrote that education training should not been treated "as a token class that allows a checkbox to be filled", but that students must be educated "about ethical impact of their own research." Two respondents wrote about education in the context of the right to work and the disappearance of traditional jobs. One respondent proposed more education about the right to work as a basic human need. Another called for "good re-education options … for people whose jobs become obsolete" in "automation-proof areas".

Exchange of Best Practices

One respondent proposed "better knowledge exchange across sectors, disciplines, and countries regarding practical experiences with the use of Algorithmic Decision-Making (ADM) systems and measures taken to ensure their ethical development and use." This includes organizations developing and deploying SIS, as well as local and national governments.

Stakeholder Dialogue and Scrutiny

Five respondents proposed more dialogue with and public scrutiny from stakeholders. One respondent specifically called for "more diverse and deeply connected networks of AI scientists, programmers/developers and decision makers/politicians" working together. Another respondent noted that the "discourse and exchange of information between industry players and policy-makers, between

businesses and researchers should not be a cat-mouse game"; trust is needed on both sides. Another respondent advocated for "pushing for broad public debate on the issue outside of the technical and legal communities."

Employing 'Fairness' Officer or Ethics Board
Two respondents proposed supervisory boards or committees to help developers "think of all possible side effects at the start of an AI-project," including unintended side effects.

Ethical Mindset
One respondent proposed that adopting an "ethical mindset" should be the objective of AI providers and private companies.

3rd-Party Testing and External Audits
Two respondents proposed creating a system of third-party testing and/or external audits. In order for this proposal to be successful, one respondent argued "we need competence building among these organizations" that would conduct the audits. Additionally, this respondent noted it is "not enough to simply provide information"; "knowledge and time available of the decision subjects and the concrete settings need to be considered and findings from behavioural science should be taken into account."

Create Open Source Tools
One respondent proposed establishing interdisciplinary research projects to create open source "tools that allow for transparency, explainability, and bias mitigation."

Standardisation
One respondent called from some form of standardisation that requires "certain controls and addresses breaches"; the controls must "be able to be documented and demonstrable."

Individual Action
One respondent advocated that individuals should "have a choice" to "avoid AI." For example, individuals should be presented with simple processes "to refuse cookies."

Retaining 'Unsmart' Products and Services
One respondent called for keeping "the possibility of 'unsmart' products next to the smart ones."

Q5: Which should be the top 3 criteria for society to select and prioritise the most appropriate measures?

There were 31 responses to Q5.

About half of respondents seemed to misunderstand the question; instead of identifying criteria that should guide the development of measures, the respondents identifying criteria to guide the development of new technologies. As such, the two most frequently mentioned criteria – societal impact and transparency – are better characterized as criteria for the technology itself. It was interesting to note that more traditional criteria for evaluating measures – like costs, feasibility, and effectiveness – were mentioned only a few times.
Societal Impact

Twelve respondents identified the level and/or type of societal impact as a consideration criterion for appropriate measures. Different language used to describe this impact included: "the level of risk to society" ; "harm to users" ; "effect of data on personal lifes [*sic*]"; "impacts to society"; "what can be lost

in terms of public interest" and "who is benefitted"; and "impact on social cohesion". One respondent called for measures to "take into consideration any technological innovation which may be forced," citing examples including cashless systems and "robots degrading the job market." One respondent called for measures to be "human and planet centric". One respondent called for measures that favor "social / global advantages rather than individual benefits"; One respondent called for measured to be guided by the "value it has for civilians and never the value it has for the government or corporate" (aka public value management). One respondent called for measures with broad reach that "focus on systems with high impacts on people's lives (i.e. with a focus on the public sector)". One respondent called for measures that "think long term and global benefits (rather than short term and local)".

Transparency

Ten respondents identified the level of transparency as a consideration criterion for appropriate measures, with nine of those using that specific term. One response that didn't use the word 'transparency' called for "knowledge of AI decision making". Respondents called for transparency: in "what is being done and its effects"; "of the methods and tools that the companies are using" "on the possibility [and] technical progress"; and "when this control is not working or visibility of how the control is working". Additionally, one respondent proposed having the "ability to back to the source when possible". To help create transparency, one respondent notes that "future developers (in their student years) should be informed and taught on ethical issues more deeply and should be motivated to think more critically". One respondent advocated for prioritizing transparency over costs.

Respect for Human Rights

Five respondents identified level of respect for human rights as a consideration criterion for appropriate measures, with one respondent noting that measure should "ensure the human rights are not diluted" because "we are free human beings and we belong to ourselves".

Enforcement/Monitoring/Oversight

Five respondents identified enforceability and oversight as a consideration criterion for appropriate measures. One respondent referred specifically to regulatory frameworks, while another called for the "ability to ensure compliance with or without regulation". One respondent called for measures that can offer "fast protection for those affected", like "functioning oversight mechanism". One respondent called for a "method of oversight that does not depend on elected national politicians".

Impact on Minorities/Vulnerable Groups

Four respondents identified the level and/or type of impact on vulnerable individuals and groups, including minority communities, as a consideration criterion for appropriate measures. One respondent specifically advocated for the need to "prioritise the availability of access to public and some private services for the digitally-excluded, e.g., the poor, less educated, and those in rural areas with weak connectivity."

Degree to Which Human Decision-Making is Preserved

Four respondents identified the degree to which human-decision making is preserved as a consideration criterion for appropriate measures. Two respondents focused on retaining the possibility of human decision-making; one respondent focused on human influence in the decision-making process; and one respondent emphasized "not reducing human interactions".

Fairness

Three respondents identified the fairness as a consideration criterion for appropriate measures. One respondent posed the question: "Are the positive and negative effects distributed fairly?" (36), while another called for fairness "to as great a proportion of the human race as possible (preferably everyone)". If there is unfairness, "can the negative effects be compensated otherwise?".

Upholding Democratic Values
Three respondents identified the impact on democratic processes as a consideration criterion for appropriate measures.

Effectiveness
Three respondents identified effectiveness as a consideration criterion for appropriate measures. One respondent specifically referred to cost-effectiveness.

Feasibility of Implementing
Three respondents identified the feasibility (or practicality) of implementing as a consideration criterion for appropriate measures. One respondent argued "a measure is useless if it is impossible to execute in practice" and advocated for considering: "impact on the economy; effectiveness if adopted by a single region first; and effect on the competitiveness of a single country".

Non-Discriminatory
Two respondents identified the non-discrimination or inclusivity as a consideration criterion for appropriate measures.

Acceptability
Two respondents identified broad acceptability as a consideration criterion for appropriate measures. One respondent specified that measures must be met "with the fewest gut-level rejections", "move the most people to approval" and "attract the greatest expert support".

Impact on Innovation
Two respondents identified the impact on innovation as a consideration criterion for appropriate measures, which one respondent specifically calling for measures that support innovation.

Impact on Environment/Climate Change
Two respondents identified the impact on the environment as a consideration criterion for appropriate measures. One respondent specifically called for measures that promote technologies to stop climate change.

Reliability
One respondent identified reliability as a consideration criterion for appropriate measures, calling for measures that can be "sustained over time."

Proportionality
One respondent identified proportionality as a consideration criterion for appropriate measures, calling for measures that are tailored to the level of "risk of the processing to individuals' rights and well-being" and meaning that smaller organization may need a "lighter touch."

Flexibility
One respondent identified flexibility as a consideration criterion for appropriate measures, calling for measures "that are intended to have some longevity (eg legislation) are not tied to the specifics of current technologies."

Based on Data/Science
One respondent identified the degree to which there is a basis in science and/or data as a consideration criterion for appropriate measures, rejecting measures that are responses to "the, often flawed and biased, opinions of and prejudices of voter and interest groups."

Long-Term Impact
One respondent identified the long-term impact as a consideration criterion for appropriate measures.

Constructiveness
One respondent identified constructiveness as a consideration criterion for appropriate measures.

Trustworthiness
One respondent identified trustworthiness as a consideration criterion for appropriate measures, suggesting that measures "should be put up by politicians to secure a trustful relations in AI development."

Based on Multi-Disciplinary Stakeholder Dialogue
One respondent identified the need to have multi-disciplinary and international networks of stakeholders, including "lawyers, political actors, academics, and technical experts" to inform the creation of new measures

Specific Areas to Address

Responsibility/Liability
Two respondents called for measures to clearly establish the legal responsibility and liability of all actors involved, including AI systems. One respondent specifically noted "the use of technology does not absolve one of responsibility".

Determine Role for AI
One respondent called for measures to determine the role of AI and regulate "what AI will be able to do (i.e. where and how it should be used in place of humans) in the future in the job market."

Promote Privacy
Three respondents called for measures to promote privacy and personal data awareness. One respondent specifically proposed measures to "understand and promote the importance of crypto and data management".

Data Usage
Two respondents called for measured to address how data is used. One respondent proposed laws to regulate or restrict bulk data collection (e.g. "who has the right to collect and keep such personal information, and for what specific purpose, and for how long"). The other proposed measures to "prevent misuse of data for better profits".

Precautionary Principle
One respondent called for a measure to promote the precautionary principle.

Appendix E: Round 2 Responses

Question 1 (Ethical and Human Rights Issues) - Average scores (reach, significance, attention, and overall)

Ethical and Human Rights Issues	Reach	Significance	Attention	Average
Lack of Privacy (26 responses) Related to which type of data and how much data is collected, where from, and how it is used	4.19	3.85	3.85	**3.96**
Misuse of Personal Data (26 responses) Related to concerns over how SIS might use personal data (e.g. commercialization, mass surveillance)	4.27	4.38	3.50	**4.05**
Lack of Transparency (25 responses) Related to the public's need to know, understand, and inspect the mechanisms through which SIS make decisions and how those decisions affect individuals	3.85	3.73	3.04	**3.54**
Bias and Discrimination (25 responses) Related primarily to how sample sets are collected/chosen/involved in generating data and how data features are produced for AI models; and how decisions are made (e.g. resource distribution) according to the guidance arising out of the data	3.60	4.40	3.40	**3.80**
Unfairness (26 responses) Related to how data is collected and manipulated (ie. how it is used), also who has access to the data and what they might do with it as well as how resources (eg. Energy) might be distributed according to the guidance arising out of the data	3.81	3.92	2.73	**3.49**
Impact on Justice Systems (26 responses) Related to use of SIS within judicial systems (e.g. AI used to 'inform' judicial reviews in areas such as probation)	3.15	4.04	2.54	**3.24**
Impact on Democracy (25/26 responses) Related to the degree to which all involved feel they have an equal say in the outcomes, compared with the SIS	4.08	4.15	3.16	**3.80**
Loss of Freedom and Individual Autonomy (25 responses) Related to how SIS affects how people perceive they are in control of decisions, how they analyse the	3.72	3.92	3.08	**3.57**

world, how they make decisions (e.g. impact of manipulative power of algorithms to nudge toward preferred behaviours), how they interact with one another, and how they modify their perception of themselves and their social and political environment				
Human Contact (25/26 responses) Related to the potential for SIS to reduce the contact between people, as they take on more of the functions within a society	3.42	3.24	2.50	**3.05**
Loss of Human Decision-Making (26 responses) Related to how SIS affects how people analyse the world, make decisions, interact with one another, and modify their perception of themselves and their social and political environment	3.58	3.46	2.54	**3.19**
Control and Use of Data and Systems (25-26 responses) Related to how data is used and commercialised, including malicious use (e.g. mass surveillance); how data is collected, owned, stored, and destroyed; and how consent is given	4.04	4.08	3.20	**3.77**
Potential for Military Use (25 responses) Related to the use of SIS in future possible military scenarios (e.g. autonomous weapons), including the potential for dual-use applications (military and non-military)	3.04	3.88	2.92	**3.28**
Potential for Criminal and Malicious Use (25 responses) Related to the use of SIS in criminal and malicious scenarios (e.g. cyber-attacks and cyber espionage)	3.76	4.28	3.12	**3.72**
Ownership of Data (25 responses) Related to who owns data, and how transparent that is (e.g. when you give details to an organisation, who then 'owns' the data, you or that organization?)	3.88	3.48	2.72	**3.36**
Lack of Informed Consent (24 responses) Related to informed consent being difficult to uphold in SIS when the value and consequences of the information that is collected is not immediately known by users and other stakeholders, thus lowering the	3.75	3.50	2.75	**3.33**

possibility of upfront notice				
Lack of Accountability and Liability (25 responses) Related to the rights and legal responsibilities (e.g. duty of care) for all actors (including SIS) from planning to implementation of SIS, including responsibility to identify errors or unexpected results	3.60	4.00	2.60	**3.40**
Accuracy of Predictive Recommendations (25 responses) Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS interprets an individual's personal data	3.56	3.92	2.76	**3.41**
Accuracy of Non-Individualized Recommendations (25 responses) Related to the possibility of misinterpreting data, implementing biases, and diminishing the accuracy of SIS recommendations when SIS makes a decision based on data not specific to an individual	3.64	3.24	2.36	**3.08**
Power Relations (25 responses) Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to 'act above the law'	3.84	3.84	2.92	**3.53**
Concentration of Economic Power (25 responses) Related to growing economic wealth of companies controlling SIS (e.g. big technology companies) and individuals, and unequal distribution of resources	3.88	4.08	3.28	**3.75**
Power Asymmetries (25 responses) Related to the ability of individuals to frame and partake in dialogue about issues; and the fact that few powerful corporations develop technologies, influence political processes, and have know-how to 'act above the law'	3.84	4.00	3.16	**3.67**
Lack of Access to and Freedom of Information (25 responses) Related to quality and trustworthiness of information available to the public (e.g. fake news, deepfakes) and	3.96	4.20	3.40	**3.85**

the way information is disseminated and accessed				
Accuracy of Data (25 responses) Related to using misrepresentative data or misrepresenting information (ie. predictions are only as good as the underlying data) and how that affects end user views on what decisions are made (ie. whether they trust the SIS and outcomes arising from it)	3.84	4.08	2.64	**3.52**
Integrity (23 responses) Related to the internal integrity of the data used as well as the integrity of how the data is used by a SIS	3.30	3.43	2.22	**2.99**
Impact on Health (24 responses) Related to the the use of SIS to monitor an individual's health and how much control one can have over that	3.75	4.38	3.25	**3.79**
Impact on Vulnerable Groups (24 responses) Related to how SIS creates or reinforces inequality and discrimination (e.g. impacting on the dignity and care for older people, for example how much a care robot might exert over an older person's life and 'tell them what to do'	3.21	4.04	2.75	**3.33**
Violation of End-Users Fundamental Human Rights (24 responses) Related to how human rights are impacted for end-users (e.g. monitoring and control of health data impacting right to health; manipulative power of algorithms nudging towards some preferred behaviours, impacting rights to dignity and freedom	3.75	4.08	3.04	**3.63**
Violation of Fundamental Human Rights in Supply-Chain (23 responses) Related to how human rights are impacted for those further down the supply-chain extracting resources and manufacturing devices (e.g. impacts on health, labour violations, lack of free, prior and informed consent for extractives	3.04	3.57	2.39	**3.00**
Lack of Quality Data (24 responses) Related to using misrepresentative data or misrepresenting information in building AI models	3.33	3.92	2.46	**3.24**

Disappearance of Jobs (24 responses) Related to concerns that use of SIS will lead to significant drop in the need to employ people	3.29	3.71	4.00	**3.67**
Prioritization of the "Wrong" Problems (24 responses) Related to the problems SIS is developed to 'solve' and who determines what the immediate problems are	2.79	3.29	2.33	**2.81**
"Awakening" of AI (23 responses) Related to concerns about singularity, machine consciousness, super-intelligence etc. and the future relationship of humanity vis-a-vis technology	2.78	2.74	3.52	**3.01**
Security (23 responses) Related to the vulnerabilities of SIS and their ability to function correctly under attacks or timely notify human operators about the need of response and recovery operations	3.65	4.00	2.83	**3.49**
Lack of Trust (23-24 responses) Related to using misrepresentative data or misrepresenting information (ie. predictions are only as good as the underlying data) and how that affects end user views on what decisions are made (ie. whether they trust the SIS and outcomes arising from it); also related to informed consent and that helps with trust	3.88	3.96	3.35	**3.73**
Access to Public Services (24 responses) Related to how SIS could change the delivery and accessibility of public services for all (e.g. through privatisation of services)	3.63	3.67	2.71	**3.33**
Harm to Physical Integrity (24 responses) Related to the potential impacts on our physical bodies (e.g. from self-driving cars, autonomous weapons)	3.17	3.67	3.58	**3.47**
Cost to Innovation (24 responses) Related to balancing the protection of rights and future technological innovation	2.92	3.00	2.75	**2.89**

Unintended, Unforeseeable Adverse Impacts (24 responses) Related to future challenges and impacts that are yet known	3.29	3.67	2.88	**3.28**
Impact on Environment (24 responses) Related to concern about the environmental consequences of infrastructures and devices needed to run SIS (e.g. demand for physical resources and energy)	3.71	3.79	3.00	**3.50**
Average	3.58	3.81	2.95	3.45

Do you have any further comments regarding Ethical and Human Rights Issues?

- I think I miss a point in relation to "education" and "education gap", i.e. as with any new technology , the gap between those who use and understand it and those who lack this knowledge and how to minimize it.
- Governments and the public sector do not as yet incentivise private sector companies to behave with recognition of ethical, human rights or sustainability. These elements should be part of statutory business reporting in the same way as financial indices and be included in public sector procurement.

Question 1 (Ethical and Human Rights Issues) - Top and bottom five issues (reach, significance, attention, and overall)

TOP FIVE REACH		TOP FIVE SIGNIFICANCE		TOP FIVE ATTENTION		TOP FIVE AVERAGE	
Misuse of Personal Data	4.27	Bias and Discrimination	4.40	Disappearance of Jobs	4.00	Misuse of Personal Data	**4.05**
Lack of Privacy	4.19	Misuse of Personal Data	4.38	Lack of Privacy	3.85	Lack of Privacy	**3.96**
Impact on Democracy	4.08	Impact on Health	4.38	Harm to Physical Integrity	3.58	Lack of Access to and Freedom of Information	**3.85**
Control and Use of Data and Systems	4.04	Potential for Criminal and Malicious Use	4.28	"Awakening" of AI	3.52	Bias and Discrimination	**3.80**
Lack of Access to and Freedom of Information	3.96	Lack of Access to and Freedom of Information	4.20	Misuse of Personal Data	3.50	Impact on Democracy	**3.80**
BOTTOM FIVE DESIRABILITY		BOTTOM FIVE FEASIBILITY		BOTTOM FIVE PROBABILITY		BOTTOM FIVE AVERAGE	
Violation of Fundamental Human Rights in Supply-Chain	3.04	Prioritization of the "Wrong" Problems	3.29	Lack of Quality Data	2.46	"Awakening" of AI	**3.01**
Potential for Military Use	3.04	Human Contact	3.24	Violation of Fundamental	2.39	Violation of Fundamental	**3.00**

				Human Rights in Supply-Chain		Human Rights in Supply-Chain	
Cost to Innovation	2.92	Accuracy of Non-Individualized Recommendations	3.24	Accuracy of Non-Individualized Recommendations	2.36	Integrity	**2.99**
Prioritization of the "Wrong" Problems	2.79	Cost to Innovation	3.00	Prioritization of the "Wrong" Problems	2.33	Cost to Innovation	**2.89**
"Awakening" of AI	2.78	"Awakening" of AI	2.74	Integrity	2.22	Prioritization of the "Wrong" Problems	**2.81**

Question 1 (Ethical and Human Rights Issues) – High and mid-high scoring issues (reach, significance, and attention)

High Reach (4-4.49)	High Significance (4-4.449)	High Attention (4-4.49)
• Misuse of Personal Data • Lack of Privacy • Impact on Democracy • Control and Use of Data and Systems	• Bias and Discrimination • Misuse of Personal Data • Impact on Health • Potential for Criminal and Malicious Use • Lack of Access to and Freedom of Information • Impact on Democracy • Violation of End-Users Fundamental Human Rights • Concentration of Economic Power • Accuracy of Data • Control and Use of Data and Systems • Impact on Vulnerable Groups • Impact on Justice Systems • Lack of Accountability and Liability • Power Asymmetries • Security	• Disappearance of Jobs
Mid-High Reach (3.5-3.99)	**Mid-High Significance (3.5-3.99)**	**Mid-High Attention (3.5-3.99)**
• Lack of Access to and Freedom of Information • Ownership of Data • Concentration of Economic Power • Lack of Trust • Lack of Transparency • Power Relations	• Lack of Trust • Unfairness • Loss of Freedom and Individual Autonomy • Accuracy of Predictive Recommendations • Lack of Quality Data • Potential for Military Use • Lack of Privacy	• Lack of Privacy • Harm to Physical Integrity • "Awakening" of AI • Misuse of Personal Data

- Power Asymmetries
- Accuracy of Data
- Unfairness
- Potential for Criminal and Malicious Use
- Lack of Informed Consent
- Impact on Health
- Violation of End-Users Fundamental Human Rights
- Loss of Freedom and Individual Autonomy
- Impact on Environment
- Security
- Accuracy of Non-Individualized Recommendations
- Access to Public Services
- Bias and Discrimination
- Lack of Accountability and Liability
- Loss of Human Decision-Making
- Accuracy of Predictive Recommendations
- Power Relations
- Impact on Environment
- Lack of Transparency
- Disappearance of Jobs
- Access to Public Services
- Harm to Physical Integrity
- Unintended, Unforeseeable Adverse Impacts
- Violation of Fundamental Human Rights in Supply-Chain
- Lack of Informed Consent

Question 2 (Potential Regulatory Measures) - Average scores (reach, significance, attention, and overall)

Potential Regulatory Measures	**Desirability**	**Feasibility**	**Probability**	**Average**
Creation of new international treaty for AI and Big Data (21 responses) (open for adoption by all countries)	3.86	3.05	2.38	3.0952
Better enforcement of existing international human rights law (21 responses)	4.29	3.33	2.71	3.4444
Binding Framework Convention to ensure that AI is designed, developed and applied in line with European standards on human rights, democracy and the rule of law (Council of Europe) including through a new ad hoc committee on AI (CAHAI) (22-21 responses)	4.09	3.43	3.00	3.5065

CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment (22 responses)	3.77	3.59	3.09	3.4848
Legislative framework for independent and effective oversight over the human rights compliance of the development, deployment and use of AI systems by public authorities and private entities (Council of Europe) (21 responses)	4.10	3.81	3.19	3.6984
General fund for all smart autonomous robots or individual fund for each and every robot category (EU Parliament) (19-20 responses)	2.79	3.00	2.45	2.7465
Establishment of a comprehensive Union system of registration of advanced robots within the Union's internal market where relevant and necessary for specific categories of robots and establishment of criteria for the classification of robots (22 responses)	3.27	2.95	2.32	2.8485
Algorithmic impact assessments under the General Data Protection Regulation (GDPR) (21 responses)	4.10	3.67	3.19	3.6508
Creation of new body: **EU Taskforce/Coordinating body of field-specific regulators for AI/big data** (22 responses)	3.45	3.09	2.64	3.0606
Redress-by-design mechanisms for AI (High-Level Expert Group on Artificial Intelligence (AI HLEG) (19-20 responses)	3.74	3.25	2.60	3.1956
New laws regulating specific aspects, e.g., deepfakes, algorithmic accountability. (21 responses)	3.71	3.19	2.86	3.2540
Register of algorithms used in government (21 responses)	3.86	3.19	2.67	3.2381
New national independent cross-sector advisory body (e.g. UK Centre for Data Ethics and Innovation) (22 responses)	3.50	3.82	3.45	3.5909
New specialist regulatory agency to regulate algorithmic safety (21-22 responses)	3.48	3.14	2.59	3.0678
Public Register of Permission to Use Data	3.19	2.76	2.19	2.7143

(individuals provide affirmative permission in a public register for companies to use their data) (21 responses)				
Reporting Guidelines (20 responses) (for publicly registered or traded companies based on corporate social responsibility reporting as described by GRI)	3.80	3.55	3.15	3.5000
Regulatory sandboxes for AI and big data (19 responses)	3.74	3.53	3.05	3.4386
Three-level obligatory impact assessments for new technologies (18-19 responses)	3.83	3.16	2.68	3.2251
Average	3.70	3.31	2.79	3.2645
Do you have any further comments regarding Potential Regulatory Measures? NONE				

Question 2 (Potential Regulatory Measures) - Top and bottom five issues (reach, significance, attention, and overall)

TOP FIVE DESIRABILITY		TOP FIVE FEASIBILITY		TOP FIVE PROBABILITY		TOP FIVE AVERAGE	
Better enforcement of existing international human rights law	4.26	New national independent cross-sector advisory body	3.82	New national independent cross-sector advisory body	3.45	Legislative framework for independent and effective oversight of human rights	3.6984
Legislative framework for independent and effective oversight of human rights	4.10	Legislative framework for independent and effective oversight of human rights	3.81	Legislative framework for independent and effective oversight of human rights	3.19	Algorithmic impact assessments under the GDPR	3.6508
Algorithmic impact assessments under the GDPR	4.10	Algorithmic impact assessments under the GDPR	3.67	Algorithmic impact assessments under the GDPR	3.19	New national independent cross-sector advisory body	3.5909
Binding Framework Convention	4.09	CEPEJ European Ethical Charter	3.59	Reporting Guidelines	3.15	Binding Framework Convention	3.5065

International treaty for AI and Big Data	3.86	Reporting Guidelines	3.55	CEPEJ European Ethical Charter	3.09	Reporting Guidelines	3.50 00
BOTTOM FIVE DESIRABILITY		BOTTOM FIVE FEASIBILITY		BOTTOM FIVE PROBABILITY		BOTTOM FIVE AVERAGE	
Specialist regulatory agency to regulate algorithmic safety	3.48	EU Taskforce/Coordinating body	3.09	New specialist regulatory agency to regulate algorithmic safety	2.59	Specialist regulatory agency to regulate algorithmic safety	3.06 78
EU Taskforce/Coordinating body	3.45	International treaty for AI and Big Data	3.05	General fund for all smart autonomous robots or individual fund for each and every robot category	2.45	EU Taskforce/Coordinating body	3.06 06
EU system of registration of advanced robots	3.27	General fund for all smart autonomous robots or individual fund for each and every robot category (EU Parliament)	3.00	International treaty for AI and Big Data	2.38	EU system of registration of advanced robots	2.84 85
Public Register of Permission to Use Data	3.19	EU system of registration of advanced robots	2.95	EU system of registration of advanced robots	2.32	General fund for all smart autonomous robots or individual fund for each and every robot category	2.74 65
General fund for all smart autonomous robots or individual fund for each and every robot category	2.79	Public Register of Permission to Use Data	2.76	Public Register of Permission to Use Data	2.19	Public Register of Permission to Use Data	2.71 43

Question 2 (Potential Regulatory Measures) – High and mid-high scoring issues (reach, significance, and attention)

High Desirability	High Feasibility	High Probability

(4-4.49)	(4-4.49)	(4-4.49)
• Better enforcement of existing international human rights law • Legislative framework for independent and effective oversight of human rights compliance • Algorithmic impact assessments • Binding Framework Convention	-	-
Mid-High Desirability (3.5-3.99)	**Mid-High Feasibility (3.5-3.99)**	**Mid-High Probability (3.5-3.99)**
• Creation of new international treaty for AI and Big Data • Register of algorithms used in government • Three-level obligatory impact assessments for new technologies • Reporting Guidelines • CEPEJ European Ethical Charter • Redress-by-design mechanisms • Regulatory sandboxes • New laws regulating specific aspects • New national independent cross-sector advisory body	• New national independent cross-sector advisory body • Legislative framework for independent and effective oversight of human rights compliance • Algorithmic impact assessments • CEPEJ European Ethical Charter • Reporting Guidelines • Regulatory sandboxes	• New national independent cross-sector advisory body

Question 3 (Potential Technical Measures) - Average scores (reach, significance, attention, and overall)

Potential Technical Measures	Desirability	Feasibility	Probability	Average
Methodologies for systematic and comprehensive testing of AI-based systems (incldluding fairness of decisions) (20 responses)	4.55	3.70	3.45	3.9000
Techniques for providing explanations for output of AI models (e.g., Layerwise relevance propagation for neural networks) (20 responses)	4.50	3.60	3.50	3.8667
Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models (20 responses)	4.10	3.65	3.30	3.6833

AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model) (17-18 responses)	3.94	3.33	3.28	3.5174
Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties (20 responses)	4.50	3.75	3.60	3.9500
Tools for verifying and certifying publicly available services based on machine learning models (19-20 responses)	4.40	3.58	3.16	3.7123
Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models) (19-20 responses)	4.20	3.70	3.00	3.6333
Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public) (19-20 responses)	4.55	3.15	3.05	3.5842
Average	**4.3425**	**3.5575**	**3.2925**	**3.7309**
Do you have any further comments regarding Potential Technical Measures? ● A probability score of 5 indicates that such tools already exist (and can and should continue being developed and improved) ● The thinking here is too black and white; "good" vs "biased" datasets and "good" versus "biased" algos. A dataset can be unbiased for a certain use case and thus "good", but when applied to a different use case it may turn out biased. Also, individual algorithms may work fine but when stacked on top of each other the emergent behavior goes wrong. A bigger systems perspective is needed! ● Maybe I miss some tools for monitoring, following up and assessing human trust on AI-based systems; and / or other metrics on ethics.				

Question 3 (Potential Technical Measures) - Top and bottom five issues (reach, significance, attention, and overall)

TOP THREE DESIRABILITY		TOP THREE FEASIBILITY		TOP THREE PROBABILITY		TOP THREE AVERAGE	
Methodologies for systematic and comprehensive testing of AI-based systems	4.26	Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties	3.75	Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties	3.60	Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties	3.9500
Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public)	4.10	Methodologies for systematic and comprehensive testing of AI-based systems	3.70	Techniques for providing explanations for output of AI models	3.50	Methodologies for systematic and comprehensive testing of AI-based systems	3.9000
Techniques for providing explanations for output of AI models	4.10	Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models)	3.70	Methodologies for systematic and comprehensive testing of AI-based systems	3.45	Techniques for providing explanations for output of AI models	3.8667
BOTTOM THREE DESIRABILITY		BOTTOM THREE FEASIBILITY		BOTTOM THREE PROBABILITY		BOTTOM THREE AVERAGE	
Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc.	4.20	Tools for verifying and certifying publicly available services based on machine learning models	3.58	Tools for verifying and certifying publicly available services based on machine learning models	3.16	Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models)	3.6333

services and models)							
Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models	4.10	AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model)	3.33	Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public)	3.05	Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public)	3.5842
AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model)	3.94	Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content (available and easy-to-use for the general public)	3.15	Reputation information about publicly available services based on machine learning models (e.g. including a black list of known faulty, vulnerable, inaccurate, etc. services and models)	3.00	AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks (e.g. functionality to allow a model's owner to easily determine whether training data can be reverse-engineered from the model)	3.5174

Question 3 (Potential Technical Measures) – High and mid-high scoring issues (reach, significance, and attention)

Very High Desirability (4.5-5)	Very High Feasibility (4.5-5)	Very High Probability (4.5-5)
• Methodologies for systematic and comprehensive testing of AI-based systems • Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content	--	-

• Techniques for providing explanations for output of AI models • Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties		
High Desirability (4-4.49)	**High Feasibility (4-4.49)**	**High Probability (4-4.49)**
• Tools for verifying and certifying publicly available services based on machine learning models Lack of Privacy • Reputation information about publicly available services based on machine learning models • Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models	-	-
Mid-High Desirability (3.5-3.99)	**Mid-High Feasibility (3.5-3.99)**	**Mid-High Probability (3.5-3.99)**
• AI-as-a-service, providing in-built mechanisms to mitigate against common adversarial attacks	• Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties • Methodologies for systematic and comprehensive testing of AI-based systems	• Tools for verifying and certifying labelled datasets for accuracy, bias and other important properties • Techniques for providing explanations for output of AI models

- Reputation information about publicly available services based on machine learning models
- Easily understandable description of the model's inputs (including input validity checks), training data, requirements, and potential limitations for services based on machine learning models
- Techniques for providing explanations for output of AI models
- Tools for verifying and certifying publicly available services based on machine learning models

Question 4 (Other Potential Measures) - Average scores (reach, significance, attention, and overall)

Other Potential Measures	Desirability	Feasibility	Probability	Average
Certification (e.g. initiative for IEEE Ethics Certification Program for Autonomous and Intelligent Systems) (19-20 responses)	3.90	3.74	3.47	3.7035
Citizen Juries to evaluate risk of various AI technologies and propose appropriate tools (19-20 responses)	3.10	2.95	2.42	2.8228
Education Campaigns (e.g. Finnish Element of AI course; Dutch Nationale AI Cursus) (19-20 responses)	4.45	4.37	3.74	4.1851
Ethical Codes of Conduct (e.g. EU High Level Expert Group Guidelines for Trustworthy AI, SHERPA guidelines) (19-20 responses)	3.85	4.21	4.26	4.1079
Ethical Mindset adopted by companies (19-20 responses)	4.35	3.53	2.95	3.6079
Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases (e.g. rules governing recommendation systems: how	3.95	3.32	2.95	3.4044

they should work, what they should not be used for, how they should be properly hardened against attacks, etc.) (19-20 responses)				
Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments (e.g., AI robots resembling dogs, sex robots) (19-20 responses)	3.55	3.26	2.84	3.2184
Exchange of Best Practices (19-20 responses)	4.60	4.53	4.16	4.4281
'Fairness' Officer or Ethics Board employed within companies using/developing SIS (19-20 responses)	3.90	3.79	3.16	3.6158
Framework, Guidelines, and Toolkits for project management and development (e.g. UK Data Ethics Framework; IBM AI Fairness 360 Open Source Toolkit; Dutch Data Ethics Decision Aid (DEDA) tool) (19-20 responses)	4.20	4.21	4.00	4.1368
Grievance Mechanisms for complaints on SIS (19-20 responses)	4.50	4.26	3.21	3.9912
High-level Expert Groups (e.g. UN AI for Good Global Summit) (17-18 responses)	3.61	4.39	4.18	4.0588
Individual Action (e.g. participating in conferences to raise awareness; protecting oneself by refusing cookies online) (20 responses)	4.10	3.85	3.40	3.7833
International Ethical Framework (e.g. OECD Principles on AI) (19-20 responses)	4.05	3.95	3.74	3.9114
Investigative Journalism about issues concerning SIS (19-20 responses)	4.70	4.37	4.37	4.4789
More Open Source Tools that allow for transparency, explainability, and bias mitigation (19-20 responses)	4.55	3.79	3.79	4.0430

NGO Coalitions on particular issues (e.g. Campaign to Stop Killer Robots) (19 responses)	4.05	4.00	3.84	3.9649
Open Letters to governments and the public (e.g. 2015 Open Letter on AI) (19 responses)	3.84	4.00	3.84	3.8947
Public Policy Commitment by company to be ethical (19 responses)	4.11	4.05	3.68	3.9474
Public "Whistleblowing" Mechanisms for the reporting of bias, inaccuracies, or ethical impacts of systems based on machine learning models (19-20 responses)	4.50	3.63	3.16	3.7632
Retaining 'Unsmart' Products and Services by keeping them available to purchase and use (18 responses)	4.06	3.72	2.83	3.5370
Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations (e.g. self-driving vehicles and other systems) (19 responses)	4.37	3.37	3.32	3.6842
Self-Regulation by Company (e.g. Twitter's self-imposed ban on political ads) (20 responses)	4.05	3.50	3.20	3.5833
Stakeholder Dialogue and Scrutiny with scientists, programmers, developers, decision makers, politicians, and the public at large (18-19 responses)	4.32	4.00	3.33	3.8830
Standardisation (e.g. IEEE P7000 series of standards for addressing ethical concerns during system design). (18-19 responses)	4.16	3.56	3.47	3.7290
Third-party Testing and External Audits (e.g. of data used for training for quality, bias, and transparency) (19 responses)	4.21	3.84	3.32	3.7895
Average	**4.12**	**3.85**	**3.49**	**3.8182**
Do you have any further comments regarding Other Potential Measures? ● **Ethical principles are the easy part, they are "feel good principles". The issue**				

is in the translation to tangible guidelines to action, where 1) the principle may be multi-interpretable and 2) principles may turn out to be conflicting with each other. ● **What is needed is not necessarily more sets of ethical principles but rather an international consensus around the key elements of these and effective mechanisms to deploy them. Tools to ensure eg transparency, explainability and fairness are available but there is likely to be reluctance on the part of SIS orgnisations to adopt them, if this is seen to impair innovation and performance - hence some lower scores for probability above. This needs to be integrated into business practice through compulsory, documented self assessments and effective certification and assurance schemes.**	

Question 4 (Other Potential Measures) - Top and bottom five issues (reach, significance, attention, and overall)

TOP FIVE DESIRABILITY		TOP FIVE FEASIBILITY		TOP FIVE PROBABILITY		TOP FIVE AVERAGE	
Investigative Journalism	4.70	Exchange of Best Practices	4.53	Investigative Journalism	4.37	Investigative Journalism about	4.4789
Exchange of Best Practices	4.60	High-level Expert Groups	4.39	Ethical Codes of Conduct	4.26	Exchange of Best Practices	4.4281
More Open Source Tools	4.55	Education Campaigns	4.37	High-level Expert Groups	4.18	Education Campaigns	4.1851
Grievance Mechanisms for complaints on SIS	4.50	Investigative Journalism	4.37	Exchange of Best Practices	4.16	Framework, Guidelines, and Toolkits	4.1368
Public "Whistleblowing" Mechanisms	4.50	Grievance Mechanisms for complaints on SIS	4.26	Framework, Guidelines, and Toolkits	4.00	Ethical Codes of Conduct	4.1079
BOTTOM FIVE DESIRABILITY		BOTTOM FIVE FEASIBILITY		BOTTOM FIVE PROBABILITY		BOTTOM FIVE AVERAGE	
Ethical Codes of Conduct	3.85	Self-Regulation by Company	3.50	Ethical Mindset adopted by companies	2.95	Self-Regulation by Company	3.5833
Open Letters to governments and the public	3.84	Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations	3.37	Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases	2.95	Retaining 'Unsmart' Products and Services	3.5370
High-level Expert Groups	3.61	Ethical Rules pertaining to the creation or use of machine learning models with	3.32	Ethical Rules pertaining to the use or treatment of AI agents in robotics or	2.84	Ethical Rules pertaining to the creation or use of machine learning models with potential malicious	3.4044

		potential malicious applications, covering preventive and reactive cases		virtual environments		applications, covering preventive and reactive cases	
Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments	3.55	Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments	3.26	Retaining 'Unsmart' Products and Services	2.83	Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments	3.2184
Citizen Juries	3.10	Citizen Juries	2.95	Citizen Juries	2.42	Citizen Juries	2.8228

Question 4 (Other Potential Measures) – High and mid-high scoring issues (reach, significance, and attention)

Very High Desirability (4.5-5)	Very High Feasibility (4.5-5)	Very High Probability (4.5-5)
• Investigative journalism • Exchange of best practices • More open source tools • Grievance mechanisms for complaints on SIS • Public "whistleblowing" mechanisms	• Exchange of best practices	-
High Desirability (4-4.49)	**High Feasibility (4-4.49)**	**High Probability (4-4.49)**
• Education campaigns • Rules on how decisions in systems that have the capability to cause physical harm should be made in difficult situations • Ethical Mindset adopted by companies • Stakeholder Dialogue and Scrutiny • Third-party Testing and External Audits • Framework, Guidelines, and Toolkits for project management and development	• High-level expert groups • Education campaigns • Investigative journalism • Grievance mechanisms for complaints on SIS • Ethical codes of conduct • Framework, Guidelines, and Toolkits for project management and development • Public Policy Commitment by company to be ethical • Stakeholder Dialogue and Scrutiny	• Investigative journalism • Ethical codes of conduct • High-level expert groups • Exchange of best practices • Framework, Guidelines, and Toolkits for project management and development

• Standardisation • Public Policy Commitment by company to be ethical • Individual action • Retaining 'Unsmart' Products and Services by keeping them available to purchase and use • International ethical framework • NGO coalitions on particular issues • Self-Regulation by company		
Mid-High Desirability (3.5-3.99)	**Mid-High Feasibility (3.5-3.99)**	**Mid-High Probability (3.5-3.99)**
• Ethical Rules pertaining to the creation or use of machine learning models with potential malicious applications, covering preventive and reactive cases • Certification • Fairness' Officer or Ethics Board employed within companies using/developing SIS • Ethical Codes of Conduct • Open Letters to governments and the public • High-level expert groups • Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments	• International Ethical Framework • Individual actions • Third-party Testing and External Audits • Fairness' Officer or Ethics Board employed within companies • More open source tools • Certification • Retaining 'Unsmart' Products and Services by keeping them available to purchase and use • Public "whistleblowing" mechanisms • Standardisation • Ethical mindset adopted by companies • Self-regulation by company	• NGO coalitions on particular issues • Open Letters to governments and the public • More open source tools • Education campaigns • International ethical framework • Public policy commitment by company to be ethical