SHERPA

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

Fostering an Ecosystem of Artificial Intelligence for Human Flourishing

Recommendations of the SHERPA project

This project has received funding from the

European Union's Horizon 2020 Research and Innovation Programme

Under Grant Agreement no. 786641

Table of Contents

Introduction

SHERPA (Shaping the ethical dimensions of smart information systems– a European perspective, www.project-sherpa.eu) is an EU project that focuses on ethical and human rights aspects of smart information systems, those technologies that use artificial intelligence and big data analytics.

The SHERPA project is action-oriented and aims to make a positive contribution to the broader societal debate around AI and big data. One of its unique aspects is that it has a strong emphasis on advocacy and dedicated resources to highlight the project's findings and promote any recommendations arising from the work of the project. Much of the work in developing options and evaluating and prioritising them is geared towards supporting these activities.

This document describes the process of collecting and categorising possible options and mitigation measures and using them to generate SHERPA project recommendations.

Background

The SHERPA project and its outputs have to be interpreted before the very crowded background of academic, technical, media, civil society and policy discussions of AI, big data and their ethical implications. Particularly noteworthy activities are the EU High Level Expert Group on AI and the guidelines it has published, the IEEE work on ethical AI, bottom-up initiatives by researchers, such as the Asilomar group as well as industry-driven initiatives such as the Partnership on AI.

In light of this very crowded field, SHERPA needs to avoid duplication of efforts, ensure that its recommendations are developed cognisant of parallel developments and can be delivered to the appropriate audiences.

This document outlines the principles of the SHERPA recommendations, how they are arrived at and invites comments to help shape the eventual recommendations.

Assumptions and Axioms

The following points inform the subsequent discussion

- SHERPA is unlikely to find the magic bullet. There is no one solution to all the ethical and human rights issues of SIS.
- The eventual outcome of this exercise is an "intelligent mix" of various options. SHERPA's contribution can be to inject some intelligence (and empirical understanding) into the mix.
- In order to have an impact, SHERPA needs to be visible. This requires, among others,
    - A clear and simple storyline

- Clear and simple supporting materials

Logic of Recommendations Development

The recommendations should be relevant and address a current and pressing issue. The following process expressed in the flowchart below can help identify these.

1. Compile set of ethical / human rights issues that arise from SIS.
   Receives input from earlier SHERPA activities, including case studies, scenarios, technical analysis, ethical and human rights analysis, Delphi Study round 1 and the online survey
2. Investigate which measures already exist to address these.
   Uses input from all WP3 activities, Delphi Study round 1, Input from WP2 activities, including Stakeholder Board interaction
3. Are measures efficient and effective? What are strengths and limitations?
   Investigated through stakeholder engagement (WP2) and evaluation mechanisms, notably focus groups in T2.2
4. Rank open issues, which ones are most pressing?
   Input from stakeholder activities, Delphi Study round 2
5. Develop consortium recommendations
   Consortium deliberation taking into account all activities and insights generated by prior steps.

The following flowchart is meant to represent the logic of this process. The red / pink rectangles represent activities that are undertaken by the SHERPA project as part of the Description of Action, the SHERPA project contract..

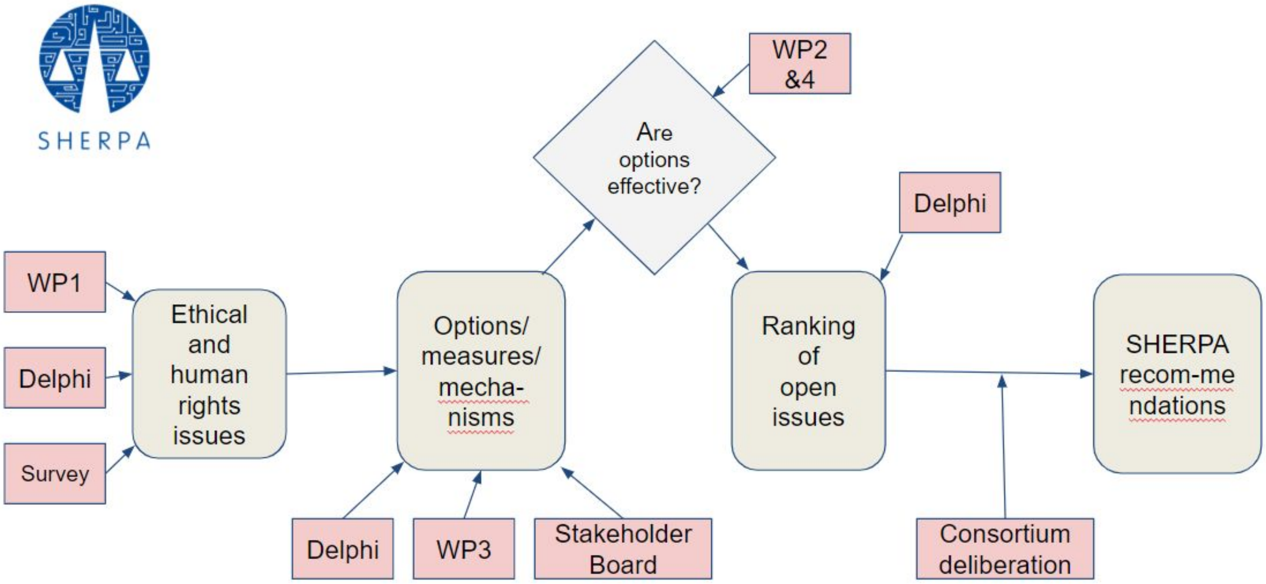

Figure X: Flowchart for identifying Recommendations

Inclusion criteria

The following criteria for including recommendations into the overall SHERPA recommendations are considered by the SHERPA consortium:

1. Recommendations need to be actionable. It needs to be clear who they are addressed to and what we want them to do.
2. Recommendations need to have a clear audience / target group
3. Recommendations should not be redundant
4. Recommendations should not be already implemented
5. Recommendations should be clearly linked to SHERPA work.

Following this logic, the first step in developing the recommendation is to identify ethical issues.

Ethical and Human Rights Issues

This section covers the identification of these issues and their categorisation.

Identification of Ethical and Human Rights Issues

The identification of ethical and human rights issues was the core of WP1. During the case study research (T1.1) participants were asked which issues they encountered. The scenarios (T1.2) were focused on these issues. Ethical (T1.4) and human rights analysis (T1.5) focused on them. The technical analysis (T1.3), while more focused on technical issues took these issues into account. Various stakeholder engagement events, including interviews and the first Stakeholder Board meeting focused on them. The first round of the Delphi Study provided an opportunity to the expert panel to suggest issues to be considered. The following figure represents a consolidated overview of all the ethical issues identified in all of these activities. It is based on the list of issues that was used during the second round of the Delphi Study, which aimed to include all insights gathered by the consortium up until the time of delivering it in early March 2020. A full list of the issues, including brief definitions is available in appendix 1.

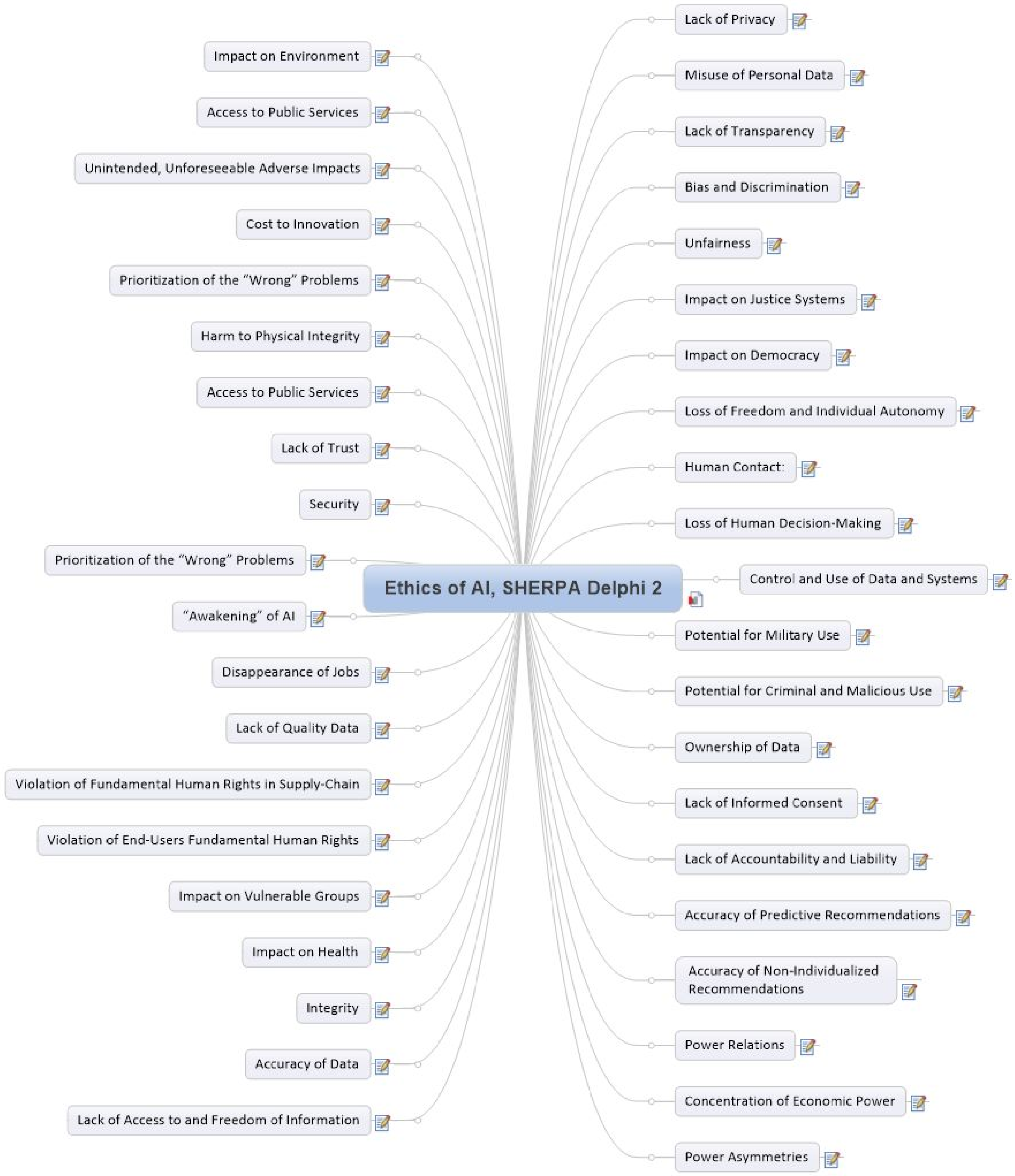
Ethics of AI, SHERPA Delphi 2
Impact on Environment
Access to Public Services
Unintended, Unforeseeable Adverse Impacts
Cost to Innovation
Prioritization of the “Wrong” Problems
Harm to Physical Integrity
Access to Public Services
Lack of Trust
Security
Prioritization of the “Wrong” Problems
“Awakening” of AI
Disappearance of Jobs
Lack of Quality Data
Violation of Fundamental Human Rights in Supply-Chain
Violation of End-Users Fundamental Human Rights
Impact on Vulnerable Groups
Impact on Health
Integrity
Accuracy of Data
Lack of Access to and Freedom of Information
Lack of Privacy
Misuse of Personal Data
Lack of Transparency
Bias and Discrimination
Unfairness
Impact on Justice Systems
Impact on Democracy
Loss of Freedom and Individual Autonomy
Human Contact:
Loss of Human Decision-Making
Control and Use of Data and Systems
Potential for Military Use
Potential for Criminal and Malicious Use
Ownership of Data
Lack of Informed Consent
Lack of Accountability and Liability
Accuracy of Predictive Recommendations
Accuracy of Non-Individualized Recommendations
Power Relations
Concentration of Economic Power
Power Asymmetries

Figure 2: Consolidated overview of ethical and human rights issues

It is worth noting that this is a descriptive list that is based on accounts of what respondents during the various SHERPA activities highlighted as ethical issues. It does not imply or presuppose a particular position in philosophical ethics, nor does it reflect a judgment by the SHERPA consortium with regards to what is or should be regarded as an ethical issue. The aim behind creating this list was to provide a comprehensive starting point for the exploration of possible mitigation measures. As a consequence it is likely that there is some overlap between these issues. The list furthermore does not distinguish between ethical and human rights issues.

In order to be able to work with the issues and identify possible mitigation strategies, it is helpful to not work with a long list but come up with a categorisation or classification of these issues, which we discuss in the next section.

Categorisation of Ethical and Human Rights Issues

An initial categorisation of the ethical issues that were identified as part of the work undertaken in WP1 was proposed and discussed by the consortium during the general assembly in Cyprus (9/10.10.2019). It forms part of a paper (currently under development) that describes the insights from WP1.

The categorisation (represented in figure 2, below) suggests that ethical issues of AI can be divided into three categories:

- Specific issues arising from machine learning
  These issues are linked to the specific characteristics of machine learning as a core AI technique underpinning much of the current AI discourse (opacity, black box nature, need for large training data sets).
- General questions about living in a digital world
  These issues are typically higher level concerns that are not subject to resolution on a local level but that are caused by larger, often societal-level structures, which can have consequences for AI or which can be affected by AI in a way to raise concerns. Examples of such macro level issues include the economic and political power concentration of particular actors (notably the big Internet companies) which is supported and exacerbated by novel technologies like AI, questions of the legitimate use of AI in contested areas, like autonomous weapon use in warfare or the degree to which technologies can and should structure the autonomous choice of individuals.
- Metaphysical questions
  The final group of issues are the meta-level ones, which are issues that touch on deeper conceptual and philosophical questions, such as the change of human nature due to technology or the possibility of technology acquiring agency or consciousness and subsequent questions of the possibility or likelihood of developments such as superintelligence or the so-called singularity.

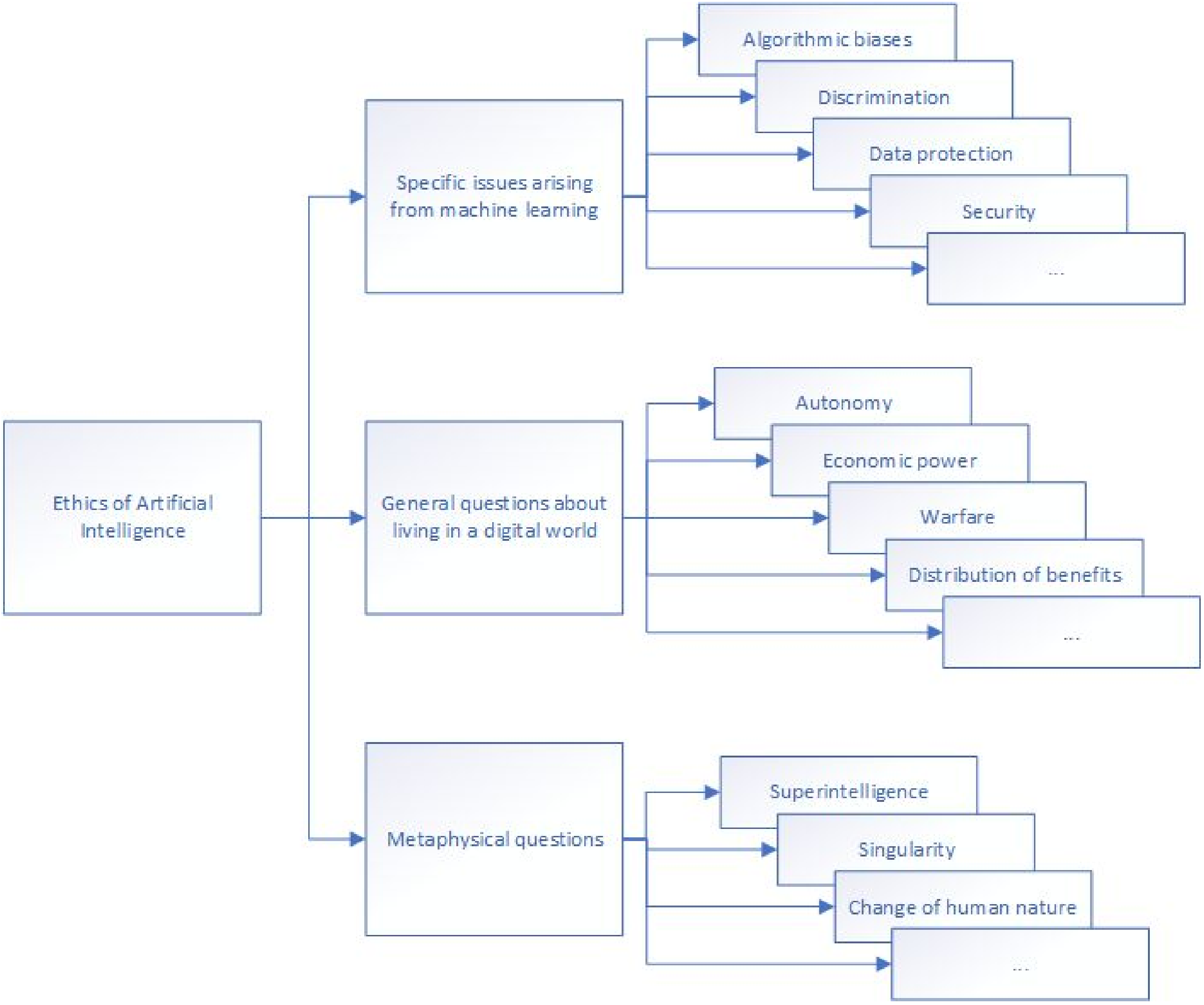

Figure 2: categorisation of ethics and human rights aspects of SIS (Source: Stahl et al. under review)

This categorisation is similar to the distinction of micro, meso and macro level issues suggested by (Haenlein and Kaplan, 2019), but it is not identical.

This categorisation of issues has a number of advantages.

1. The categories provide a first tentative pointer to the set of stakeholders who are best placed to address the issues.
2. It allows the consortium to check whether measures are already in place to address the issues and compare SHERPA proposals with existing proposals.
3. It lends itself to understandable communication with stakeholders involved in the SHERPA project and thereby to structuring feedback

It is important that the categorisation displayed in figure 1 is an initial idea and subject to revision and development. One can argue, for example, that the specific issues of machine learning are a subset of the the broader category of issues arising in a digital world. The three categories are not clearly delineated.

The consortium tested the categories using a subset of the ethical issues, namely those arising from the case studies. This analysis showed that the ethical issues can usefully be allocated to the main categories. It also shows that intermediate levels of categories can also be used to make further sense of the categories.

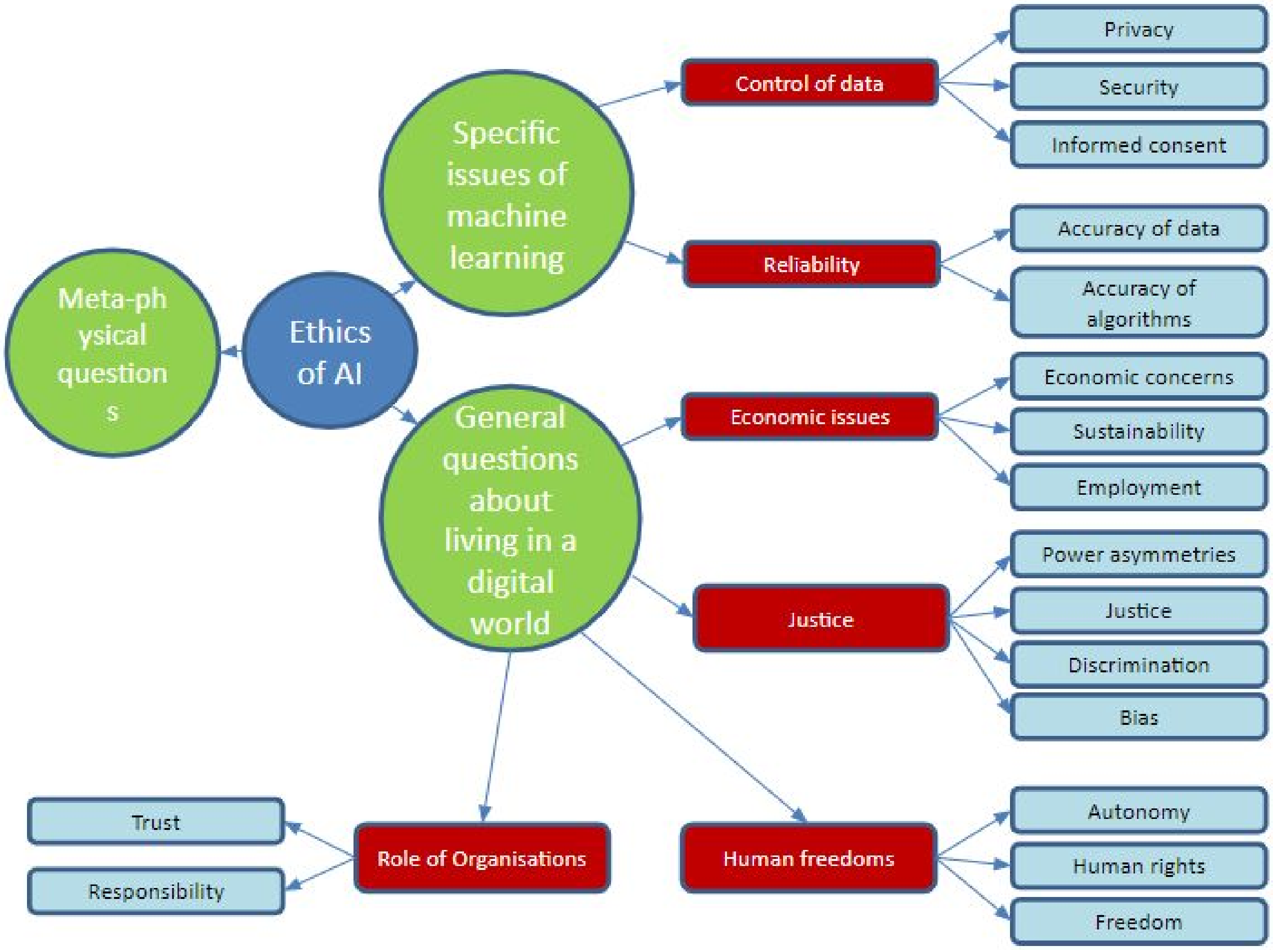

Figure 3: Ethical issues identified in the case studies (source: Ryan et al., under review)

The figure indicates that during the case studies there was very little attention to the metaphysical issues. This is not surprising, as the cases focused on organisational use of current technology which is typically designed to perform specific tasks and therefore does not come near to general AI capabilities that would form part of the metaphysical concerns. Participants were also not asked about these issues. The findings thus do not imply a position on whether and to what degree the metaphysical issues are relevant and in need of attention.

While there is thus scope for further revision and development of the categorisation, an initial attempt to categorise all of the ethical issues so far identified by SHERPA (see appendix 1) gives a similar picture. While individual allocation of issues to categories or subcategories may be debatable, the overall schema can accommodate all ethical issues.

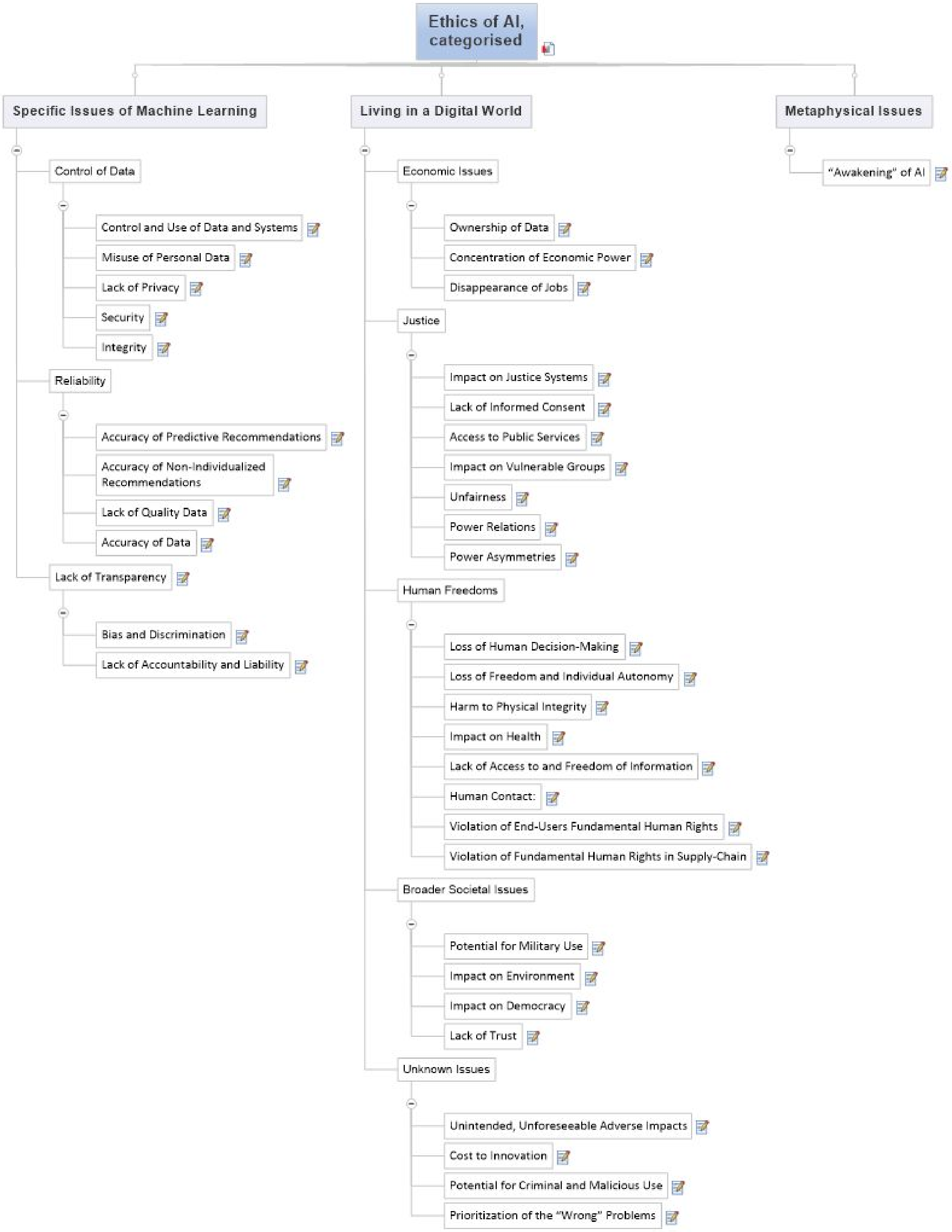
Ethics of AI, categorised
Specific Issues of Machine Learning
Control of Data
Control and Use of Data and Systems
Misuse of Personal Data
Lack of Privacy
Security
Integrity
Reliability
Accuracy of Predictive Recommendations
Accuracy of Non-Individualized Recommendations
Lack of Quality Data
Accuracy of Data
Lack of Transparency
Bias and Discrimination
Lack of Accountability and Liability
Living in a Digital World
Economic Issues
Ownership of Data
Concentration of Economic Power
Disappearance of Jobs
Justice
Impact on Justice Systems
Lack of Informed Consent
Access to Public Services
Impact on Vulnerable Groups
Unfairness
Power Relations
Power Asymmetries
Human Freedoms
Loss of Human Decision-Making
Loss of Freedom and Individual Autonomy
Harm to Physical Integrity
Impact on Health
Lack of Access to and Freedom of Information
Human Contact:
Violation of End-Users Fundamental Human Rights
Violation of Fundamental Human Rights in Supply-Chain
Broader Societal Issues
Potential for Military Use
Impact on Environment
Impact on Democracy
Lack of Trust
Unknown Issues
Unintended, Unforeseeable Adverse Impacts
Cost to Innovation
Potential for Criminal and Malicious Use
Prioritization of the "Wrong" Problems
Metaphysical Issues
"Awakening" of AI

Figure 4: Categorisation of all Ethical Issues

Mitigation Measures

Recommendations arising from the SHERPA projects should be able to address the ethical issues. The project undertook a number of activities to identify and develop recommendations.

In compiling the possible mitigation measures, SHERPA took a similar approach to the one used in the compilation of ethical issues. This means it drew on the core activities of the project, notably those undertaken in WP3. In addition it looked at the discussion in the literature and made use of the first round of the Delphi Study to allow the expert panel to suggest possible mitigation measures.

SHERPA specific measures

Guidelines

SHERPA has already developed two sets of guidelines, one for users, one for developers of AI. These were submitted as a deliverable, are available on the website and are the subject of several of the SHERPA focus groups.

Regulatory options (including regulator)

The consortium investigated current discussions of regulatory activities on a national, European and international level. It evaluated the regulatory options with a view to identifying which ones are most likely to be successful.

Standardisation

SHERPA interacted with standardisation organisations in various different ways, contributing to several standardisation processes.

Technical measures

One particular aspect that received the attention of the project was the technical side, which focused on questions of cybersecurity of AI, in particular, on resilience to model poisoning.

Consortium Brainstorm

As the first step in identifying candidates for recommendations to be adopted by SHERPA, the consortium agreed to collect ideas from all partners, based on all work undertaken by the consortium so far. The full file includes the text of the recommendations. The graphical part of it is shown in the following figure:

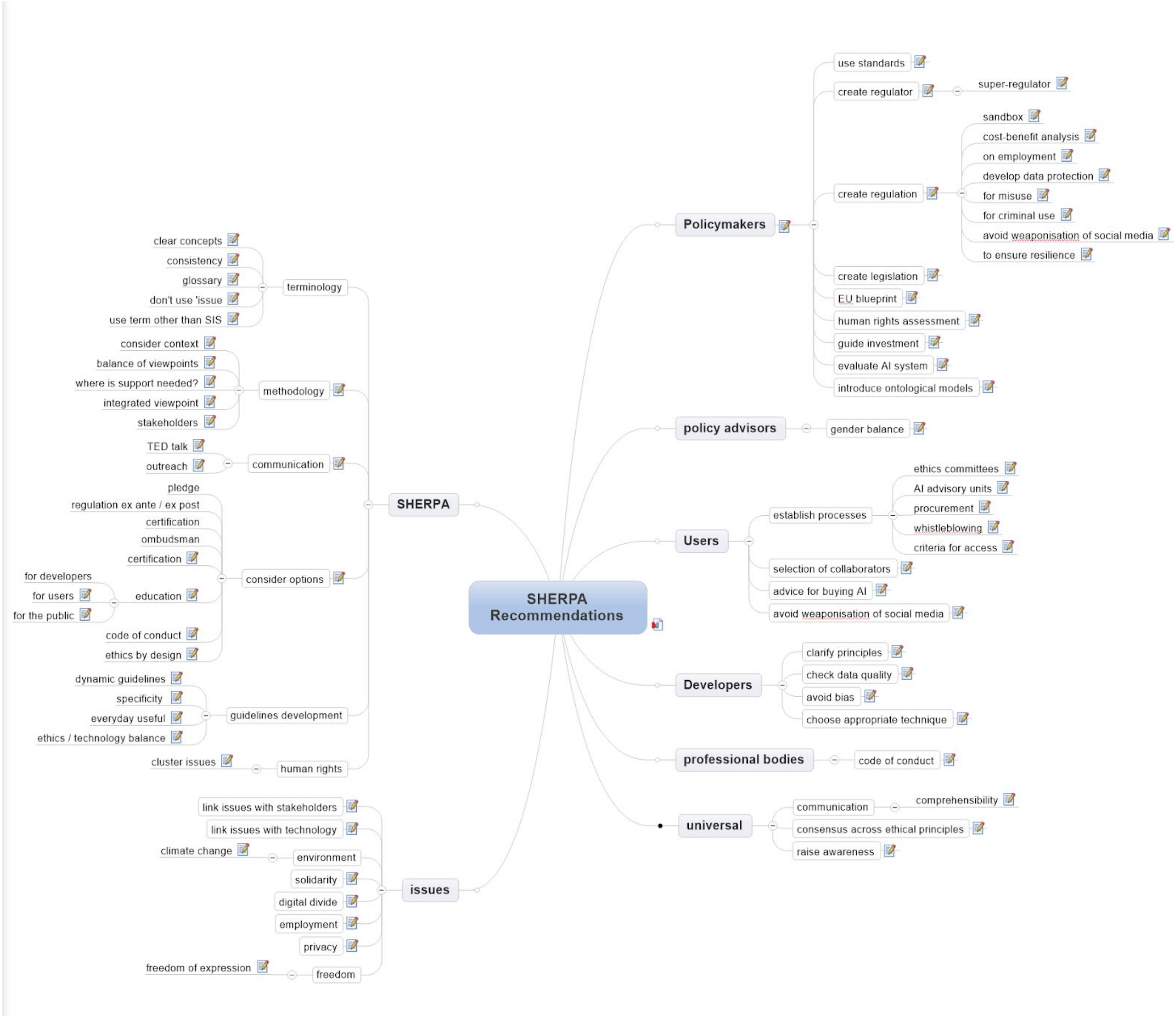

Figure 1: Mindmap of SHERPA recommendations on 19.12.2019

Delphi Study

In the Delphi Study round 1 respondents were asked “Which current approaches, methods, or tools for addressing these issues are you aware of?” The table below summarises their responses

International Measures	● United Nations and international organisations (4) ● Non-binding frameworks (2)
Regional Measures	● Regulations (13) ● High-Level Expert Group (4)
Domestic Government [1] **Measures**	● Regulation (3) ● National Policies & Public-Sector Frameworks (5) ● Awareness and Education Campaigns (2)[2]
Industry-developed Initiatives	● Ethical Codes, Guidelines & Toolkits (8) ● Technical Measures ● Stakeholder Participation ● Standardisation ● Policy Commitments ● Self-Regulation
NGO & Civil Society-developed Initiatives	● Educational Tools (2) ● Guidance and Frameworks (3) ● NGO Coalitions (1) ● Open Letter (1)
Investigative Journalism	
Individual Action	

Table X: Summary of responses to Q2 of Delphi Study round 1

Literature

There is a vast literature on ethics of AI, often containing discussions of possible mitigation measures. The mindmap below is an attempt to capture the principal ideas.

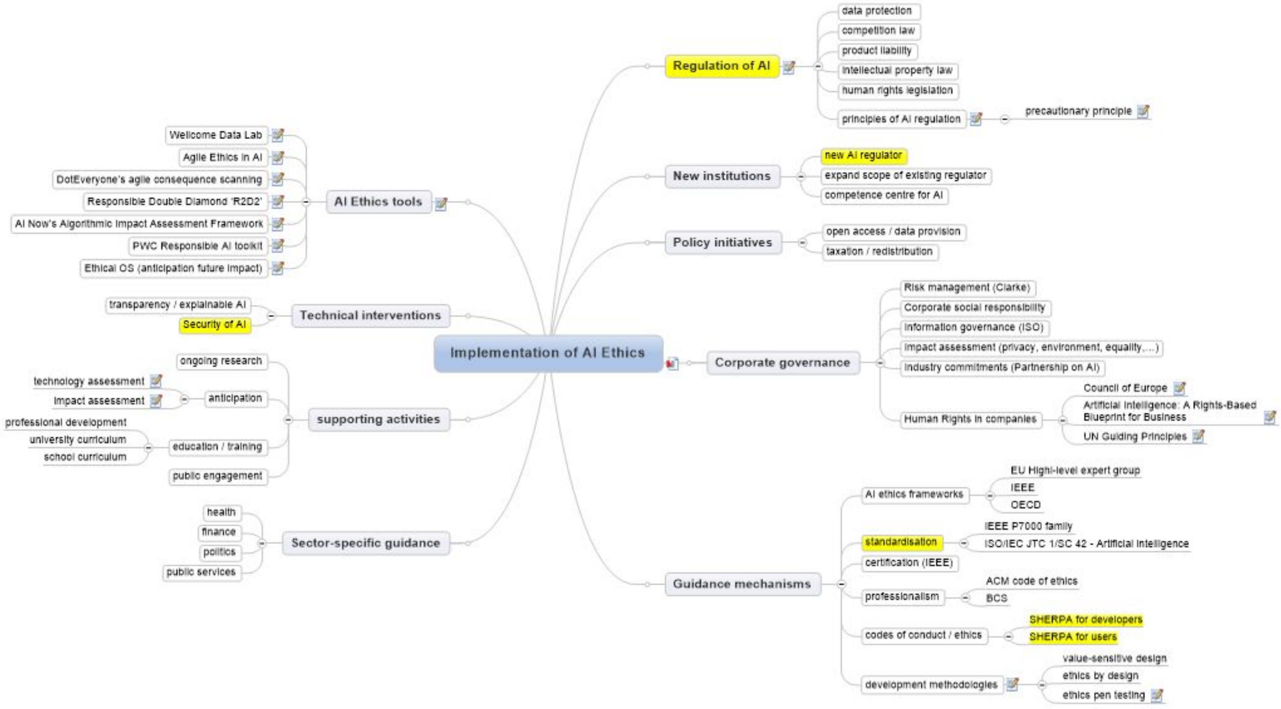

Figure Y: AI Ethics recommendations from the literature (not systematic). Yellow: SHERPA work

Summary of Mitigation Measures

The following figure shows the summary of the mitigation measures drawing on all insights and inputs listed above. The list was included in the second round of the Delphi Study (for a detailed list see appendix 2). The measures are divided into regulatory measures, technical measures and other measures.

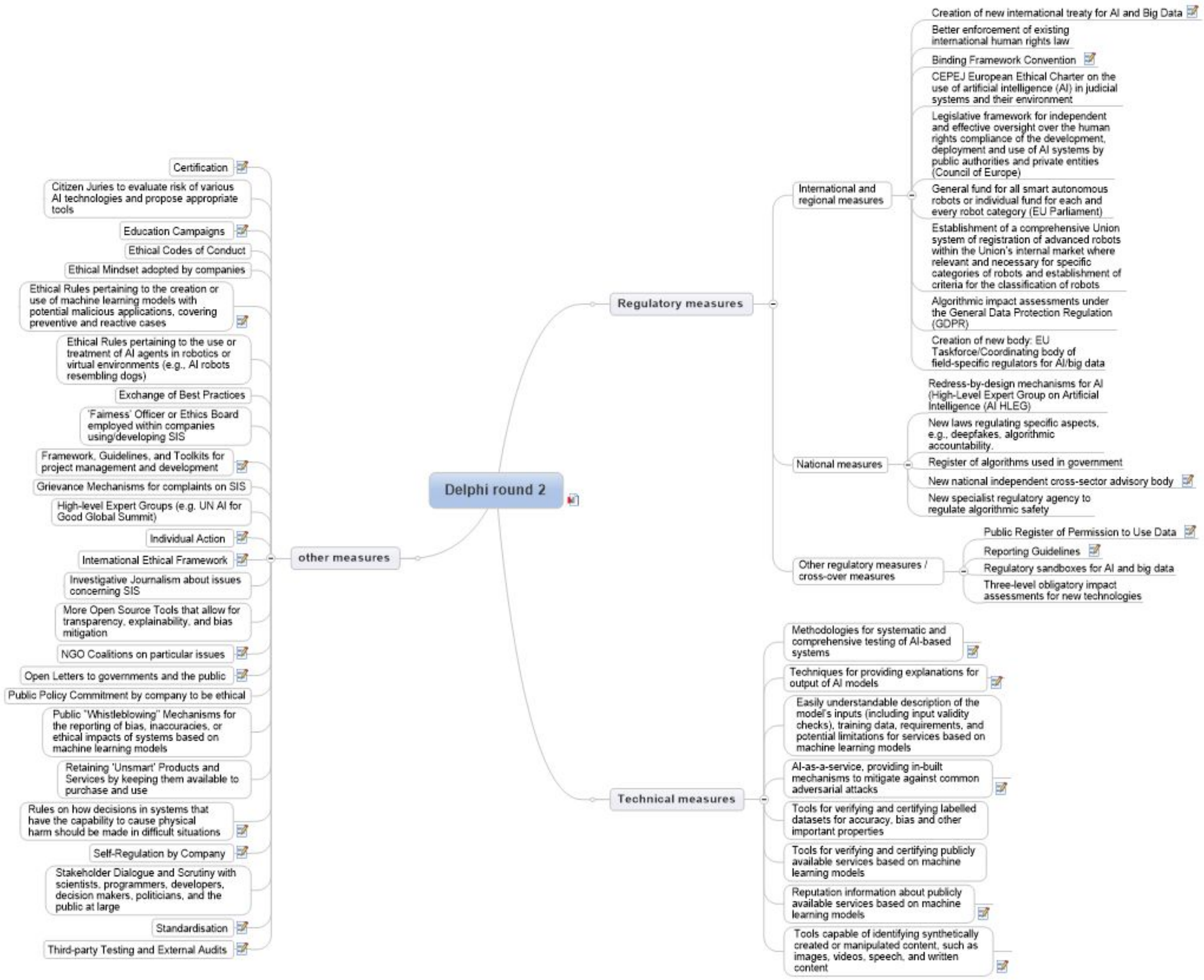
Delphi round 2
Regulatory measures
International and regional measures
Creation of new international treaty for AI and Big Data
Better enforcement of existing international human rights law
Binding Framework Convention
CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial 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
National measures
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
New specialist regulatory agency to regulate algorithmic safety
Other regulatory measures / cross-over measures
Public Register of Permission to Use Data
Reporting Guidelines
Regulatory sandboxes for AI and big data
Three-level obligatory impact assessments for new technologies
Technical measures
Methodologies for systematic and comprehensive testing of AI-based systems
Techniques for providing explanations for output of AI 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
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
Tools for verifying and certifying publicly available services based on machine learning models
Reputation information about publicly available services based on machine learning models
Tools capable of identifying synthetically created or manipulated content, such as images, videos, speech, and written content
other measures
Certification
Citizen Juries to evaluate risk of various AI technologies and propose appropriate tools
Education Campaigns
Ethical Codes of Conduct
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
Ethical Rules pertaining to the use or treatment of AI agents in robotics or virtual environments (e.g., AI robots resembling dogs)
Exchange of Best Practices
'Fairness' Officer or Ethics Board employed within companies using/developing SIS
Framework, Guidelines, and Toolkits for project management and development
Grievance Mechanisms for complaints on SIS
High-level Expert Groups (e.g. UN AI for Good Global Summit)
Individual Action
International Ethical Framework
Investigative Journalism about issues concerning SIS
More Open Source Tools that allow for transparency, explainability, and bias mitigation
NGO Coalitions on particular issues
Open Letters to governments and the public
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
Self-Regulation by Company
Stakeholder Dialogue and Scrutiny with scientists, programmers, developers, decision makers, politicians, and the public at large
Standardisation
Third-party Testing and External Audits

Figure X: Overview of Mitigation Measures included in Delphi Study round 2

Stakeholder Mapping

The next step will include a mapping of stakeholders to relevant mitigation measures. A first attempt is provided in the following figure.

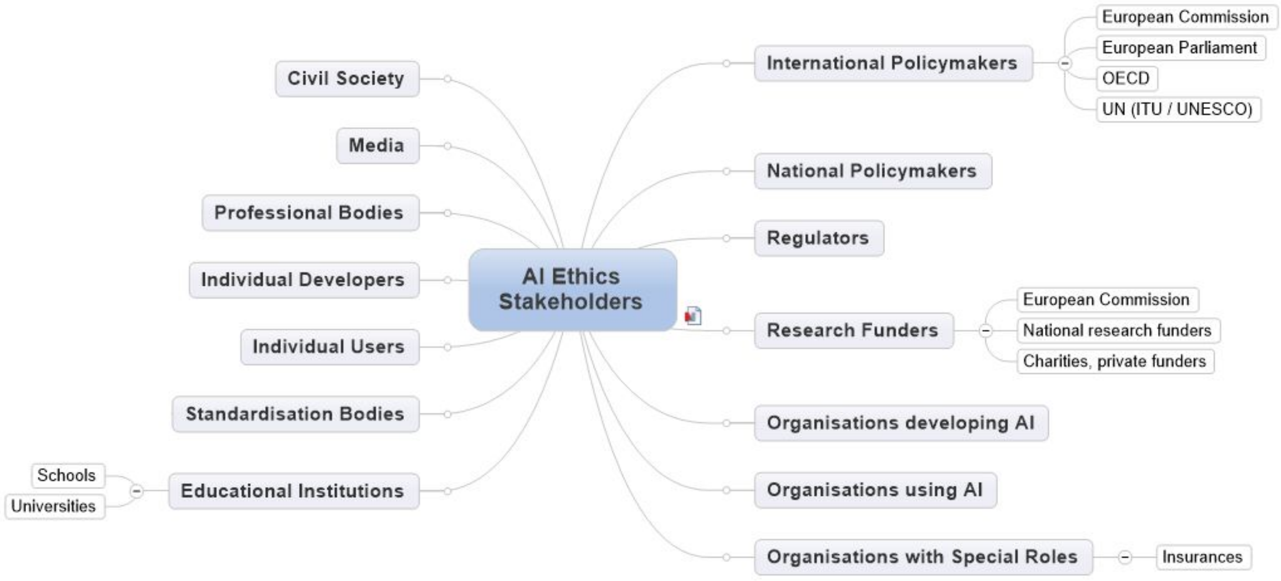

Figure X: Stakeholders of AI Ethics

A stakeholder mapping between the stakeholders and the mitigation measures will be undertaken as part of the next steps of development of the recommendations.

What Next? An Ecosystem of AI for Human Flourishing

The collection, categorisation and description of the ethical and human rights issues as well as the broad range of possible mitigation measures indicates that there is no simple way of addressing the issues. There are multiple reasons for this, including:

1. Knowledge and awareness of issues and measures
   No individual or group possesses the full overview of the underlying technologies, their implications, possible impacts and regulatory options. The complexity of the way in which AI and society interact precludes a centralised solution.
2. Distribution of responsibilities
   AI finds its place in an existing network of responsibilities. All of the AI ethics stakeholders are already part of existing responsibilities, many of which have direct bearing on their work with AI. Many of these will need to be developed and modified to be sensitive to the specific challenges of AI.
3. Technical progress
   SHERPA has provided a snapshot of current and likely future uses of SIS. It is to be

expected that the underlying technologies will continue to develop and new issues will arise due to new technical capabilities and applications.

The appropriate way of dealing with ethics and human rights issues of AI therefore

1. Should involve all stakeholders of AI research and development
2. Should be based on multidisciplinary and multi-sector approaches
3. Requires the coordination of the various actors and stakeholders
4. Needs to balance local, regional and international aspects

It therefore seems appropriate to use the **metaphor of an ecosystem**. In order to ensure that AI is developed and used to promote human flourishing, an ecosystem of stakeholders and their responsibility should be promoted that is conducive to such flourishing. The ecosystem contains the various inhabitants, but it also depends on the natural / technical and social environment. Members of an ecosystem influence one another in non-linear and non-trivial ways and the survival and success of the ecosystem depends on its inhabitants, but also on the maintenance of the environment in which it finds itself.

An ecosystem of AI already exists. In order to address the issues, this ecosystem needs to develop and maintain the capacity to understand and proactively deal with these issues. In line with one of the SHERPA publications (currently under review), it is suggested that a suitable way to characterise this ecosystem would be as an ecosystem that is conducive to human flourishing. The concept of human flourishing has been introduced into the ethics of AI debate previously and offers a good way of highlighting the way in which technology can be developed and used with a high degree of sensitivity towards ethics and human rights.

We therefore suggest that the SHERPA recommendations should be geared towards establishing an ecosystem of AI for human flourishing. This provides a simple way of characterising the intended outcomes and leaves freedom to develop more detailed recommendations for particular stakeholders. It also allows the SHERPA consortium to focus on those aspects where it has particular strengths and expertise.

Questions to be Considered

Based on this analysis, there are a number of questions we can ask of stakeholders (including members of the consortium) that will allow SHERPA to progress on its pathway towards developing useful recommendations. These questions include:

1. Is the overall narrative in this document plausible and does it add value to the AI ethics discourse and the SHERPA project? How could it be improved? Which aspects need more / less emphasis?

2. What are the most important steps to
    a. Establish this ecosystem?
    b. Prepare pathways towards acceptance of the ecosystem?
    c. Maintain and stabilise the ecosystem?
3. What are the biggest gaps in the ecosystem at the moment?
4. What needs do specific AI ethics stakeholders (including you) have in navigating the ecosystem?
5. What can a project like SHERPA contribute to the development and acceptance of the ecosystem?

Please provide answers to these questions using this link:

http://bit.ly/SHERPA-Recommendations-Feedback

Appendices

Appendix 1: Ethical Issues, according to Delphi Study round 2 questionnaire

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*
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)*
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)*

Appendix 2: Mitigation Measures, Delphi Study round 2

Potential Regulatory Measures

International and Regional Measures

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

National Measures

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
Other regulatory measures/cross-over measures
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

Potential Technical Measures
Methodologies for systematic and comprehensive testing of AI-based systems (including 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)

Other Potential Measures

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)