D5.4: Multi-stakeholder Strategy and Practical Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributors	Anais Resseguier, Trilateral Research
	Anais.resseguier@trilateralresearch.com
	Philip Brey, Brandt Dainow, Anna Drozdzewska, University of Twente
Other contributors	Nicole Santiago, David Wright, Trilateral Research
Annex contributors	See Annexes
Reviewers:	Rowena Rodrigues, Trilateral Research Konrad Siemaszko, Helsinki Foundation for Human Rights

Due date	28 February 2021
Delivery date	26 February 2021
	Embargoed for publication until 1 September 2021
Type	Report
Dissemination level	PU = Public
Keywords	artificial intelligence, robotics, ethics, ethics by design, strategy for ethical AI and robotics, ethics tools

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

After having conducted extensive ethical and legal studies and consulted a wide range of stakeholders on AI and robotics and the social implications of these technologies, the SIENNA project has developed practical ethical guidance documents for AI and robotics. Considering the numerous high-level guidance documents developed in the field over the last couple of years, including the High-Level Expert Group on Artificial Intelligence (AI HLEG), the SIENNA project, together with numerous stakeholders, identified the *need for more operational guidance documents*. The main objective of these documents is to provide practical ethics guidance to ensure AI and robotics are developed, deployed and used in ways that respect key ethical principles and values. SIENNA has developed six different guidance documents. The first guidance document is a Multi-stakeholder Strategy for Ethical AI and Robotics which is a comprehensive multi-stakeholder framework to ensure ethical AI and robotics. The others are:

- Ethics by Design and Ethics of Use for AI and Robotics
- Industry Education and Buy-In for AI Ethics
- Research ethics guidelines for Artificial Intelligence
- AI Ethics Education, Training and Awareness Raising
- Ethics at Attention to Context: Recommendations for AI ethics

Acknowledgments

We acknowledge the inputs received from various stakeholders. See the Acknowledgement section p. 10.

Document history

Version	Date	Description	Reason for change	Distribution
V0.1	5 February 2021	First draft for Quality Assurance	Development of report	Quality Assurance reviewers
V0.2	26 Feb 2021	Final version for submission	Finalisation	European Commission
V0.3	4 March 2021	Final version for submission	Section duplication removed	European Commission
V0.4	7 July 2021	Revised final version	Revision requests from reviewer	European Commission

Information in this report that may influence other SIENNA tasks

Linked task	Points of relevance
Task 6.4: Obtain buy-in for SIENNA proposals from EU and international institutions	Task 6.4 will disseminate and/or exploit the recommendations advanced here through liaising with relevant organisations and experts to obtain buy-in for them.
Task 6.6: Formulate a sustainability plan	Task 6.6 will ensure sustainability of the ethics proposals developed in D5.4
Task 6.3: Methods for translating ethical analysis into instruments for the ethical development and deployment of emerging technologies	Task 6.3 will draw from methods used in T5.4 and related outputs to develop generalised methods to develop ethics tools for the development and deployment of emerging technologies.

Table of contents

Executive summary

After extensive ethical and legal studies and consultation with a wide range of stakeholders, including the public, on AI and robotics and the social implications of these technologies, the SIENNA project has developed practical ethical guidance documents for AI and robotics. Considering the numerous high-level guidance documents developed in the field over the last couple of years, including the High-Level Expert Group on Artificial Intelligence (AI HLEG), the SIENNA project, together with numerous stakeholders, identified the *need for more operational guidance documents*. The main objective of these documents is to provide practical ethics guidance to ensure AI and robotics are developed, deployed and used in ways that respect key ethical principles and values. SIENNA developed six different guidance documents. These documents, i.e., the substantive content of this deliverable, are in the Annexes of this report. We developed these documents as individual texts so that they could easily be shared to different target audiences. The main text of the deliverable reports on the process undertaken to develop these documents.
These documents are:

- Annex 1: Multi-stakeholder Strategy for Ethical AI and Robotics
- Annex 2: Ethics by Design and Ethics of Use for AI and Robotics
- Annex 3: Industry Education and Buy-In for AI Ethics
- Annex 4: Research ethics guidelines for Artificial Intelligence
- Annex 5: AI Ethics Education, Training and Awareness Raising
- Annex 6: Ethics at Attention to Context: Recommendations for AI Ethics

Annex 1 - **Multi-stakeholder Strategy for Ethical AI and Robotics** which presents an extensive framework for the ethical development, deployment, and use of AI and robotics technologies. The strategy addresses all stakeholders in society, particularly researchers, users, regulators, educators, the media, and the general public. All have a role in bringing about ethical AI and robotics. It identifies a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that promotes adhesion to ethical principles and protection of social values. Within this general strategy, we pay particular attention to methods and procedures for ethical research and innovation (R&I) in AI and robotics.

Annex 2 - **Ethics by Design and Ethics of Use for AI and Robotics** - offers guidance for an Ethics by Design and Ethics of use approach when developing and using AI-driven systems, including robots. It offers a way to include ethical principles and procedures into the design and development processes. It aims at ensuring ethical problems are not generated in the first place by using ethically-focused activities during the design, development and deployment phases of a project.

Annex 3 - **Industry Education and Buy-In for AI Ethics** - outlines proposals for encouraging industry to adopt ethical AI and robotics in design, development, sales, staffing and use. The central mechanisms presented are a set of certifications for products and people, the development of a certification business eco-system, and the use of proven market mechanisms to build customer demand for certified products and people. This document recommends that AI and robotics products should be certified as ethical upon creation (a “product certification”), and that their deployment in a working environment should also be certified as maintaining that ethical status (an “installation certification”). We further recommend the development of professional certification for staff appropriate to their roles.

Annex 4 – **Research Ethics Guidelines for Artificial Intelligence** - proposes a set of research ethics guidelines for artificial intelligence (AI) and robotics and discusses how these can serve as a basis for stand-alone guidance protocols for R&D in AI and robotics. It also discusses how they can be incorporated into broader research ethics frameworks for computer and science and for other disciplines. This proposal covers twenty-seven research ethics guidelines that were grouped into six categories, under the categories of human agency, privacy & data governance, fairness, social and environmental well-being, accountability & oversight, and transparency, and "special topics" guidelines for specific techniques, products, and application domains in the AI and robotics field. We also propose an Ethics by Design approach, which provides a comprehensive way of integrating ethical guidelines and criteria into design methodologies. After presenting these proposals, we propose how they can be used as a basis for stand-alone research ethics protocols for AI and robotics, and next how they can be integrated into broader research ethics frameworks for computer and information sciences, and then for research ethics frameworks that span multiple fields.

Annex 5 - **AI Ethics Education, Training and Awareness Raising** - outlines proposals for developing awareness and skills in ethical AI and robotics systems within society. It covers the various needs, and proposes solutions, within Higher Education curricula, industry training, product certification and the means by which to raise awareness within the general public. We sought to minimise risk and increase the chances of success by limiting our proposals to established methodologies and proven approaches. With regard to Higher Education, we outline the case for integration of ethical education within Computer Science and Engineering, and covering the needs of other disciplines, such as law and business, to understand relevant ethical AI concerns within their speciality. Finally, we provide a best practice case study of a module taught at Harvard University which exemplifies the ideal format we recommend. Moving to commercial industry, we briefly outline policy proposals for the development of commercial certification schemes for products and people. We then discuss methods by which these schemes can obtain popular support, such that it becomes profitable for AI developers to pursue such certifications. This document concludes with a brief discussion of methods to promote general public awareness of ethical AI and its issues.

Annex 6 - **Ethics at Attention to Context: Recommendations for AI Ethics** - shows that current AI ethics guidance and initiatives tend to be dominated by a principled approach to ethics. Although this brings value to the field, it also entails some risks, especially in relation to the abstraction of this form of ethics that makes it poorly equipped to engage with and address deep socio-political issues and practical impacts. Thus, this document seeks to complement the existing principled approach to ethics with an approach to ethics as attention to context and relations. It makes practical recommendations to promote ethical AI by drawing from an approach to ethics as attention to context.

List of tables

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
CSO	Civil Society Organisations (CSOs)
D	Deliverable
DH-BIO	Committee on Bioethics (CoE)
EC	European Commission
EP	European Parliament
EU	European Union
GDPR	General Data Protection Regulation
ICT	Information and Communications Technology
R&I	Research and Innovation
RRI	Responsible Research and Innovation
STOA	Science and Technology Options Assessment Panel of the European Parliament
T	Task
WP	Work Package

Table 1: List of acronyms/abbreviations

Glossary of terms

Term	Explanation
Artificial intelligence (AI)	The science and engineering of machines with capabilities that are considered intelligent (i.e., intelligent by the standard of *human* intelligence).
Autonomy	The value of a person's ability to decide and act on her own authentic desires and preferences, without being unduly influenced, coerced or manipulated by others.
Ethical AI and robotics	In the context of this strategy, steps to promote and work toward an ethical development, deployment and use of AI and robotics.
Ethics by design	The systematic inclusion of ethical guidelines, recommendations and considerations into design and development processes.
Robotics	The field of science and engineering that deals with the design, construction, operation, and application of robots.

Table 2: Glossary of terms

1. Introduction

This report was developed within the SIENNA project, a European Horizon 2020-funded project[1] focussing on the ethical, legal and social dimensions of three technological areas: (1) human genetics and genomics, (2) human enhancement technologies, and (3) artificial intelligence and robotics. The project conducted extensive analysis of ethical and legal aspects of these technology areas, reviewed their present and expected applications, socio-economic impacts and analysed key concepts and demarcations of the fields and performed studies on the public awareness and acceptance of these areas and of their current coverage by research ethics committees and in ethical codes. Moreover, the project has proposed general ethical frameworks for the three fields.[2] Based on this work and further engagement with stakeholders, SIENNA has developed ethics proposals to promote ethics in the three technology areas and to ensure these technologies are designed and used in ways that respect ethical principles and values. This report presents ethics proposals for AI and robotics as part of Task 5.4: Multi-stakeholder strategy and practical tools for ethical AI and robotics.

2. Operational guidance documents

The scope of T5.4 was amended due to the development, over the last couple of years (especially 2019-2020), of codes, guidelines, and other guidance documents for AI both at the national and international levels.[3] This changed task description responded to feedback from stakeholders who advised us that no additional high-level guiding documents were needed for AI and robotics. Rather, stakeholders recommend providing more practical, detailed and operationalised recommendations to ensure ethical AI and robotics. Thus, Task 4.7 began by proposing a general strategy for ethical AI and robotics and an ethics by design methodology for research and innovation for AI and robotics drawing from AI HLEG requirements and SHERPA guidelines.[4]

Task 5.4 carried forward the work of Task 4.7 and further expanded, detailed and complemented it. Task 5.4 has developed concrete tools to implement proposals formulated in D4.7.[5] In our effort to make these documents as relevant and practical, we developed these tools as standalone documents so that they could be shared to the relevant audiences independently. They are included in the Annex of this Deliverable and include:

- **Annex 1: Multi-Stakeholder Strategy for Ethical AI and Robotics**: this strategy draws from section 2 of D4.7, [6] and is further developed, complemented and specified using inputs from

[1] https://www.sienna-project.eu

[2] For all SIENNA publications, please visit the SIENNA website: https://www.sienna-project.eu/publications/deliverable-reports/

[3] See the list of documents collected as part of D4.3: Tambornino, Lisa, Dirk Lanzareth, Rowena Rodrigues, David Wright, SIENNA D4.3 Survey of REC approaches and codes for Artificial Intelligence & Robotics, August 2019: https://zenodo.org/record/4067990#.YDUBni1Q2u5

[4] Brey, Philip, Philip Jansen, Jonne Maas, Björn Lundgren, Anaïs Resseguier, SIENNA D4.7 An Ethical framework for the development and use of AI and robotics technologies, March 2020, submitted public deliverable from the SIENNA project, awaiting approval from the European Commission. Brey, P., Lundgren, B., Macnish, K. and Ryan, M. (2019). *Guidelines for the development and use of SIS.* Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833.

[5] Brey, SIENNA D4.7, op cit. 2020.

[6] Ibid.

stakeholders in this report. It starts from the recognition that ethical AI and robotics requires the involvement of wide range of actors of society through a diversity of initiatives. It then highlights a series of methods to promote ethical AI and robotics. The Multi-stakeholder Strategy is the umbrella document. All the other documents presented are tools that fit into this general strategy.

- **Annex 2: Ethics by Design and Ethics of Use for AI and Robotics:** This draws from the ethics by design method developed in section 3 and annexes of D4.7[7] and further complements and adapts it.
- **Annex 3: Industry Education and Buy-In for AI Ethics.** This is a new document produced in T5.4, building upon the Multi-stakeholder strategy. It outlines proposals for encouraging industry to adopt ethical AI and robotics in design, development, sales, staffing and use.
- **Annex 4: Research ethics guidelines for Artificial Intelligence.** This is a new document produced in T5.4 building on the proposal made to the European Commission for an ethics self-assessment process for AI projects funded by the EU. It proposes a set of research ethics guidelines for AI and robotics and discusses how these can serve as a basis for stand-alone guidance protocols for R&D in AI and robotics.
- **Annex 5: AI Ethics Education, Training and Awareness Raising.** This is a new document produced in T5.4, building upon the Multi-stakeholder strategy. It outlines proposals for developing awareness and skills in ethical AI and robotics systems within society. It covers the various needs, and proposes solutions, within Higher Education curricula, industry training, product certification and the means by which to raise awareness within the general public.
- **Annex 6: Ethics as Attention to Context: Recommendations for AI Ethics.** This is a new document produced in T5.4, building upon the Multi-stakeholder strategy. It shows that current AI ethics guidance and initiatives tend to be dominated by a principled approach to ethics that has values but also limitations. It seeks to complement this approach with an approach to ethics as attention to context and relations.

3. Process to develop SIENNA proposals for ethical AI and Robotics

The SIENNA proposals for ethical AI and robotics were developed by Trilateral Research and the University of Twente based on previous work conducted in SIENNA and taking into consideration stakeholder input and public views[8] gathered throughout the project. Drafts of these proposals were reviewed by various stakeholders in a number of meetings, workshops, and written feedback process. SIENNA attempted to reach out to a wide range of stakeholders. These included:

- Written feedback on the Multi-Stakeholder strategy by SIENNA partners in August 2020
- Online discussion on the Multi-Stakeholder strategy with stakeholders on 24 August 2020
- SIENNA online workshop on Multi-Stakeholder Strategies for Ethical AI on 8-9 September 2020
- Public commentary via the SIENNA website 11-25 January 2021

[7] Ibid.

[8] SIENNA explored public acceptance and awareness of AI and robotics in two different studies, a qualitative one and a quantitative one: Hamly, Rebecca, SIENNA D4.5 Public views on artificial intelligence and robots across 11 EU and non-EU countries, Aug 2019, https://doi.org/10.5281/zenodo.4068220; Kantar, SIENNA D4.6: Qualitative research exploring public attitudes to AI and robotics, Aug 2019, https://doi.org/10.5281/zenodo.4081247

- SIENNA online consortium and Advisory Board meeting, 14 January 2021

The feedback received was progressively integrated into the drafts of SIENNA proposals by Trilateral Research and the University of Twente. The Acknowledgements section below lists all individuals who have provided input.

4. Conclusion

This deliverable presents SIENNA proposals to ensure consideration and practical implementation of ethical aspects in the development, deployment and use of AI and robotics. It responds to an identified need in both the technical and ethics community working on AI and robotics for more operational guidance for ethical AI and robotics. In an effort to make our proposals effective and impactful, we developed these as stand-alone documents focused on specific methods to promote ethics in these technology fields and targeted at particular stakeholders to whom these guidance documents will be the most relevant and useful.

The proposals and tools presented here can be used by stakeholders in their own practice and help reduce potential harms caused by AI and robotics and bring beneficial outcomes. SIENNA hopes that they will lead to practical uptake in various organisations and institutions developing, deploying and using AI. Finally, as the multi-stakeholder strategy has shown, there is a wide variety of actors and methods required for ethical AI and robotics and SIENNA hopes that the tools developed here are further complemented with other ethics tools in the future.

Acknowledgements

Several stages of review of the SIENNA proposals took place in written and oral forms, with a diversity of experts and stakeholders.

Type of reviewer	Reviewer	Affiliation	What was reviewed
WRITTEN REVIEW – August 2020			
SIENNA partner	Robert Gianni	University of Maastricht	Multi-stakeholder strategy
SIENNA partner	Lisa Tambornino	EUREC Office GUG	Multi-stakeholder strategy
SIENNA partner	Javier Valls Prieto	University of Granada	Multi-stakeholder strategy
External stakeholder	Ana Valdivia	King's College London	Ethics by design framework
ONLINE PANEL DISCUSSION ON MULTI-STAKEHOLDER STRATEGY – 24 August 2020			
External stakeholder	John Havens	IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	Multi-stakeholder strategy
External stakeholder	Rene von Schomberg	European Commission	Multi-stakeholder strategy
External stakeholder	Catherine Tessier	ONERA – French Aerospace Lab	Multi-stakeholder strategy
SIENNA MULTI-STAKEHOLDER STRATEGIES FOR ETHICAL AI WORKSHOP – 8-9 September 2020			
External stakeholder	Norberto Andrade	Facebook	SIENNA proposals from D5.4
External stakeholder	Natalie Bertels/ Ernestina Menasalvas	BDVA	SIENNA proposals from D5.4
External stakeholder	Declan Brady	Council of European Professional Informatics Societies	SIENNA proposals from D5.4
External stakeholder	Cansu Canca	AI Ethics Lab	SIENNA proposals from D5.4
External stakeholder	Raja Chatila	Institute of Intelligent Systems and Robotics (ISIR)	SIENNA proposals from D5.4

Type of reviewer	Reviewer	Affiliation	What was reviewed
External stakeholder	Leonardo Calini/Jochem de Groot	Microsoft	All SIENNA proposals from D5.4
External stakeholder	Dafna Feinholz	UNESCO	All SIENNA proposals from D5.4
External stakeholder	Iban Garcia del Blanco	Member of European Parliament. Group of the Progressive Alliance of Socialists and Democrats	All SIENNA proposals from D5.4
External stakeholder	Chiara Giovannini	ANEC	All SIENNA proposals from D5.4
External stakeholder	Alexei Grinbaum	CEA-Saclay	All SIENNA proposals from D5.4
External stakeholder	Anna Jobin	Swiss Young Academy	All SIENNA proposals from D5.4
External stakeholder	Albena Kuyumdzhieva	Ethics and Research Integrity Sector at European Commission	All SIENNA proposals from D5.4
External stakeholder	Katherine Mayes	TechUK	All SIENNA proposals from D5.4
External stakeholder	Clara Neppel	IEEE	All SIENNA proposals from D5.4
External stakeholder	Johnny Soraker	Google	All SIENNA proposals from D5.4
External stakeholder	Bernd Stahl	De Montfort University (Coordinator of SHERPA project)	All SIENNA proposals from D5.4
SIENNA partner	Lisa Tambornino,	EUREC Office GUG	All SIENNA proposals from D5.4
External stakeholder	Krista Varantola	ALLEA	All SIENNA proposals from D5.4
External stakeholder	Diane Whitehouse	EHTEL	All SIENNA proposals from D5.4
SIENNA CONSORTIUM MEETING – 14 January 2021 (online)			
SIENNA scientific advisory board	Marc Coeckelbergh	University of Vienna	All SIENNA proposals from D5.4
SIENNA scientific advisory board	Roger Brownsword	TELOS, King's College London and Bournemouth University	All SIENNA proposals from D5.4

Type of reviewer	Reviewer	Affiliation	What was reviewed
SIENNA scientific advisory board	Allen Buchanan	University of Arizona	All SIENNA proposals from D5.4
PUBLIC CONSULTATION – 11-25 January 2021			
External Stakeholder	Sergio Guida	Science4People-Design Lab Pro / Legal Design ITLab Pro	Ethics as Attention to Context. Recommendations for AI Ethics
External Stakeholder	Xaroula Kerasidou	Lancaster University	Ethics as Attention to Context. Recommendations for AI Ethics
External Stakeholder	Ana Valdivia	King's College London	Ethics as Attention to Context. Recommendations for AI Ethics
External Stakeholder	French Transhumanist Association - Technoprog		SIENNA proposals from D5.4

sienna.

Annex 1 - Multi-Stakeholder Strategy for Ethical AI and Robotics

Multi-stakeholder Strategy for Ethical AI and Robotics

Annex 1 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Philip Brey, University of Twente
	p.a.e.brey@utwente.nl
Other contributors	Anais Resseguier, Trilateral Research
Reviewers	Elisabeth de Castex, Fondation pour l'innovation Politique Ebru Dogan, Institut Vedecom
Date	February 2021
Type	Annex to D5.4 deliverable
Dissemination level	PU = Public
Keywords	AI ethics; Robot ethics; Multi-Stakeholder strategy; ethical tools; technology ethics

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Table of contents

1. Introduction

1.1 Background

This report was developed within the SIENNA project, a European Horizon 2020-funded project on the ethical and human rights dimensions of emerging technologies.[1] A major focus of the SIENNA project is on the ethical and human rights aspects of AI and robotics. We performed extensive studies on AI and robotics: a state of the art, social and economic impacts, ethical aspects, legal and human rights context, currently existing ethical codes and guidelines regulating these technologies, and finally, a public awareness and acceptance study.[2] Based on our previous studies, we hereby propose an extensive framework for the ethical development, deployment, and use of AI and robotics technologies.

1.2 Objectives

This report proposes a Multi-Stakeholder Strategy for ethical AI and robotics. It includes a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that promotes respect for ethical principles and protection of social values. This strategy addresses all stakeholders in society, particularly researchers, developers, technology users, regulators, educators, the media, and the general public. All have a role in ethical AI and robotics. Within this general strategy, we pay particular attention to methods and procedures for ethical research and innovation (R&I) in AI and robotics. In R&I, major decisions are made about what technological solutions to develop and which ones not to develop, and R&I often comes with prescriptions about deployment and use as well. However, we will also pay attention to methods for ethical deployment and use, and to the role of organisations that market and use AI and robotics technologies, as well as policy makers, regulators and educators.

This document proposes an overall strategy to promote ethical AI and robotics. It starts with an identification of relevant stakeholders and categories of methods for obtaining ethical AI & robotics.

[1] See https://www.sienna-project.eu/.

[2] See SIENNA reports: Jansen, Philip, Broadhead, S., Rodrigues, R., Wright, D., Brey, P., Fox, A, and Wang, N., SIENNA D4.1: State-of-the-art Review: Artificial Intelligence and robotics, 2019, https://doi.org/10.5281/zenodo.4066571. Rodrigues, Rowena, Siemaszko, K., and Warso, Z., SIENNA D4.2: Analysis of the legal and human rights requirements for AI and robotics in and outside the EU, 2019, https://doi.org/10.5281/zenodo.4066811. Lisa Tambornino, Lanzerath, D., Rodrigues, R, and Wright, D., SIENNA D4.3: Survey of REC approaches and codes for Artificial Intelligence & Robotics 2019, https://doi.org/10.5281/zenodo.4067989. Philip Jansen, Brey, P, Fox, A., Maas, J., Hillas, H., Wagner, N., Smith, P, Ouoch, I., Lamers, L., van Gein, H., Resseguier, A., Rodrigues, R., Wright, D., and Douglas, D., SIENNA D4.4: Ethical Analysis of AI and Robotics Technologies, 2020, https://doi.org/10.5281/zenodo.4068082. Rebecca Hamlyn, SIENNA D4.5 Public views on artificial intelligence and robots across 11 EU and non-EU countries, 2020, https://doi.org/10.5281/zenodo.4068220. Kantar (Public Division), SIENNA D4.6: Qualitative research exploring public attitudes to AI and robotics, 2019, https://doi.org/10.5281/zenodo.4081246. Philip Brey, Jansen, J., Maas, J., Lundgren, B., and Resseguier, A., SIENNA D4.7 An Ethical framework for the development and use of AI and robotics technologies, 2020, submitted public deliverable from the SIENNA project, awaiting approval from the European Commission, https://www.sienna-project.eu/digitalAssets/801/c_801912-l_1-k_d4.7_ethical-framework-for-ai--robotics_v2.pdf .

It then proceeds to discuss the categories of methods in detail and concludes by highlighting how methods can be developed further and how stakeholders can be motivated to use them.

The role of ethical principles

This report does not develop or propose general ethical principles or guidelines for AI and robotics. By now, there is already enough convergence, in our opinion, on ethical principles for AI and robotics. During 2019, in particular, many countries and international organisations proposed general ethical guidelines for AI. Notably, 2019 saw the Ethics Guidelines for Trustworthy AI from the High-Level Expert Group on Artificial Intelligence (AI HLEG)[3], the Recommendation of the Council on Artificial Intelligence of the OECD[4], the guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers[5], and the Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence China's Ministry of Science and Technology[6].

As was observed in a recent study by the EU funded H2020 project SHERPA[7], co-authored by some of the people behind this study, there is a remarkable convergence between these recent sets of ethical guidelines. Although formats and wordings may differ, the three main guidelines (AI HLEG, OECD, and IEEE – see above) are in agreement on content. Most importantly they agree on nine key ethical principles for AI and robotics: *privacy, autonomy, freedom, dignity, safety and security, justice/fairness, responsibility/accountability, well-being (individual, societal and environmental) and transparency*. In addition, none of these documents proposed major principles outside of this list. Even the Chinese guidelines converge remarkably with more "Western" guidelines: they by and large reflect these ethical principles as well.

1.3 The Multi-stakeholder Strategy for Ethical AI and Robotics in brief

A set of ethical guidelines or principles is only one component of a strategy for ethical AI & robotics. It could provide some direction to activities, but only in a very general sense. Many more elements need to be in place to achieve the objective of ethical AI & robotics. Consider, for example, the development of AI & robotics technologies. Developers and other stakeholders involved, like most people, have certain ethical views and moral leanings that they respect.[8] This may colour the development process. When given a list of ethical principles for AI, some developers may endorse them and attempt to adhere to them in their activities. This may point developers to actively focus on ethics during the development process. A set of principles, nevertheless, may not always be successful. Programmers could easily fail to properly interpret or implement them, whether this is due to a lack of training in ethics, lack of knowledge of how to apply ethical principles in technology development, lack of support from management, lack of inclusion of ethics criteria in quality assessment frameworks or corporate

[3] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[4] OECD, *Recommendation of the Council on Artificial Intelligence*, 2019. Retrieved on 8-3-2020 at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[5] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/

[6] Ministry of Science and Technology of China, *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence,* 2019. A translation can be found at: https://perma.cc/V9FL-H6J7.

[7] Ryan, M., Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van Der Puil (2019). *Ethical Tensions and Social Impacts.* Deliverable D 1.4 of the SHERPA project. https://doi.org/10.21253/DMU.8397134

[8] Miller, Catherine, and Rachel Coldicutt, "People, Power and Technology: The Tech Workers' View," London, Doteveryone, 2019, p. 16. https://doteveryone.org.uk/report/workersview.

social responsibility strategies, ignorance as to how the technology may violate the principles, or other reasons. Much more is needed to make stakeholders motivated and competent in the incorporation of ethical considerations in their practices, and to support stakeholders in collaborative practices towards this shared objective.

A sound strategy for ethical AI & robotics should in our view do three things:

- *Identify relevant stakeholders*
- *Identify methods that these stakeholders can use to contribute to ethical AI & robotics*
- *Propose ways these methods could be made available to these stakeholders, and ways to promote their use.*

An overall strategy is proposed in this report. Such a strategy is, in our view, a first step towards realising ethical AI & robotics. A second step is the successful implementation of the strategy by relevant stakeholders.

We will now proceed to identify the most relevant stakeholder categories, and then propose relevant methods for each of them, including some shared methods that apply to different actor categories. We end with a brief discussion of ways to make the methods available to stakeholders, and ways to motivate them. This strategy is complemented by a set of "practical tools" presented in the other annexes of Deliverable 5.4 – these aim at operationalising, details and making more practical the recommendations in this strategy. These tools include:

- **Annex 2: Ethics by Design and Ethics of Use in AI and Robotics**
- **Annex 3: Industry Education and Buy-In for AI Ethics**
- **Annex 4: Research ethics for AI**
- **Annex 5: AI ethics education, training and awareness raising**
- **Annex 6: Ethics as Attention to Context: Recommendations for AI Ethics**

1.4 Scope and limitations

This strategy adopts the nine key ethical principles listed above (section 1.2) as a starting point for ethical guidance. Specifically, given that this is a European Union funded project, we will take on, with minor adaptations, the version of these principles as developed by the High-Level Expert Group on AI. That is, we will adopt the ethics guidelines for trustworthy AI of the AI HLEG as our guiding set of principles, specifically its seven ethics requirements for trustworthy AI in which these nine principles are contained: *Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; and Accountability*. Because of the strong similarities between these guidelines and others used outside the European union, we expect this proposal to have relevance and applicability outside the European Union too.

Our main objective in this report, however, is to operationalise ethical guidelines to make them directly usable by stakeholders in particular practices. This is what much of this report focuses on. This strategy in itself is not sufficient to offer ethical guidance for particular products and applications, or specific contexts of use. It needs to be complemented by the practical guidelines presented in the other annexes of D5.4. More detailed guidelines will also be needed to address such issues, for example, ethical guidelines for unmanned aerial vehicles, or for healthcare applications of AI, or for predictive data analytics techniques.

It is necessary to clarify the meaning of the term "ethical" in the concept of "ethical AI and robotics" used in this strategy. This concept should not be understood as a guarantee of AI and robotics being strictly ethical, i.e., avoiding any harm and leading to beneficial outcomes. In other words, following these recommendations is not necessarily sufficient to ensure perfectly beneficial AI. A product or technology in and of itself cannot be said to be "ethical". Rather, this strategy underlines that, at all stages of the development, deployment and use of a product or technology, it is essential to pay attention to ethical aspects and to put in place mechanisms to ensure these aspects are taken into account. In that sense, "ethical AI and robotics" in the context of this strategy means ***steps to promote and work toward an ethical development, deployment and use of AI and robotics.*** Hence, this strategy is about a *method,* rather than about giving an "ethical" label to a product or technology.

Additionally, ethical AI and robotics means, at least, avoiding harm in the development, deployment and use of these technologies; at best, the pursuit of beneficial outcome, such as human well-being and flourishing or environmental protection. This strategy seeks to ensure that this primary objective persists throughout the different stages of development and use of AI and robotics and is not set aside by other pressures, especially political and financial interests.

2. Stakeholders

The following stakeholder categories are most relevant for our purposes. They have been selected on the basis of having influence on how AI & robotics technologies are developed, used, and perceived, and thereby on what their impacts and ethical aspects are:

1. AI & robotics developers
2. AI & robotics development support organisations
3. Organisations that deploy and use AI & robotics technology
4. Governance and standards organisations
5. Educational and media organisations
6. Civil society organisations and the general public
7. Organisations and units working on ethics and social impacts

We will now discuss them in turn.

2.1. AI & robotics developers

Within this broad category, we can make some further distinctions. At the organisational level, developers include firms that develop AI & robotics technologies and research institutes (universities and other research performing organisations) that engage in research and innovation in AI & robotics. At the intra-organisational level, there are various units within these institutions that are involved in the planning, support and carrying out of R&I activities. At the individual level, there are also professionals in various roles (e.g., IT project manager, IT director, hardware technician, professor in robotics) that are stakeholders in AI & robotics development.

2.2. AI & robotics development support organisations

These are organisations that support R&I activities of AI & robotics firms and research institutes. These include business and industry organisations (also known as trade organisations): organisations that support companies in a certain sector; chambers of commerce; research funding organisations;

investment banks and other investors and funders; associations of universities and research institutes; science academies and associations of science academies; professional organisations for the AI & robotics fields; advisory and consultancy firms for companies and research institutes.

2.3. Organisations that deploy and use AI & robotics technology

These are private and public organisations that use AI & robotics. Its usage can be intended to improve or support various organisational functions, including operations, finance, marketing, human resources, customer service, regulation, etc. Within these organisations, one can furthermore define various units and professional roles associated with the deployment and use of AI systems within or by the organisation, such as information technology managers, database administrators, and development operations engineers. Note, some organisations are simultaneously developers and users of AI & robotics systems. For example, technology companies like Apple and Google develop AI technologies, and use them within their own organisation.

2.4. Governance and standards organisations

These are organisations involved in developing, implementing or enforcing policies, standards and guidelines, specifically those regarding the development, deployment and use of AI & robotics technologies. Organisations also make policies and guidelines for themselves. These are not our concern here. This category refers to organisations that develop or implement guidelines, policies, regulations and standards for others. This includes, first of all, national, local and supranational governments, and government-instituted or -supported advisory and regulatory bodies. They also include intergovernmental organisations such as the United Nations, the Council of Europe, and the World Health Organization WHO). Also included in this category are national and international standards (e.g., ISO, IEEE), certification, quality assurance, accreditation and auditing organisations. Policies, standards and guidelines can also be issued by many of the AI & robotics development support organisations discussed earlier.

2.5. Educational and media organisations

Educational institutes and media organisations both have a significant role, albeit a quite different one, in shaping people's knowledge and understanding of AI & robotics, the ethical issues associated with them, and the ways in which these ethical issues can be addressed. Educational organisations, from elementary school to postgraduate education, provide the major vehicle by which individuals acquire knowledge, skills and insights regarding AI & robotics, their impacts on society, their ethical aspects, and ways to address ethical issues in their profession. Of course, it is not only educational organisations that provide education and training. Companies may, for example, organise their own in-house trainings as well. Media organisations have a large role in generating public awareness and understanding of AI & robotics and the ethical issues raised by them and therefore should also be recognized as stakeholders with respect to ethical AI & robotics.

2.6. Civil society organisations and the general public

Civil Society Organisations (CSOs) are non-governmental, not-for-profit organisations that represent the interests and will of individuals. They may be based on cultural, political, ethical, scientific, economic, religious or philanthropic considerations. They include civic groups, cultural, groups, consumer organisations, environmental organisations, religious organisations, political parties, trade unions, professional organisations, non-governmental policy institutes, activist groups, and several other kinds. Many CSOs want to, and should, have a role in public policy or influence the way that

organisations function in which they have an interest. For some of them, the development and use of AI is a concern, and as a result, these organisations may function as agents with respect to public policy and the actions of relevant other organisations. The general public, finally, can also perform as a stakeholder and should be considered as such by policy makers and other actors involved in the development, deployment and use of AI and robotics. The general public can be consulted through public opinion surveys and studies and studied through voting patterns, consumer purchases, and use or non-use of AI & robotics products and services.

2.7. Organisations and units working on ethics and social impacts

Finally, it is important to mention organisations and units working on ethics and social impacts. These may be part of the various kinds of organisations and units listed above. These include ethics research units, ethics policy units, ethics officers, research ethics committees, integrity offices and officers, corporate social responsibility teams and officers, technology impact assessors, ethics educational programmes, ethics advisory bodies, and national and international ethics committees. Although all of the listed stakeholders above have a role in ensuring ethical standards and practices, ethics organisations and units have a particular responsibility in that regard. This category also includes research institutes working on the ethics and social sciences of technology, especially AI and robotics. These are essential to closely follow technological developments and their short, medium and long terms impacts on the society. Considering the novelty and complexity of AI and robotics, it is necessary to conduct in-depth studies on ethical and social impacts of these technologies on the society and identify transformations that may remain invisible without the tools of ethics and social sciences. There still remain many unknowns and uncertainties regarding the ethical and social impacts of these technologies. The resources of ethics and social sciences are much needed to lift these and come to a better and understanding of the long-term impact on society in general and on particular groups, and to mitigate negative implications.

Finally, it is essential that all these different stakeholder groups include a diversity of profiles. Studies have shown that both fields of AI and AI ethics tend to be overly dominated by white males[9]. This is an issue that needs to be addressed as it leads to the omission of the needs and interests of other groups, especially vulnerable and/or underrepresented people. This is a particularly problematic when it comes to issues of biases, inequalities and injustice. To address this issue, women, people of colour, and other marginalised and vulnerable groups, also need to be involved to work toward ethical AI and robotics.

3. Methods

In the context of this report, methods are means by which stakeholders can take into account ethical considerations and implement ethical guidelines. Our identification of methods for ethical AI & robotics builds on earlier proposals of the AI HLEG (2019) and IEEE (2019).[10] Both reports propose

[9] Sarah Myers West, Whittaker, M., and Crawford, K., "Discriminating Systems. Gender, Race, and Power in AI", AI Now Institute, 2019. https://ainowinstitute.org/discriminatingsystems.html

[10] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE), *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/

methods for the implementation of ethical guidelines in relation to different stakeholders. The AI HLEG distinguishes between what they call technical and non-technical methods, both of which apply to all stages of the development and use lifecycle of AI systems. Technical methods include ethics by design methods, explanation methods for transparency, methods of building system architectures for trustworthiness, extensive testing and validation, and the definition of quality of service indicators. Non-technical methods include regulation, codes of conduct, standardization, certification, accountability via governance frameworks, education and awareness to foster an ethical mindset and sensitivity, stakeholder participation and social dialogue, and diverse and inclusive design teams.
The IEEE (2019) report has a chapter on "methods to guide ethical research and design" for researchers, technologists, product developers and companies, and a chapter on policies and regulations by governing institutions and professional organisations.[11] In its methods for ethical research and development (R&D) chapter, it considers both individual and structural approaches, and distinguishes between three overall approaches: interdisciplinary education and research, corporate practices on AI & robotics, and responsibility and assessment. In its policy chapter, the IEEE advocates methods such as the founding of national policies and business regulations for AI on human rights approaches, the introduction of support structures for the building of governmental expertise in AI and robotics, and the fostering of AI & robotics ethics training in educational programs.

The methods proposed by the AI HLEG and IEEE are partly overlapping and partly complementary. Drawing from them, we propose seven sets of methods for the ethical development and use of AI & robotics[12], for the different classes of stakeholders that were defined earlier:

1. Methods for incorporating ethics into research and development of AI & robotics (aimed at AI & robotics developers and support organisations)
2. Methods to incorporate ethics into the deployment and use of AI & robotics (aimed at organisations that deploy and use AI & robotics technology)
3. Corporate responsibility policies and cultures that support ethical development and use of AI & robotics (aimed at both developers, deployers/users and support organizations)
4. National and international guidelines, standards and certification for ethical AI & robotics (aimed at governance and standards organisations; indirectly affecting developers, deployers/users and support organizations)
5. Education, training and awareness raising for the ethical and social aspects of AI & robotics (aimed at all stakeholders)
6. Policy and regulation to support ethical practices in AI & robotics (aimed at governance and standards organisations; indirectly affecting developers and deployers/users)
7. In-depth ethics and social sciences studies on impacts of AI & robotics (aimed at all stakeholders)

We next discuss these sets of methods in some more detail and relate them to the roles and responsibilities of stakeholders.

3.1. Methods for incorporating ethics into research and development

These are methods to make ethical considerations, principles, guidelines, analyses or reflections a part of research and development processes. They apply to the first stakeholder category identified above: AI & robotics developers. Four main classes of methods fall into this category:

[11] Ibid, pp. 124-139 and pp. 198-210 respectively.
[12] Points 1, 3-6 are taken from the SHERPA development and use guidelines: Brey, P., Lundgren, B., Macnish, K. and Ryan, M. (2019). *Guidelines for the development and use of SIS.* Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833. Point 2 is an added point.

1. Research ethics guidelines and protocols for R&I in AI & robotics
2. Ethical impact assessment methodologies for emerging AI & robotics
3. Ethics by design methodologies for AI & robotics
4. Codes of professional ethics for researchers and developers of AI & robotics technologies

We now discuss them in turn.

1. Research ethics guidelines and protocols for R&I in AI & robotics

Research ethics guidelines and protocols for AI & robotics are ethics guidelines and procedures by which researchers, developers, research ethics committees and ethics officers can ethically assess R&I proposals and ongoing R&I practices. We can differentiate between ethics guidance documents for research ethics committees and ethical checklists, assessments or guidance documents for developers. These guidelines can be used to improve R&I plans and practices to make them more ethical. As of the time of the writing of this report (February 2021), few research ethics guidelines and protocols specifically for AI and robotics were in existence (see our report D4.3 Survey of REC approaches and codes for Artificial Intelligence & Robotics).[13] While there is an abundance of general ethical guidelines for AI and robotics, few specifically focus on R&I practices and on the role of research ethics committees. The SIENNA project developed its own proposals of guidelines for research ethics committees (D5.1 Report documenting elements to open and complement operational guidelines for research ethics committees, in progress) and a research ethics protocol specifically focused on AI projects (Annex 4 of D5.4).

2. Ethical impact assessment methodologies for emerging AI & robotics

Ethical impact assessment (EIA) methodologies are methods for assessing present and potential future impacts of emerging technologies, including specific products and applications, and identifying ethical issues associated with these impacts. EIA, in short, is an approach for assessing not only present but also potential future ethical issues in relation to a technology. EIA, in its current form, was developed within the EU-funded FP7 SATORI project.[14] It has also been developed into a CEN pre-standard.[15] EIA is not just a method for AI & robotics developers, but can also be used, amongst others, by government agencies and bodies to support technology policy, and by research funding organisations to help set priorities in research funding.

3. Ethics by design methodologies for AI & robotics

Ethics by design methodologies for AI & robotics are methods for incorporating ethical guidelines, recommendations and considerations into design and development processes. They fill a gap that exists in current research ethics approaches, which is that it is often not clear for developers how to implement ethical guidelines and recommendations, which are often of a quite general and abstract nature. Ethics by design methodologies identify how at different stages in the development process, ethical considerations can be included in development, by finding ways to translate and operationalize ethical guidelines into concrete design practices. Ethics by design approaches have been in existence

[13] Tambornino, Lisa, Dirk Lanzareth, Rowena Rodrigues, David Wright, SIENNA D4.3 Survey of REC approaches and codes for Artificial Intelligence & Robotics, August 2019: https://zenodo.org/record/4067990#.YDUBni1Q2u5

[14] https://satoriproject.eu

[15] CEN, *Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework.* CEN workshop agreement, CWA 17145-2, 2017.

in computer science and engineering since the early 1990s, initially under the name Value-sensitive design (VSD)[16] and later also under the label of Design for Values.[17] Over 2020, the term "ethics by design" has come into vogue. An extensive ethics by design approach for AI was published as part of the EU Horizon 2020-funded project SHERPA.[18] As far as we can see, as of writing, no other full-blown ethics by design approaches have yet been published for AI & robotics, although the IEEE is working on one. The SIENNA project builds on the SHERPA report to present an extended approach for ethics by design that has wider applicability than the one proposed in that report. Annex 2 of the present report presents the ethics by design guidelines developed in SIENNA.

4. Codes of professional ethics for researchers and developers of AI & robotics technologies

Codes of professional ethics, also called codes of conduct, are codified personal and corporate standards of behaviour that are expected in a certain profession or field. These codes are often set by professional organisations. To our knowledge, no internationally accepted codes of ethics for either artificial intelligence specialists or robotics engineers are currently in existence, and few if any national codes for these professions exist either. Wider codes of ethics, for computer scientists and electrical engineers, are in existence and cover the AI and robotics professions as well. However, these broader codes do not address the specific challenges and responsibilities of AI and robotics specialists. In this report, we do not attempt to propose codes of professional ethics for these professions.

In the AI HLEG and IEEE reports, various other methods for incorporating ethics into R&D are mentioned. Some of these can however, in our opinion, be subsumed under ethics by design approaches. These include, amongst others, the development and use of explanation methods for transparency, extensive testing and validation, the definition of quality of service indicators, and better technical documentation. Others will be discussed under the heading of "corporate social responsibility cultures" below. One method deserves special attention, however: interdisciplinary research, which is proposed in the IEEE report. **Interdisciplinary research is, in our view, an important component of ethical AI & Robotics**, if it involves collaborations that bring engineers and scientists into contact with social science and humanity scholars, including ethicists. Such research activities allow for a better incorporation of social and ethical concerns into engineering practice, and are therefore highly advisable, at different stages of the R&D continuum.

3.2. Methods to incorporate ethics into the deployment and use of AI & robotics

After the development of AI & robotics systems, services and solutions, they are deployed and used by organisations or individuals.[19] The deployment and use of these technologies often require their own ethical guidelines and solutions, that are to some extent different from those that apply to their development. Ethical questions that are typically asked in relation to deployment and use include questions like: Is it ethical to deploy a system that is intended to do X/is capable of doing X/ can be used to do X? How can unethical uses of the system be monitored and prevented? What is the

[16] Friedman, B., Kahn, P. and Borning, A., "Value Sensitive Design and Information Systems," in *Human-Computer Interaction in Management Information Systems: Foundations* (eds. P. Zhang and D. Galletta). Armonk, NY: M.E. Sharpe, 2006.

[17] Hoven, Jeroen van den, Pieter E. Vermaas, and Ibo van de Poel, eds., *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, Springer Netherlands, 2015.

[18] Brey, P., Lundgren, B., Macnish, K. and Ryan, M., *Guidelines for the development and use of SIS,* 2019. Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833.

[19] Of course, deployment and use cycles are often followed by repeated redevelopment of systems.

responsibility of different stakeholders in preventing or mitigating unethical use? What policies to prevent unethical use should be put in place and how can they be implemented effectively?
Deployment and use scenarios come in various forms, but the following are the most typical:

(1) Deploying AI or robotics technology to enhance organisational processes. An organisation acquires AI or robotics technology and uses it to improve its organisational processes, such as manufacturing, logistics, and marketing. End-users are IT specialists or other employees.

(2) Embedding AI and robotics technology in products and services. An organisation acquires AI or robotics technology and incorporates it into products or services that it offers to customers. This is a different application of AI and robotics than its application in the development, manufacturing and marketing of products and services. For example, AI can be used to better design, manufacture or market automobiles that themselves do not contain AI technology. AI and robotics technologies can be embedded in products and services for different purposes:
   a. To enhance the value of a product or service for customers by offering enhanced functionality or usability. E.g., by powering an online dating service with AI algorithms, or by enhancing an automobile with a self-drive mode.
   b. To enhance the value of a product or service through intelligent monitoring, self-repair, communications with customer service, or data collection for future upgrades.
   c. To further the interests of the organisation or of third parties, for example, by collecting data for marketing purposes or allowing for targeted messaging.

It is not always clear who is the end-user of the AI and robotics technology in these three scenarios, since the end-user of AI or robotics technology embedded in a product or service may be different from the end-user of that product or service, and there may also be multiple end-users (e.g., Uber drivers and customers using the same AI algorithms).

Taking these scenarios into consideration, the following four methods can contribute to ethical deployment and use of AI & robotics technologies:

(1) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies for the enhancement of organisational processes

(2) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies in products and services

(3) Codes of professional ethics for IT managers, technical support specialists and other management, IT and engineering staff responsible for the deployment and use of the AI & robotics technologies in an organisation or its embedding in products and services

(4) End-user guidelines for ethical usage of (products and services that include) AI and robotics technologies

In Brey, Lundgren, Macnish and Ryan, the previously mentioned SHERPA report, proposals were made for the first and, to some extent, the second of these methods.[20] Building on two widely used models for the management and governance of information technology in organisations, ITIL (Information Technology Infrastructure Library) and COBIT (Control Objectives for Information and Related Technologies) , as well as on the ethics requirement of the AI HLEG, this report proposed operational guidelines for the deployment and use of AI systems (including AI-powered robotic systems) in organisations.

[20] Brey, P., et al., op. cit., 2019.

3.3. Corporate responsibility policies and cultures

Ethical guidelines and professional ethical codes, even when fully operationalized for particular practices, will have little impact if they are not supported by organisational structures, policies and cultures of responsibility. In Brey, Lundgren, Macnish and Ryan, specifically the division of the report with guidelines for the ethical deployment and use of AI, an attempt was made to include these wider considerations of responsibility in organisations in the guidelines that were proposed.[21] For instance, requirement 1 in that report, which targets the board of directors of companies, reads as follows:

> "Requirement 1. The board of directors should direct in its IT governance framework that IT management adopts and implements relevant ethical guidelines for the IT field and should monitor conformity with this directive. There should be an appointed representative at each level of the organisation, including the board of directors, who are 'ethics leaders' or 'ethics champions', and who should meet regularly to discuss ethical issues and best practice within the organisation. The ethics leader from the board of directors should be responsible for the ethical practice of the whole organisation."[22]

Requirements 2, 3 and 4, which targets IT management, are as follows:

> "Requirement 2. The IT management strategy should include the adoption and communication to relevant audiences of ethics guidelines for AI and big data systems, define corresponding ethics requirements within role and responsibility descriptions of relevant staff, and include policies for the implementation of the ethics guidelines and monitoring activities for compliance and performance."[23]
>
> "Requirement 3: The IT management strategy should include the design and implementation of training programs for ethical awareness, ethical conduct, and competent execution of ethical policies and procedures, and these programs should cover the ethical deployment and use of the system. More generally, IT management should encourage a common culture of responsibility, integrating both bottom-up and top-down approaches to ethical adherence."[24]
>
> "Requirement 4: Consider how the implementation of the AI and big data systems ethics guidelines, and other IT-related ethics guidelines, affects the various dimensions of IT management strategy, including overall objectives, quality management, portfolio management, risk management, data management, enterprise architecture management, stakeholder relationship management. Ensure proper adjustment of these processes. There will be different levels of risk involved, depending upon the application, so the levels of risk need to be clearly articulated to allow different responses from the organisation's ethical protocols."[25]

These guidelines, and several others that are proposed, serve as meta-guidelines for the proper implementation of ethics guidelines for AI & robotics in organisations. They point out that proper implementation of ethics considerations in organisations involves much more than the development and distribution of operationalized ethics guidelines, but also requires leadership from the top, adjustment of existing management strategy, definitions of roles and responsibilities, training of staff, monitoring and assurance activities, and encouragement of a common culture of responsibility. While these guidelines were developed for organisations that deploy and use AI & robotics technologies, they are also applicable to organisations that engage in AI & Robotics R&D.

[21] Ibid., p. 53-87.
[22] Ibid., p. 61.
[23] Ibid., p. 64.
[24] Ibid., p. 64-65.
[25] Ibid., p. 65.

3.4. National and international guidelines, standards and certification

In this report, we distinguish between *operational ethics guidelines*, which are detailed, practical guidelines developed for specific practices by specific stakeholders, and *general ethics guidelines*, which are statements of ethical principles and general guidelines that apply to a broad range of stakeholders and practices. While it is possible to develop operational guidelines without general guidelines, it is often beneficial to have shared general guidelines on the basis of which operational guidelines are developed. These guidelines can be supported by national governments and intergovernmental organisations. Currently the two most important sets of international guidelines for AI & robotics technologies are the Recommendation of the Council on Artificial Intelligence of the OECD (2019) and the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence of the European Commission (AI HLEG, 2019). These two documents currently serve as the two most important international guidance documents for ethical issues in AI & robotics.

Next to such general guidelines, directed at all stakeholders, there are also ethical guidelines that are general rather than operational, but that are focused on specific stakeholders or practices. The guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019) are a case in point. These specifically apply to design practices and are of greatest relevance to technology developers.

Standards, developed by recognised national and international standards organisations or by particular (associations of) companies or organisations, are different from ethics guidelines in two ways. First, they apply to specific products, services, processes or methods, while ethics guidelines apply to any action, thing or event that has ethical implications. Second, they define specific norms or requirements to which the phenomenon to which the standard applies must adhere. Standards are intended to leave limited room for subjectivity and interpretation and are intended to define intersubjective requirements that different stakeholders can apply, identify or assess. Standards sometimes aim to codify ethical requirements, procedures or methods.[26] Examples are ISO 26000[27], which is an international standard for corporate social responsibility, CEN CWA 17145-1[28], which is a standard for ethics assessment by research ethics committees, and CEN CWA 17145-2[29], which is a standard for the method of ethical impact assessment for R&I. Standards can also include ethical requirements, procedures or methods, while not themselves having ethics as a focus. For example, ethics is discussed in the context of the ISO 9000 and 9001 standards[30] for quality management.

For AI & robotics, a remarkable number of ethical standards are currently being developed by IEEE as part of its Ethically Aligned Design programme.[31] A total of 13 standards are in development, including standards for ethics by design, transparency of AI systems, algorithmic bias, data privacy, ethically driven robotics and automation systems, and automated facial analysis technology. The ISO also has several standards in development that focus in part or in whole on ethical issues, including standards for identifying ethical and societal concerns in AI systems, bias in AI systems, trustworthiness of AI systems, quality assurance in AI and risk assessment in AI.

[26] ISO is the International Organisation for Standardisation which develops and publishes international standards: https://www.iso.org/home.html. CEN is the European Committee for Standardization. It brings together the National Standardization Bodies of 34 European countries to develop and define standards at the European level.
[27] https://www.iso.org/iso-26000-social-responsibility.html
[28] https://satoriproject.eu/media/CWA_part_1.pdf
[29] https://satoriproject.eu/media/CWA17145-23d2017.pdf
[30] https://www.iso.org/iso-9001-quality-management.html
[31] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE), *op. cit.,* 2019.

Certification is the process by which an external third party (typically a certifying body) verifies that an object, person or organisation is in possession of certain characteristics or qualities. Amongst others, certification can be applied to persons, in professional certification, to products, to determine if it meets minimum standards, and to organizations or organizational processes, through external audits, to verify that they meet certain standards. Certification can be a means to verify and validate the quality of ethics processes and procedures in organisations. In relation to standards, in particular, certification can be a means of ensuring conformity to the requirements of the standard. IEEE is currently developing its own certification programme to certify adherence to the ethics standards it is developing. ISO does not carry out certification itself, but third-party certification organisations could in the future, assess compliance to ISO ethics-related standards for AI. The present report (D5.4) develops a certification scheme in Annex 3.

3.5. Education, training and awareness raising

Education is a powerful method for stimulating ethical behaviour in relation to AI & robotics. In professional and academic education, specifically, education that concerns ethical and social issues in AI & robotics would benefit future professionals, especially those in the AI & robotics field, those in other fields who may deploy and use these technologies in the future, and more generally, any individuals to make informed decisions about AI and robotics. Given the seriousness of ethical issues in the AI & robotics fields, a required ethics course for AI and robotics students seems advisable. Such a course could cover key ethical issues in AI & robotics, ethical guidelines and their application, responsibilities of AI and robotics professionals, and relevant standards, laws, policies, and approaches for ethical AI & robotics. Methodologies for ethics by design could be part of such a course, but for these to be used by future professionals in actual design practice, it might be better if these were to be incorporated in the standard design methodologies used in these fields.

Most professionals who develop and use AI & robotics did not have ethics education in these areas in their professional education. For them, continuing education programmes that include ethics of AI and/or robotics would be valuable. Such training programmes could even be accompanied by professional certification, for example, certification in ethics by design methodology, algorithmic bias avoidance, preparing for ethics review, or all-round ethical practice in AI or robotics. Next to external organisations setting up such training and education programmes, organisations could of course also organise their own in-house training in ethics for AI & robotics.

Turning now from educational institutions to the media, we should acknowledge that media organisations and journalists (including independent ones) have a large role in generating public awareness and understanding of AI & robotics, including the ethical issues raised by them. These are complicated technologies that are difficult to understand for the general public. Since they are expected to have major impacts on people's lives, a proper understanding of them and the ethical issues they raise is important. A certain degree of awareness of the technologies and their social and ethical impacts is also essential to ensure proper public oversight over them. Media companies and journalists are an important type of organisation that can provide such an understanding to the general public. Therefore, relevant media stories on AI & robotics and its social and ethical dimensions, whether in print, podcast, television or other formats, are important.[32] While media organisations and

[32] See in particular the media analysis conducted as part of SIENNA Deliverable D4.4 and the public perception studies conducted in SIENNA in D4.5 and D4.6: Jansen, P., et al, op. cit., 2020. Hamlyn, R., op. cit., 2020. Kantar (Public Division), op. cit., 2019.

journalists have a major responsibility here, AI & robotics developers also have a responsibility to be transparent and communicate with the public about these issues, and governments in ensuring that sufficient information is provided.

3.6. Policy, regulation and governance

While policy can be made by any kind of organisation, our concern is with public policy, as made by governments, as well as the laws and regulations created by them. The key question here is: what policies, laws and regulations should governments develop, if any, to stimulate the ethical development, deployment and use of AI & robotics? Policies, laws and regulations can relate to ethical criteria in three ways: they can explicitly institute, promote or require ethics guidelines, procedures, or bodies; they can have a focus on upholding certain moral values or principles without explicitly identifying them as ethical (e.g., well-being, privacy, fairness, sustainability, civil rights); and they either explicitly or implicitly take on board ethical considerations in broader social and economic policies.

Governments are currently at a decision point for AI & robotics policy. What should they do, and how can they avoid regulating too little as well as regulating too much? Decisions that relate to ethics include the following:

- Whether or not to issue, or support the issuing of, ethical guidelines for AI & robotics
- Whether or not to put any ethical guidelines for AI & robotics into law
- Whether or not to revise existing institutional structures to better account for ethical issues or to create new governmental bodies or unites for ethical and social issues in AI & robotics
- Whether or not to mandate ethics standards, certification, education, training, ethical impact assessments or ethics by design methods in relation to ethics of AI & robotics
- Whether and how to introduce new legislation and regulations for morally controversial AI & robotics technologies, such as automated tracking, profiling and identification technologies, behaviour and affect recognition technologies, and automated lethal weapons
- How to include ethical considerations concerning AI & robotics in policies, laws and regulations, both ones that pertain to AI & robotics specifically and more general ones that need to be updated to account for AI & robotics, such as in the areas of consumer protection, data protection, criminal law, non-discrimination provisions, civil liability and accountability
- What financial support and funding to provide, if any, for ethics research, ethics education, ethics dialogue, ethics awareness raising and other ethics initiatives in relation to AI & robotics
- How to regulate the government's own use of AI & robotics so as to ensure ethical conduct

See also SIENNA report D5.6, *Recommendations for the enhancement of the existing EU and international legal framework*, which contains our proposals to support ethical AI & robotics.[33]

3.7. Studies on the ethical and social impacts of AI and robotics

The last method that we wish to highlight to ensure ethical considerations are taken into account in the development and use of AI and robotics concerns the **need to ensure ongoing studies on the ethical and social impacts of these technologies** on the society and individuals. There are still many consequences of the technology that we do not fully comprehend nor are able to mitigate properly.

[33] Konrad Siemaszko, Rowena Rodrigues, & Santa Slokenberga, "SIENNA D5.6: Recommendations for the enhancement of the existing legal frameworks for genomics, human enhancement, and AI and robotics" , 2020. https://doi.org/10.5281/zenodo.4121082

These include aspects related to bias and discrimination, the impact of the rising level of surveillance, or questions related to human agency and autonomy. We can only get to a fuller understanding of these issues through these ethical and social studies on the short, medium and longer term.

Finally, a general remark regarding these methods: it remains to be seen whether ethical AI & robotics are best served by specific ethics standards, certification, design methodologies, audits, policies and other methods, or whether it is better to integrate ethics concerns into broader standards, policies, audits, etc. This probably varies from situation to situation but should receive proper attention as an issue to account for.

4. Making methods available and motivating stakeholders

In the preceding discussion of methods, we made a number of suggestions regarding the responsibility of different stakeholders for developing and making available different types of methods. Obviously, governments responsible for developing government policies, laws and regulations, and universities would lead the development of ethics courses in degree programmes in AI and robotics. In other cases, it may not be immediately obvious which stakeholder would be responsible for developing and advocating a particular method. Which stakeholder would be responsible for developing methods of ethical impact assessment, for example, or for developing operational ethics guidelines for the deployment and use of AI in organisations? Often, this is a matter of particular stakeholders stepping up and taking on such responsibilities. For instance, the IEEE embarked on an extensive programme to develop ethical guidelines, methods, standards and certification for the design and deployment of AI and robotics systems, but it nevertheless chose to do so.
A recent study has shown that developers themselves do often care about the ethical implications of what they develop.[34] As the study shows, developers generally rely on their own ethical compass to guide them in their work and to ensure their outputs do not lead to ethical issues or negative social impacts and actually brings beneficial outcomes. It can be expected that what this study has shown about developers can be extended to most people, and therefore to the various categories of stakeholders listed in Section 2. In that sense, any formalised attempts at ensuring ethical aspects are taken into account in the development and use of AI and robotics products aim at nourishing and supporting the existing ethical sense, rather than at imposing guidance from outside.

However, relevant stakeholders may fail to step up, leaving a responsibility vacuum in society due to which important methods for ethical AI & robotics are not being developed and implemented. If this is to occur, then governments are often seen as the responsible stakeholder to step in and enact policies, laws and regulations that help fill this vacuum. While there are some limitations, governments, after all, have a particular responsibility to promote the public good, protect human rights, and support fair socioeconomic conditions, and have the powers to stimulate and compel other stakeholders to act responsibly and in the public interest.

[34] Miller, Catherine, and Rachel Coldicutt, "People, Power and Technology: The Tech Workers' View," London, Doteveryone, 2019, p. 16. https://doteveryone.org.uk/report/workersview.

5. Conclusion

This report proposed a Multi-stakeholder strategy for ethical AI and robotics. It showed that a strategy for ethical AI and robotics should contain three components: (1) an identification of relevant stakeholders; (2) an identification of methods that these stakeholders can use to contribute to ethical AI & robotics, and (3) proposals of ways in which these methods can be made available to these stakeholders, and ways to motivate them to use them. Subsequently, these three components were elaborated in the report. Seven main classes of relevant stakeholders were defined, including AI & robotics developers; AI & robotics development support organizations; organisations that deploy and use AI & robotics technology; governance and standards organizations; educational and media organizations; civil society organizations and the general public; and organisations and units working on ethics and social impacts.

Seven types of methods for ethical AI & robotics were discussed and related to these classes of stakeholders: methods for ethical development and design, methods for ethical deployment and use, corporate responsibility policies and cultures, national and international guidelines, standards and certification, policy and regulation actions (by governments), and education, training and awareness raising; studies on the ethical and social impacts of AI and robotics. Finally, it briefly discussed how these methods can be made available to stakeholders.

This strategy provided an overview of the various actors and methods required to promote ethical AI and robotics. It can be used to ensure all necessary actors are engaged toward this effort to ensure AI and robotics are developed, deployed and used in a way that respects ethical values and principles and avoids harmful impacts on the society and individuals. It is also our hope that it will motivate more stakeholders to take a role toward ethical AI and robotics.

6. References

13th Annual State of Agile Survey, 7 May 2019. https://www.stateofagile.com/

Alix, *Working Ethically At Speed*, 7 May 2018. https://medium.com/@alixtrot/working-ethically-at-speed-4534358e7eed

Beck, Kent, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, … David Thomas, *Manifesto for Agile Software Development,* 2001. http://agilemanifesto.org/

Brey, Philip, Björn Lundgren, Kevin Macnish, and Mark Ryan, *Guidelines for the development and use of SIS,* 2019. Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833.

Brey, Philip, Philip Jansen, Jonne Maas, Björn Lundgren, and Anais Resseguier, SIENNA D4.7 An Ethical framework for the development and use of AI and robotics technologies, 2020, submitted public deliverable from the SIENNA project, awaiting approval from the European Commission, 2020. https://www.sienna-project.eu/digitalAssets/801/c_801912-l_1-k_d4.7_ethical-framework-for-ai--robotics_v2.pdf .

CEN, *Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework.* CEN workshop agreement, CWA 17145-2, 2017.

Fitzgerald, Brian, Gerard Hartnett, and Kieran Conboy, Customising agile methods to software practices at Intel Shannon. *European Journal of Information Systems*, 15(2), 200-213, 2006.

Friedman, Batya, Peter Kahn and Alan Borning, "Value Sensitive Design and Information Systems," in P. Zhang and D. Galletta (eds.), *Human-Computer Interaction in Management Information Systems: Foundations*, Armonk, NY: M.E. Sharpe, 2006.

Gonçalves, Luis, *What Is Agile Methodology*, 1 May 2019. https://luis-goncalves.com/what-is-agile-methodology/

Hamlyn, Rebecca, SIENNA D4.5 Public views on artificial intelligence and robots across 11 EU and non-EU countries, 2020, https://doi.org/10.5281/zenodo.4068220

AI HLEG (High-Level Expert Group on Artificial Intelligence), *Ethics Guidelines for Trustworthy AI*, 2019. Downloaded on 8-3-2020 at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Hoven, Jeroen van den, Pieter E. Vermaas, and Ibo van de Poel, eds., *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, Springer Netherlands, 2015.

Huldtgren, Alina, "Design for Values in ICT", in Jeroen van den Hoven, Pieter Vermaas, and Ibo van de Poel (eds.), *Handbook of ethics and values in technological design. Sources, Theory, Values and Application Domains,* Springer, 2015.

IEEE (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) (). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/

Jansen, Philip, Philp Brey, Alice Fox, Jonne Maas, Bradley Hillas, Nils Wagner, Patrick Smith, Isaac Ouoch, Laura Lamers, Hero van Gein, Anais Resseguier, Rowena Rodrigues, David Wright, and David Douglas, SIENNA D4.4: Ethical Analysis of AI and Robotics Technologies, 2019, https://doi.org/10.5281/zenodo.4068082.

Jansen, Philip, Stearns Broadhead, Rowena Rodrigues, David Wright, Philip Brey, Alice Fox, Ning Wang, SIENNA D4.1: State-of-the-art Review: Artificial Intelligence and robotics, 2019, https://doi.org/10.5281/zenodo.4066571.

Kantar (Public Division), SIENNA D4.6: Qualitative research exploring public attitudes to AI and robotics, 2019, https://doi.org/10.5281/zenodo.4081246.

Kneuper, Ralf, *Software Processes and Life Cycle Models: An Introduction to Modelling.* Cham, Switzerland: Springer Nature Switzerland, 2018.

Liversidge, Ed, The Death Of The V-Model, *Harmonic Software Systems*, June 25, 2015. http://harmonicss.co.uk/project/the-death-of-the-v-model/.

Miller, Catherine, and Rachel Coldicutt, "People, Power and Technology: The Tech Workers' View," London, Doteveryone, 2019, p. 16. https://doteveryone.org.uk/report/workersview.

Ministry of Science and Technology of China (2019). *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence*. A translation can be found at: https://perma.cc/V9FL-H6J7.

OECD (2019). *Recommendation of the Council on Artificial Intelligence.* Retrieved on 8-3-2020 at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Rodrigues, Rowena, Konrad Siemaszko, and Zuzanna Warso, SIENNA D4.2: Analysis of the legal and human rights requirements for AI and robotics in and outside the EU, 2019, https://doi.org/10.5281/zenodo.4066811.

Ryan, Mark, Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van Der Puil (2019). *Ethical Tensions and Social Impacts.* Deliverable D 1.4 of the SHERPA project. https://doi.org/10.21253/DMU.8397134

Tambornino, Lisa, Dirk Lanzerath, Rowena Rodrigues, and David Wright, SIENNA D4.3: Survey of REC approaches and codes for Artificial Intelligence & Robotics 2019, https://doi.org/10.5281/zenodo.4067989.

Wellens, Koen, The Seductive and Dangerous V-Model, *Testing Experience* magazine, December 2008. Expanded version of article available at http://www.clarotesting.com/page11.htm.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford, "Discriminating Systems. Gender, Race, and Power in AI", AI Now Institute, 2019. https://ainowinstitute.org/discriminatingsystems.html

sienna.

Annex 2 - Ethics By Design and Ethics of Use in AI and Robotics

Ethics by Design and Ethics of Use in AI and Robotics

Annex 2 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Philip Brey, *University of Twente* (p.a.e.brey@utwente.nl)
Other contributors	Brandt Dainow, *University of Twente*
Reviewers	Ansgar Koene, University of Nottingham

Date	February 2021
Type	Report (Annex 2 to D5.4 deliverable)
Dissemination level	PU = Public

Keywords	Ethical issues; artificial intelligence; AI; robotics; robots; ethics; software design;

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This document offers guidance for an *Ethics by Design* approach when developing AI-driven systems, including robots. The Ethics by Design approach offers a way by which to include ethical principles and procedures into the design and development processes. Historically, ethical problems in AI have only been detected after the system has been deployed. Essentially, Ethics by Design seeks to make the ethical aspects of AI and robotics systems integral requirements of the system on the same level as reliability or security. The aim of Ethics by Design is to ensure ethical problems are not generated in the first place by using ethically-focused activities throughout the design, development and deployment phases of a project. We first detail foundational ethical values which all AI-driven systems should comply with. We then extrapolate these into specific features an AI system should possess. In order to make this document as useful as possible, these features are, as much as possible, presented as tasks to be performed. Finally, we show how to apply these concerns within each stage of the design, development and deployment stage.

Table of Contents

Executive summary

The Ethics by Design approach offers a way by which to include ethical principles into the design and development processes of AI-driven systems. Historically, ethical problems in AI have too often been detected after the system has been deployed. As part of the "By Design" movement, Ethics by Design is intended to prevent systems being created with ethical issues at all, just as Privacy by Design seeks to prevent systems being created which have privacy issues. Ethics by Design makes ethical aspects of the system integral requirements on the same level as reliability or security. In addition to formal ethical assessment before they are built, this requires changes in the way systems are developed by using ethically-focused activities and tools throughout of the design, development and deployment phases of a project. These activities are detailed in this document, along with the ethical values these activities uphold.

The ethical values upon which Ethics by Design is based, are drawn from previous research into responsible innovation within the EU and from international standards such as the Universal Declaration of Human Rights. These values are grouped into six categories, such as fairness, accountability and transparency. Applying these to AI and robotics, we then develop "ethical requisites," which are the conditions that a system must meet in order to achieve its goals ethically. Ethical requisites are instantiations of values within AI and robotics systems and development cycles. Asimov's Three Laws of Robotics are an example of ethical requisites. Ethical requisites may be met in many ways; through functionality, in data structures, in the process by which the system is constructed, and so forth. For example, one way the value of fairness can be met as an ethical requisite is to require that a system does not exhibit racial bias. While many ethical requisites are aspects of the system itself, some are concerned with the way in which the system is developed. For example, the value of transparency requires that developers can explain how they tested for and removed bias from a dataset. It is not sufficient for developers to be satisfied there is no bias. If others suggest that the system is biased, developers must be able to show what processes they used to remove bias and the analysis they undertook to determine why those processes, and not others, were used.

We then derive from these ethical requisites sets of ethical guidelines to be followed at different stages of the design, development and deployment of the system. These guidelines are concrete tasks which must be performed in order to achieve the ethical requisites.

The main ethical requisites for AI and robotics systems can be summarised as follows:

- Because each individual has an inherent worth, AI systems should not negatively affect human autonomy, freedom or dignity, nor limit participation in democratic processes.
- Because AI systems rely on data, it is important they do not violate the right to privacy and that the data used is representative and accurate.
- Systems should be developed with an inclusionary, fair, and non-discriminatory agenda.
- Because AI and robotics systems can have significant effects on individuals, society, and the environment, steps need to be taken to ensure they do not directly cause harm, rely on

harmful technologies or processes, or influence others to act in ways which cause harm to individual, societal or environmental well-being.
- Human oversight and accountability are required to ensure conformance to these principles and address non-compliance.
- Systems should be as transparent as possible because only then are accountability and human oversight possible.

List of figures

List of tables

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
GDPR	General Data Protection Regulation
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
SHERPA	Shaping the ethical dimensions of smart information systems– a European perspective. An EU-funded Horizon 2020 project.
XAI	Explainable Artificial Intelligence

Table 1: List of acronyms/abbreviations

Glossary of terms

Term	Explanation
Accountability	Accountability applies to both individuals and institutions. It means being able to explain the reasons behind your actions and a willingness be held responsible for them.
AI	Artificial Intelligence
Auditability	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes.
Autonomy	Ethical AI is concerned with human autonomy, of which there are three types. Moral autonomy is determining what is morally good and bad. Political autonomy refers forming one's own political opinions. Personal autonomy refers to deciding how one should live, especially by what values one should make decisions.
Bias	Bias is an unfair or unjustified prejudice towards or against a person, group of people, object, or position. Bias is a danger because it causes unfair outcomes in AI systems
Discrimination	The act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong, especially gender, race, age, sexual orientation, national origin, religion, income, property, health, or disability.
Diversity	Diversity is the organisation of people based on identity markers like gender, race, age, cultural heritage, ability, and education.
Ethics	Ethics are moral principles that govern a person's behaviour. It is also a branch of philosophy dealing with these principles. Applied ethics deals with the use of moral principles in real-life situations. AI Ethics is an example of applied ethics focused on the issues raised by AI.
Ethics assessment	The assessment, evaluation, review, appraisal or valuation of plans, practices, products and uses of research and innovation that makes use of ethical principles or criteria.
Ethical AI	Ethical AI refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental human rights, ethical principles and related core values.
Ethical impact assessment	An approach for judging the ethical impacts of research and innovation activities, outcomes and technologies that incorporates both the means for a contextual identification and evaluation of these ethical impacts and the development of a set of guidelines or recommendations for remedial actions aimed at mitigating ethical risks and enhancing ethical benefits, typically in consultation with stakeholders.
Ethical requisite	A key term in this document. An ethical requisite is a requirement relating to ethical aspects of the system and the development thereof. Ethical requisites must be met in order to be compliant with the demands for responsible, trustworthy, ethical AI.
Explainability	Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
Informed consent	Permission freely given and granted in full knowledge of the possible consequences.

Term	Explanation
Oversight	The ability to oversee, supervise, and watch carefully over something – in this context, to oversee the functionality and output of AI systems.
Personal data	Information relating to an identified or identifiable natural person, directly or indirectly, by reference to one or more elements specific to that person. The GDPR specifically mentions racial or ethnic origin, political opinions, religious beliefs, trade union membership, genetic data, biometric data, health, and sexual orientation.
Personal data processing	Any operation or set of programmatic operations to personal data.
Privacy by design	Privacy by Design is an approach taken when creating new technologies and systems. Privacy by Design encompasses IT systems, business practices and physical design. The approach is characterized by proactive anticipation of privacy invasive events so as to prevent them from occurring, rather than fixing them afterwards.
Profiling	According to Article 4(4) of the GDPR, 'profiling' means automated processing of personal data to evaluate personal aspects relating to a person, such as personal preferences, interests, or movements.
Pseudonymisation	According to Article 4 of GDPR, 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person
Reproducibility	Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions.
Stakeholders	All those that research develop, design, deploy or use AI, as well as those (directly or indirectly) affected by AI – including but not limited to companies, organisations, researchers, public services, institutions, civil society organisations, governments, regulators, social partners, individuals, citizens, workers and consumers.
Traceability	Traceability of an AI system refers to the capability to keep track of the system's data, development and deployment processes, typically by means of documented recorded identification.
XAI	Explainable AI. XAI refers to initiatives, including procedures and coding tools, in response to AI transparency and trust concerns. XAI aims to produce explainable models while also maintaining a high level of learning performance; and enable human users to understand, trust and manage AI systems

Table 2: Glossary of terms

1. Introduction

This guidance document offers guidance for people who wish to develop AI-based systems (including robotics), and a set of potential assessment criteria and values for those who have a concern for the ethical status of AI systems.[1] This document outlines an "Ethics by Design" approach, which aims for the systematic inclusion of ethical values, principles, requirements and procedures into design and development processes. Traditionally, ethical issues in AI systems have been discovered after the systems have been completed, usually only once they start to cause harm. The Ethics by Design approach is intended to ensure ethical problems are not generated in the first place. This requires specific ethically-focused activities at each stage of the design, development and deployment phases of a project. These activities are detailed in this document, as are the ethical values these activities uphold, and to which all AI and robotics projects should comply. However, it is important to bear in mind the importance of any particular ethical value will depend on the type of application and the relevance of that value to it. Not all values are equally important, so judgement must be exercised when considering them.

Using a generic model of the design process, we then offer detailed explanations of the factors which will ensure the system is ethical at each stage of the design process. We start with six ethical values. We then explain how each value can be embodied in AI and robotics systems as an "ethical requisite," (a project requirement). For example, in order to be fair, it is an ethical requisite that the system does not discriminate against particular racial groups. This document then explains how system developers can meet these ethical requisites at each stage of the design and development process.

Finally, we present an *Ethics of Deployment and Use* approach for proper inclusion of such guidelines at different stages in the deployment and use of these systems.

2. Ethics by Design Principles

This section explains the principles by which ethical concerns can be factored into the design process. Chapter 4: Ethical Deployment and Use is intended for those deploying AI or robotics systems.

2.1 The Ethics by Design approach

Ethics by Design is an example of the Value Sensitive Design approach[2]. However, as of writing, no detailed proposals for Ethics by Design approaches have been published. Our Ethics by Design approach is based on the findings of the EU-funded SHERPA project, which also takes an Ethics by Design approach. We moreover build on the ethics principles proposed in the High-Level Expert Groups on AI's *Ethics Guidelines For Trustworthy AI* [3], as well as from the SHERPA and SIENNA reports.

Many AI projects experience ethical issues only after they are deployed and start causing harm. Ethics by Design is intended to prevent ethical issues from arising in the first place, rather than trying to fix

[1] The authors of this report acknowledge the input of various experts and stakeholders to this text. Please see the Acknowledgement section of SIENNA D5.4 (Feb 2021) for a list of these people.
[2] Friedman, Kahn, and Borning, "Value Sensitive Design: Theory and Methods".
[3] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines For Trustworthy AI*.

them after the damage has been done. Ethics by Design is intended to prevent ethical issues from occurring by proactively using moral principles as requirements of the system, termed "ethical requisites". Since many cannot be achieved unless the system is constructed in particular ways, ethical requisites sometimes apply to development processes and tools rather than the system being produced. Ethics by Design is therefore something which affects the planning and creation processes by which to build AI-driven systems.

2.2 5-Layer Model of Ethics by Design

Ethics by Design can be described in a five-layer model. This model is similar to many others in Computer Science in that higher levels are more abstract, with increasing levels of specificity going down the levels.

1. **Ethics by Design Values** – These are the primary ethical values by which we want to guide the ethical status of an AI or robotics system. Where a system violates these values, it may be considered unethical. Values are to be upheld and enhanced. Privacy and fairness are examples of such values.
2. **Ethical Requisites** – Ethical requisites are the conditions that a solution or application must meet in order to achieve its goals ethically. In Ethics by Design, ethical requisites are instantiations of values within AI and robotics systems. Values may be instantiated in many ways; through functionality, in data structures, in the process by which the system is constructed, and so forth. For example, one way the value of fairness can be instantiated as an ethical requisite is to require that a system does not exhibit racial bias. Asimov's Three Laws of Robotics are an example of ethical requisites.
3. **Ethics by Design Guidelines** Whereas ethical requisites are concerned with the system, ethical guidelines are concerned with the steps by which it is created. Ethics by Design works on the basis that there are steps in the development process which are common to all design methodologies. The Ethics by Design approach offers a generic description of these phases in the development process and maps the ethical requisites onto these phases. This yields specific guidelines (usually formulated as tasks) at each phase which ensure that the final system instantiates the ethical requisites and therefore does not violate any ethical values. For example, the guidelines state that during the data gathering stage, data should be screened for fairness and any discriminatory biases found corrected.
4. **AI Methodologies** – There are a variety of methodologies used in AI and robotics projects. They are, at least partially, distinguished by the manner in which they organise the development process. Each methodology offers its own steps and sequence. Here Ethics by Design maps its principles onto the components of each individual methodology. If a project uses a different methodology, a developer can refer to the generic model. By mapping the steps in the generic development process to their own methodology, they can then allocate each guideline to the appropriate steps in their methodology.
5. **Tools & Methods** – The Tools and Methods layer accommodates specific programmatic artefacts and processes deployed within the development process to undertake Ethics by Design. It is possible some could be specific to a particular methodology and inapplicable to others, but at this stage, those which have emerged in the development community are tuned to ethical requisites and useable under any methodology. For example, Datasheets for

Datasets [4] are employed to interrogate the ethical characteristics of data, and so can be used at any stage which works with that data and for any norm relating to data. They can thus be deployed at multiple stages of the development process and are methodology-neutral.

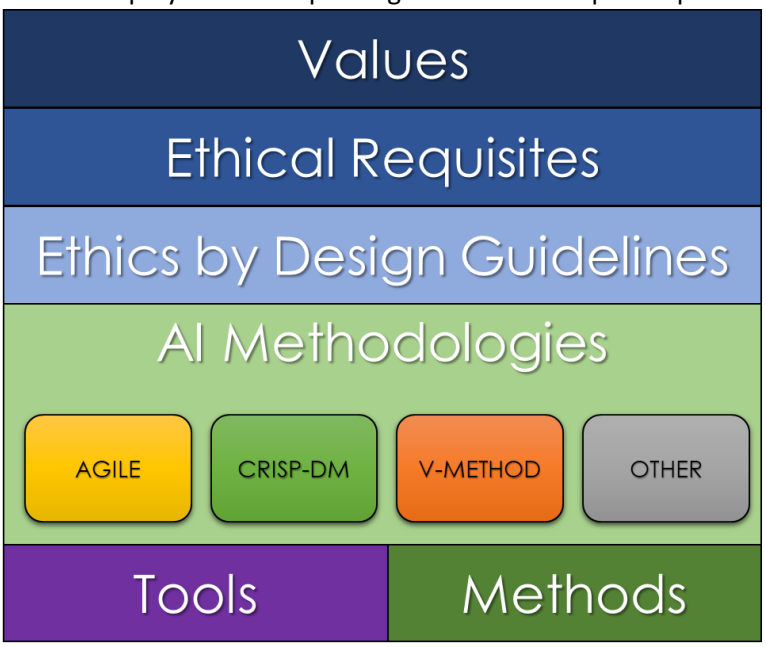

Figure 1:The 5-layer Model of Ethics by Design

2.3 Values and Ethical Requisites of Ethics by Design

Ethics by Design is based on ethical values such as privacy and fairness. These are then instantiated as concrete ethical requisites against which systems can be evaluated. This section will outline both the values and the ethical requisites that were derived from them. The requirements below are to be used as guides to what actions should be taken in the development process.

The requirements under the Ethics by Design approach can be grouped into six value categories:

- Human Agency
- Privacy and Data Governance
- Fairness
- Well-being
- Accountability and Oversight
- Transparency

Under each category we describe the values and provide examples of corresponding ethical requisites for AI and robotics systems.

[4] Gebru et al., "Datasheets for Datasets".

The following format is used:

Title of Value

Several paragraphs of text to explain the value.

General ethical requisites

A generalised description of operational and other features which are required of AI systems in order to meet the ethical requisites.

Human Agency

Human agency encapsulates the values of autonomy, dignity and freedom. These are the fundamental rights upon which the EU is founded. They are also the rights enshrined in the UN Declaration of Human Rights. Respecting autonomy means allowing people to decide for themselves what is right and wrong and the way they should live their life as a consequence. Dignity means every human being possesses an intrinsic worth which should never be compromised by others, including AI. Humans derive dignity from their capacity to determine what is right and wrong for themselves (their autonomy). This means they have the right not to be treated as a tool in the service of others or as "a means to an end", but as a unique entity of inherent worth.

Human autonomy can take many forms. This is because autonomy means each person deciding for themselves what their own personal form of autonomy is. Consequently what constitutes human autonomy is as varied as people. As a result, systems can restrict human autonomy without doing anything - simply by not catering for the full range of human variation in lifestyle, values, beliefs and all the other aspects of our lives which make us unique. This is often done with the best of intentions – restricting choices or decisions to those which the developers consider optimal, often simply because they never realised other people might think differently. This is a particular problem with personalisation services, which may not cater for some lifestyle choices, or which fail to respect cultural norms in other societies.

Respecting freedom means leaving people free to exercise their autonomy and live with dignity. Most importantly, freedom requires individuals have the ability to decide for themselves any matter which they think is so important that they to want to decide it for themselves. Respecting dignity and autonomy means no one can tell another person that an aspect of their freedom is not important if that person thinks it is. In addition to the freedom to act, this includes freedom from constraints which conflict with one's autonomy, such as coercion, deception and manipulation.

General ethical requisites

- It should be clear to people whether they are interacting with an AI system. They should be informed about the system's abilities and how to judge and interact with them. This means that if an AI system is interacting with people, it should have specific features which inform people of the system's presence and abilities, including its limits.
- AI systems should not subordinate, coerce, deceive or manipulate people, and should not create attachment or stimulate addiction.
- Furthermore, AI systems should not limit freedom of expression, access to information, freedom of assembly and association, or any other rights.
- AI systems should not be designed for uses in which human beings are objectified or dehumanised.

Privacy & Data Governance

As a value, data governance means humans must actively manage their personal data and the manner in which the system uses it. Data governance includes issues relating to quality and accuracy of data, access to data, as well as other data rights such as ownership. Ethical issues can arise from both non-personal data (e.g. racial bias) and personal data (where the data subject's rights and freedoms must be safeguarded).

General ethical requisites

- The processing of personal data requires careful consideration of the rights and freedoms of the data subjects. These should be safeguarded at all times. For more information and guidance please see the EU's *Guidance Note On Ethics And Data Protection*.[5]
- AI systems should support the right of an individual to withdraw consent for the use of their personal data. This means there needs to be a mechanism in place to allow them to object to its use.
- Where personal data is processed, an AI system should process it lawfully, fairly and transparently, in line with data minimisation principle.
- GDPR and similar regulations require that technical and organisational measures be in place to safeguard the rights and freedoms of the data subjects through measures such as anonymization, pseudonymisation, encryption, and aggregation.
- Strong security measures to prevent unauthorised access, data breaches and data leakages should be set in place (such as limiting access to qualified personnel, mechanisms for logging data access and making modifications).
- Data should be acquired, stored and used in a manner which can be audited by humans.

Fairness

'Fairness' is used here in a philosophical sense, not to be confused with mathematical fairness or use of the term within computational modelling. Fairness in this context has three possible meanings, depending on the context; sameness, deservedness, and compliance. Sameness means that each person is treated the same. Deservedness means ensuring an equitable distribution so that each gets what they deserve. Fairness as compliance means operating in compliance with relevant rules.

Fairness means that all people have the right to be treated appropriately and not on the basis of irrelevant characteristics. In this sense, non-discrimination is the application of fairness in the context of human characteristics. In particular, people should not be treated unfairly on the basis of aspects of their identity which are inalienable and cannot be taken away from them. The most important of these are gender, race, age, sexual orientation, national origin, religion, health and disability.

General ethical requisites

- *Avoidance of algorithmic bias:* AI systems should be designed to avoid bias in both input data and algorithm design. Bias is a specific concern which needs specific mitigation techniques. AI development should contain specific steps to ensure data about people is representative and reflects their diversity. Similarly, the development process should have formal plans to look for and avoid errors in the selection of input data and in the algorithmic design which could cause certain groups of people to be represented incorrectly or unfairly. This needs to consider

[5] See https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf

inferences drawn by the system which have the potential to unfairly exclude certain groups of people from consideration.

- *Universal accessibility:* AI systems should be accessible to, and usable by, different types of end-users with different abilities and means - taking into account relevant criteria such as cognitive ability, special needs and access to certain types of hardware or software. Wherever possible Universal Design principles (Smith and Preiser, Universal Design Handbook.) should be used throughout the planning and development processes.
- *Fair impacts:* Before an AI system is created, developers should formally assess possible social impact on relevant groups. If necessary, steps should be taken to ensure the system does not cause them to be discriminated against or stigmatized, or otherwise have their interests affected in a negative way.

Personal and Social Well-being

The term 'well-being' covers a range of properties. Something has well-being when its needs are met and it is able to function properly. The values of autonomy and freedom mean that people can only achieve well-being if they are able to work towards their ambitions and live whatever they consider to be a "meaningful" life.

General ethical requisites

- AI systems should take the welfare of end-users and other stakeholders into account and not, on balance, reduce their well-being. An AI system developer should be able to identify who the end-users will be and any other possible stakeholders before constructing the system. Such planning should consider whether the system could reduce their well-being and, if necessary, how this risk will be mitigated.
- AI and robotics development should be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. Adverse environmental impacts should be avoided. When planning or purchasing a system, one should consider the environmental impact of the system and, where possible, the steps which can be taken to reduce it. In the case of robotics systems this should include considerations of the materials used, their origins and what will happen to them when the device is decommissioned.
- AI and robotics systems should not reduce safety. Robotic systems which will share an environment with humans or other animals, should possess appropriate safety features. AI systems which can control actuators, open or close doors or windows, activate lighting or signage, or make other changes to the physical environment should contain safety features to ensure the system does not trigger a change which could harm someone (such as opening a window while someone is leaning on it). Robots should have safety-aware collision avoidance mechanisms. Software systems may also need some form of safety planning where relevant. Ideally, such software should be compliant with IEEE P1228 (Standard for Software Safety).

Accountability and Oversight

Human oversight as a value requires humans are able to understand, supervise and control the design, development, deployment and operation of AI and robotics systems. Oversight depends on accountability because one cannot understand or control something unless one has information about it. Accountability means there are mechanisms to explain how, and why, a system exhibits particular characteristics.

General ethical requisites

- AI systems should allow for human oversight regarding their decision cycles and operation, unless the developer can clearly provide compelling reasons why such oversight is not required. Therefore, AI-driven systems should include concrete functionality which will enable humans to understand the decisions made by the system and allow humans to override them or correct erroneous learning outcomes.
- To ensure ongoing ethical status one needs to be able to detect ethically undesirable effects of the system on end-users or subjects. The organisation using it needs to have a plan for how to stop those effects. This means the system must also be designed with mechanisms to correct the AI behaviour so that these effects do not recur.
- There needs to be a formal ethical risk assessment for any proposed AI system. There also needs to be a procedure in place for risk assessment and mitigation after deployment. This is largely the responsibility of the operating organisation, but the system will need to include features to implement this.
- The operating organisation will need formal procedures so that third parties (such as suppliers, end-users and workers) can report potential ethical concerns about the AI system. Mere reporting is not enough; it needs to be evaluated and actioned. The requirement for transparency of AI systems means there also needs to be a procedure by which to communicate with those raising concerns what has been done with their information.
- The operating organisation will need processes by which data subjects can complain if they feel they have been negatively affected by the system. There need to be mechanisms for redress. The AI system needs to have functionality which can implement such redress if necessary.
- AI systems should be auditable by independent third parties, through the establishment of mechanisms which facilitate auditability. Increasingly, this is becoming a legal requirement. Ideally, development processes should follow best practice in XAI[6]. Before a designer starts construction, they plan for the facility for ethical audit of the system. This is not limited to auditing the decisions (or other outcomes) of the system itself, but will need to consider tools and procedures used during the development process, including learning models, data sources, annotation processes and decisions made to address potential ethical issues (such as bias in datasets). Where relevant, AI systems should generate human-accessible logs of the AI system's internal processes, input, output, and positive and negative impacts.

Transparency

Transparency directly enables human agency, data governance, oversight and human governance. Transparency includes *all* elements relevant to an AI system: the data, the system and the processes by which it is designed, deployed and operated. Without this level of transparency, a decision cannot be contested, or even understood. This would make it impossible to correct errors and unethical occurrences.

General ethical requisites

- The degree to which transparency is needed depends on the context and the severity of the consequences. However, it is important to note this is a judgement call, not a precise calculation, and others may not set boundaries or assess severities in the same manner, so the precautionary principle dictates it is better to go too far than not far enough. This is why we recommend, if

[6] Explainable AI. See https://ieeexplore.ieee.org/abstract/document/8466590 for an overview.

possible, that these decisions are made by a carefully constructed group, whose composition is sufficiently diverse so as to ensure a representative range of perspectives behind these decisions. At minimum this calls for a mix of genders and ages. Where the formation of a formal group is not possible, it is recommended you take steps to ensure you understand the full range of positions others may take, rather than simply rely on your own opinions.

- There is a general requirement for traceability across all areas of ethical AI and robotics. A system design (and the processes of construction) should include measures to facilitate the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit.
- The purpose, capabilities, limitations, benefits and risks of the AI system and of the decisions conveyed by it should be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly. This ethical requisite is fundamental to meeting a number of the requirements above and is referenced in those sections. Wherever it is necessary that people can audit, query, dispute or seek to change AI or robotics activities, such as decisions or learning processes, the system will need formal mechanisms to make this possible. However, this does not end with the system itself; it needs to include governance and other organisational processes (including in the development stage, such as code documentation) by which to receive and assess and address requests from third parties.
- The design and development processes will involve making decisions about ethical issues, such as how to remove bias from a dataset. The requirement for transparency means AI development processes (and tools) need components to keep records of such decisions, so that it is possible to trace how these ethical obligations were met. This information may be required for audits, for disputing or resolving decisions made by the system, for correcting unexpected ethical issues which arise after system deployment and so that organisations can learn from the experience and improve their handling of ethical issues.
- It needs to be made clear to end-users that they are interacting with an AI system – especially for systems that simulate human communication, such as chatbots. An AI system should have specific features to do this. These features should not depend on particular educational backgrounds, technical knowledge or other skills which cannot be assumed of all people.
- Decisions made by an AI system should be explainable to users. Where possible this should include the reasons why the system made a particular decision. However, with some systems this may not be possible. Nevertheless, the system (or those deploying it) should always have a mechanism by which to explain what the decision was and what data was used to make that decision.

2.4 Conclusion

Ethics by Design is an approach for ensuring that an AI or robotics system complies with important ethical values. These values give rise to ethical requisites which a system must comply with. Some of these relate to the functionality, while others relate to the processes by which systems are constructed. While many of these values are based on fundamental rights enshrined in EU charters and legislation, they are not specific to the EU alone, but reflect a growing global consensus. They are sometimes backed by legal force, but conformance cannot be achieved simply by adhering to legal obligations. As with Privacy by Design, Ethics by Design calls for more than just specific features or

functionality in the system. Supporting organisational processes are also required, as are specific features in development tools and methodologies.[7]

3 How to apply Ethics by Design in AI development - a practical guide for system developers

This following section uses a generic model of system development to detail the specific points any AI-driven system should address in order to achieve the ethical requisites discussed above. The aim of this section is to be of immediate practical use when designing a system. We therefore list these as concrete tasks a designer should undertake as much as possible.

Ethics by Design uses a generic model of the design process by which software systems are produced. Under this approach the ethical concerns to be addressed are treated as general system requirements, just like reliability – as requirements any and all systems must achieve. Just like reliability, these requirements of an ethical nature place obligations not only on the system's features, but also on the development processes and tools themselves. Our model of the development process explains how to embody these ethical factors in the design and development processes. This model positions ethical requisites in the phase of the construction process to which they relate as concrete tasks to be undertaken. By mapping their own development methodology to the generic model, a developer can determine the relevant ethical requisites for each element of their methodology. Ethical requisites will then be instantiated in that methodology as tasks, goals, constraints and similar guidelines. If these are adhered to, the chances of ethical concerns surfacing are minimised because each step in the development process will contain measures to prevent them arising in the first place. This is the essence of Ethics by Design – don't allow ethical issues to arise in the first place.

This chapter describes the generic model, then outlines the steps required to use it so as to incorporate Ethics by Design into the development process.

In this section we directly address developers. However, managers can use this material to understand what is required of their developers, while purchasers of AI systems can use it to understand what they should look for when seeking ethical AI products.

Preliminary Note for Development Managers

Moving to Ethics by Design is a form of business process reengineering, with all the attendant difficulties – political, organisational, managerial and financial.

Ethical AI cannot be achieved purely through the characteristics of the system produced. Ethics by Design is, by definition, embedded into the development process itself. This means developers must change how they work. This is likely to require alterations to team structures, such as adding additional roles, creating new communication channels, building additional review and decision processes. It will also require additional tools and alterations to existing ones. Some tools, such as version control systems, may be completely incapable of modification to the degree necessary and may need to be

[7] Cavoukian, *Privacy by Design: The 7 Foundational Principles*.

replaced. Some aspects of development will be more onerous than before because they will add additional documentation requirements or additional considerations. In many ways, moving development to Ethics by Design is similar to moving from one type of programming language to a different type, such as moving from a procedural language to an object-oriented one.

Many developers, especially senior ones, depend for their status amongst their workmates on their expertise in the way the organisation codes. It is well-known that the best developers will be 10 or 100 times more effective than the average. No organisation should lose them to the competition because they are unhappy with the changes required. Best practice is to use these individuals to lead these changes, rather than impose them from above. Resistance is inevitable unless the political dynamics of the development teams are taken into consideration.

As a result, it cannot be assumed that moving to Ethics by Design is simply a question of adding some considerations to the design of a system. Significant changes in the work environment are inevitable. A manager's responsibility is to ensure these changes are not disruptive, but rather enhance the motivation of staff and the productivity of the development processes. Only when the organisation's culture values ethical AI to the same degree it values reliability, profitability or customer satisfaction can full compliance with Ethics by Design methodologies be expected of all staff. This requires commitment from the highest levels of management.

3.1 Generic model for design

Ethics by Design is premised on the basis that development processes for AI and robotics systems can be described with a generic model. This model involves six broadly described tasks (sometimes also called phases). While the six are presented here in a list format, *this is not necessarily a sequential process*. For example, some methodologies, such as Agile, use cyclic models.

The six tasks in the generic model are:

1. *Specification of objectives.* This is the determination of what the system is for and what it should be capable of doing.
2. *Specification of requirements*. This is development of the technical and non-technical requirements and constraints by which to build the system. This includes initial determination of required resources, together with an initial risk assessment and cost-benefit analysis, resulting in a design plan.
3. *High-level design.* This is the development of a high-level architecture and is sometimes preceded by the development of a conceptual model.
4. *Data collection and preparation.* Data must be collected, verified, cleaned, formatted and integrated.
5. *Detailed design and development.* This involves the actual construction of a full working system. For software development, this will involve programming and coding. Robotic systems will also include a manufacturing component.
6. *Testing and evaluation.* This is the process of testing of the system and evaluation against the original objectives and requirements.

We will now briefly indicate how the ethical requisites can be instantiated in procedures within each task.

Specification of objectives

As part of the specification of objectives task, the system's objectives need to be evaluated against the ethical requisites presented in *Section 2.3: Values and Ethical Requisites of Ethics by Design*. Some objectives are not ethically permitted under any circumstance. For example, a system cannot be ethical if its objective is to destroy people's freedom because the objective of the system itself is to directly violate an important value. If it is possible to adapt the objectives so as to make the system ethical, this should be done before proceeding further. If the aim is fundamentally incompatible with the ethical requisites, the project cannot proceed. *Not everything which can be done should be done.* It is possible that whether the system meets its ethical requisites or not depends on specific methods construction or the exact manner in which some functionality is implemented. If this is the case, proceed, but maintain an ethical watch over the rest of the process and understand that some aspects of more detailed design will have ethical importance.

Specification of requirements

During this phase, design requirements and constraints, selected resources and design plans are assessed against the ethical requisites. At this phase one should determine how features of the system and the construction process facilitate meeting the ethical requisites. For example, it may be found that transparency cannot be achieved using a particular coding methodology or that version control systems need additional components to record decisions taken regarding code changes. Make modifications to enable attainment of the ethical requisites. Ensure that the ethical requisites are included in the final list of product requirements. Ideally, stakeholders should be included in this process.

High-level design

High-level design is concerned with the development of the technical and non-technical requirements of the proposed system, and the mechanisms by which this will be achieved, such as version control systems. This often includes initial determination of required resources, together with an initial risk assessment and cost-benefit analysis. This frequently involves high-level architectural design, such as overall database and application layer architectures, perhaps some critical schemas, information flow and security requirements. In many cases this will also include a hierarchical breakdown of the required sub-systems and critical sub-functions within the system, though some will consider this a part of detailed design.[8] Under the Ethics by Design approach, the high-level architecture is developed in accordance with the ethical requisites. Ethical requisites should be treated just the same as any other requirements for the system. Issues that may be particularly relevant in this design phase are those relating to transparency, autonomy, privacy and fairness. Design should include functionality by which to programmatically support ethical requisites, such as keeping logs of internal data manipulation by the system. The requirements for transparency and human oversight will typically require additional features beyond what is required to achieve the system's aim.

Data collection and data preparation

Data collection is an especially critical phase as far as ethics are concerned. Fairness and accuracy are the primary concerns here. It should be assumed any data gathered is biased, skewed or incomplete until proven otherwise. In general, data gathered from human activity within any society, such as

[8] Kission, Ding, and Jerraya, "Structured Design Methodology for High-Level Design".

written communication, employment patterns or criminal sentencing, can be assumed to reflect the biases in that society. Data can never be *assumed* to be accurate, representative or neutral; it must be demonstrated that it is.

Preparation of data itself may introduce issues. Steps should be taken to ensure testing, learning and algorithmic manipulation do not introduce new biases or other ethical issues (such as de-anonymisation). A frequent problem arises where testing does not accurately reflect the real-world use after deployment. For example, many facial recognition systems perform poorly with darker-skinned people due to testing on purely Caucasian populations.

Detailed design and development

In the detailed design and development phase, actions which will incorporate the ethical requisites are added to the various subtasks within the detailed design, as well as to the development infrastructure (tools, methodologies, procedures, and anything else which effects exactly *how* something is built).

Testing and evaluation

As part of the testing and evaluation phase, an ethical assessment is performed to see if the system meets its ethical requisites. It may be that the system achieves its functional requirements, but not all ethical requisites. If this is the case, the system cannot be considered to have been successfully completed. However, the whole point of Ethics by Design is to avoid such an outcome. If rigorously applied, the Ethics by Design approach should prevent ethical issues getting to this stage of the development process. It is recommended that stakeholder consultation or involvement takes place during this phase.

Mapping the generic model onto your design methodology

The Ethics by Design approach is an addendum to design methodologies. It is intended to be grafted into whatever methodologies are being used in the project. For this reason, the Ethics by Design approach is intentionally methodologically neutral. Ethics by Design provides ethical guidelines by which a system can be designed and developed in a manner which ensures it is ethically safe at every stage of its life cycle. However, to be of practical use, the approach must be integrated into the design methodology. The generic model for design identifies six classes of task which must be accomplished in the creation of any AI or robotics system. While they have been presented as a list, that does not imply that they necessarily form a sequence. The development of every system must accomplish these tasks or the system cannot be created, but some, such as Agile, vary the sequence. Consequently, any design methodology must include these tasks in some way. Ethics by Design can therefore be integrated into any design methodology by reference to our generic model.

The steps for integrating the Ethics by Design approach to any design methodology are as follows:

1. Identify where each of the generic tasks is undertaken within the target methodology.
2. All values are relevant to all tasks. However, not all ethical requisites will be relevant to all elements of the target design methodology. Identify which of the ethical requisites provided here are relevant to which element of the design methodology.
3. This will result in lists of ethical requisites under each element of the design methodology. Cross-reference these lists against the values to determine whether additional ethical

requisites are needed to fully cover the scope of the project. Formulate additional requirements as appropriate.

4. Review the project's aim, including final functionality and output, data sources and other forms of input and the context in which the system will be used. Consider whether the lists of ethical requisites are sufficient to cover these or whether additional requirements are necessary. It is highly likely that the intended use will generate context-specific ethical requisites. Other defining aspects of the project may also generate the need for specific ethical requisites. This is an especially important consideration where the system will offer unprecedented capabilities or have significant impact on people's lives.
5. If possible (and appropriate), develop formal systems for ensuring Ethics by Design within each element of the design methodology. At its most basic, this can consist of checklists containing the ethical requisites for each design methodology element. However, additional tools or systems may be required. Some of these may be publicly available "AI ethics" tools, such as Model Cards [9] or XAI components which are open source[10]. Others may be available as add-ons to existing development tools. Some may simply require additional configuration in existing development systems. For example, Git repos can simply be configured to include ethics-related documentation and tools.

3.2 Design Phase: Specification of objectives

General Notes and ethics guidelines

While each project is unique, Ethics by Design outlines a set of standardised requirements which all AI, robotics and big data systems should meet. For obvious reasons, an important first step is to ethically assess the objectives of a development projects (i.e., what kind of technology is being developed and what it its intended functionality and purpose) against the ethical requisites, before any details of the individual project are considered. Sometimes, objectives are unethical or even illegal. For example, it cannot be an objective of a system to deceive people by collecting personal biometric data from them without their consent and using AI to hide this activity.

The two ethics guidelines for this design phase are the following:

- Assess whether the formulated objectives for the design project will permit the system to meet the relevant ethical requisites. An ethical risk assessment, to be performed later in the specification of requirements phase, can also be applied retrospectively to the objectives, as it may point to further potential ethical issues with them. It is recommended that a professional AI ethicist, if available, is enlisted to assess the objectives, in collaboration with members of the development team.
- If a project has external stakeholders it is important to plan how to include them in the early phases of the project, especially the specification of objectives and specification of requirements phases. The early inclusion of stakeholders increases the chance that their values, preferences and needs are taken into account, and thereby increases the likelihood that the resulting technology is successful and trusted and attains its ethics requisites. In particular, stakeholders may be aware of

[9] Mitchell et al., "Model Cards for Model Reporting".

[10] Source code for Model Cards can be obtained from https://github.com/tensorflow/model-card-toolkit. The git repository for XAI contains a growing set of tools and methodologies at https://ethicalml.github.io/xai/index.html

other ethical issues which could arise from the use of the system. Stakeholders should be consulted about their preferences regarding what the objectives and requirements should be, their beliefs about what ethical issues are at stake and their recommendations about how these ethical issues should be dealt with. Moreover, it is recommended that project members and stakeholders represent appropriate diversity in terms of e.g., gender, age, ethnicity, cultural heritage and viewpoints. In this way an appropriately diverse range of ideas and preferences will inform design choices.

Ethical Requisites of Design Objectives

The objectives of the proposed system should be checked against the ethical requisites listed below to see if they are potentially violated. Potential violations differ in their degree of seriousness. If it is possible to adapt the objectives so that the system does comply with the ethical requisites, this should be done before proceeding further. If the aim is fundamentally incompatible with the ethical requisites, the project should not proceed. It is likely that whether the system meets ethical requisites or not depends on specific methods of implementation or construction. If this is the case, proceed, but pass these concerns to those designing the development architecture and maintain an ethical watch over the rest of the development process. Other violations may be only potential violations or be less serious in nature. These concerns do not mean the objective should be abandoned, but that concrete steps will have to be taken to avoid the system becoming unethical. For example, a voice recognition system which is trained only on people with a strong regional accent may be less reliable for people with a different accent. The solution in this case would be to ensure the training data includes a wide variety of accents.

In the assessment of objectives, also consider the proposed system's potential for misuse. Where possible, modify the system's objectives to reduce such potential. If the potential misuse is significant, conduct a social risk assessment outlining the risks, the elements of the design which will need to be included to mitigate this, and any procedures required to reduce this risk once the system is deployed and operational.

VALUE: Human Agency

- Check whether the objectives adhere to the human agency requirements. Serious ethical non-compliance is an issue for systems that limit human rights, subordinate, deceive or manipulate people, violate bodily or mental integrity, create attachment or addiction, or that hide the fact people are interacting with an AI system.

VALUE: Privacy & Data Governance

- Check whether the objectives are compatible with privacy and data governance requirements. Non-adherence to any of these would result in serious non-compliance.
- An additional consideration is whether the initial plans for what personal and non-personal data will be used is lawful, fair and appropriate. For example, it would be both unfair and inappropriate to build a system which assesses people by irrelevant characteristics. If the proposed data source is unfair or inappropriate, either change the data source or modify the objective so that the unfair/inappropriate data source is not needed.

VALUE: Fairness

- Check whether the objectives are compatible with the fairness requirements. Particularly important is the consideration of whether violation of any of these requirements would cause

people to be significantly disadvantaged socially or politically, reduce the control that they have over aspects of their lives, such as work or lifestyle, or would likely result in discrimination or stigmatisation, either through direct actions by the system, or by likely uses to which it would be put. If so, this would constitute serious non-compliance.

VALUE: Well-being

- Check whether the objectives are compatible with the well-being requirements. Particularly serious are those violations that cause people to suffer physical, psychological or financial harm, support processes that are known to cause significant environmental damage, or that are likely to cause significant damage to social processes and institutions (for example, by contributing to misinformation of the public). Less serious violations are, for example, systems that are likely to inhibit communication and impoverish interpersonal relationships. If there is significant potential social or environmental damage which could result from use of the technology, a social and/or environmental impact assessment should be done (for projects that are of sufficient scale).

VALUE: Accountability & Oversight

- Most of the ethical requisites for accountability and oversight do not apply to objectives, but rather to the architecture and detailed design of the system. However, all systems should have an objective of allowing for human oversight and intervention regarding decision cycles and operations. If it does not, change the objectives or provide compelling reasons why such oversight is not required.

VALUE: Transparency

- The ethical requisites for transparency do not usually apply to objectives, but rather to the requirements, architecture and detailed design of the system. So they only have to be considered at this stage to determine the degree to which the system's objectives permit the required transparency to be built into the system.

3.3 Design Phase: Specification of requirements

General Notes

The primary function of the Requirement Specification phase is to arrive at a development plan that includes design specifications for the system, design the development infrastructure, determine staff resources required, set milestones and other deadlines and so forth.

Most organisations have a standardised set of development tools used for all projects. The organisational and management structures and procedures are usually tuned to these tools, as are the development methodologies. Changing these can be more challenging than building systems. Nevertheless, it cannot be assumed that any tool, process or organisational elements will be appropriate according to the ethical requisites of Ethics by Design. Some of the ethical requisites present new problems during development. For example, it is no longer sufficient to merely correct datasets for bias, developers also need to document that this has been done and how. It may even be necessary to document the reasoning which led to the use of a particular technique. Consequently, requirements such as the capacity for human oversight and audit may impose a need to document many internal processes to a greater degree than has previously been the case. For example, while

documentation within code has always been considered best practice, it has rarely been necessary and unavoidable to the degree Ethics by Design requires.

It must therefore be recognised it is unlikely that development systems, methods, tools or even organisational structures used on previous projects will be suitable without modification. Systems like git can easily accommodate the additional documentation requirements with a little planning, but others may be completely incapable of delivering the ethical requisites required. Even with a system like git, additional management procedures will be required to ensure developers produce the required documentation, and this will require staff training. As a result, it must be recognised there is likely to be a need to adapt (or even replace) aspects of customary development systems so that they become capable of delivering the project's ethical requisites.

In some cases, it may not be technically possible to meet every ethical requisite due to lack of suitable development tools. However, one should be extremely rigorous in investigations for suitable tools and cannot merely decide that the traditional methodologies are insufficient as an excuse not to bother. The requirements here are common demands of many AI projects. Consequently, tools to meet these needs are developing rapidly. For example, Model Cards [11] and Datasheets for Datasets [12] have been produced specifically to provide ethical documentation of important AI development processes. Meanwhile DARPA's Explainable AI (XAI) [13] is a rapidly developing set of methodologies and tools by which build effective machine learning techniques which are also explainable to humans and allow for human governance. In these and other cases, such tools are Open Source and freely available to all.[14]

The degree to which a technical inability to meet the ethical requisites blocks a project also depends on the particular ethical requisite in question and the system's functionality. For example, a system which approves personal loans must be able to explain each individual decision in a human-readable format because individual people will be profoundly affected by its decisions. By contrast, a system which manages a city's traffic lights has only a very limited impact on the life of individuals, so the need for transparency is much lower. Where it is genuinely technically impossible to meet a relevant ethical requisite, the importance of the requisite for that particular system will be a factor in the ethical status of the product.

Ethics guidelines

- An ethical assessment should be done of proposed design specifications, constraints, selected resources and infrastructure. For example, an early choice of deep learning techniques for a system that requires transparency and explainability may be judged not to be the best choice. For example, a design specification that a system use authentication via facial recognition may be undesirable from a privacy point of view.
- Once a complete design plan has been produced, an ethical risk and impact assessment should be performed to assess specific ethical risks that may result from development, deployment and use of the system. Steps should be planned and carried out to mitigate ethical risks. The ethical

[11] Mitchell et al., "Model Cards for Model Reporting".
[12] Gebru et al., "Datasheets for Datasets".
[13] Gunning, "Explainable Artificial Intelligence (Xai)".
[14] A working toolkit for producing Model Cards is available at https://github.com/tensorflow/model-card-toolkit. A template for Datasheets for Datasets is available at https://github.com/AudreyBeard/Datasheets-for-Datasets-Template. A python library of XAI tools is available at https://github.com/EthicalML/xai.

assessment of objectives and requirements undertaken earlier can be important constituents of this assessment, but these only assess individual elements of the plan, rather than the plan as a whole. This risk assessment should be updated at later points in the development process as more information relevant to it comes in. A professional AI ethicist, if available, should be able to perform such an assessment in collaboration with members of the development team. Ethical risk assessments are scalable; a simple assessment can often be completed within a time constraint of days and with limited resources, whereas a detailed assessment may involve extensive foresight analysis, stakeholder consultation, mapping of potential risks and development of mitigating actions. Ethical risk assessment should be planned and budgeted for at the appropriate point in the development processes. This assessment needs to be scaled to the innovative nature of the project, the severity of ethical risks that were already identified at the stage of ethics review, and the overall budget of the development project.[15]

- Ensure that relevant ethical requisites are covered in the list of design specifications. For this purpose, consider inclusion of an Ethical Requisites document for the project. At the Objectives stage this document will only cover ethical aspects of the overall system and the most obvious features of the development process. However, it can be refined and added to during the high-level design and detailed design stage of the project.

3.4 Design Phase: High-level Design

In high-level design, the architecture for a system or software product is specified. The following ethics guidelines apply to this phase in the development cycle.

VALUE: Human Agency

- Verify that the chosen architecture allows for an interface based on human-centric design principles which leave meaningful opportunities for human choice, and that it allows for freedom of expression and information.

VALUE: Privacy & Data Governance

- Verify that the chosen architecture supports the ethical requisites for privacy and data governance. Ensure the development architecture contains processes, procedures and tools to ensure that personal data is not exposed during development such that it violates the right to privacy. For example, error logs may needlessly include the personal data being accessed when a bug is encountered, or developers may be given direct access to database contents when all they need is the ability to query the database. It is especially important to ensure developers do not have access to identifiable personal information except where absolutely necessary. The GDPR (or other regulations) require that who has such access is formally documented.
- Ensure there are formal processes to guarantee the selection of data for the system will be fair, accurate and unbiased. Plan for an initial assessment of data sources before they are brought into the system. Design a mechanism to document and justify how the initial data selection was determined sufficient for external audit.
- Data imported into a system may not exactly match what was sought for. It cannot be safely assumed that the data obtained is the data wanted. It may be that the datasets do not contain

[15] A standard for ethical impact assessment has been developed in CEN working document CWA 17145:2017-2, retrievable at https://satoriproject.eu/media/CWA17145-23d2017.pdf

the data they were supposed to, the data may be incomplete or corrupt, or methods of importing or normalizing the data alter it in unanticipated ways which render it less than optimal. Design formal processes to check for and correct bias (or errors) after importing any data.

VALUE: Fairness

- Does the high-level design suggest that some users of the system will obtain better functionality than others? If so, prepare a formal justification for this differential access or modify the design because this could be challenged by stakeholders who feel disadvantaged or even legally challenged under disability access requirements.
- Examine the initial interface design and other touchpoints to see whether it is assuming a one-size-fits-all approach to users. If so, see if this makes using the system more difficult for some people. If this is the case, either modify the design to fix this or prepare a formal justification as to why this is impossible to defend against disability, discrimination or other access challenges.
- Undertake an accessibility assessment of the interface and other touchpoints. Ensure that, where relevant, the system meets accessibility standards.

VALUE: Well-being

- An initial rough environmental assessment should have been conducted during the objectives phase. Once high-level design of the system is complete, this assessment should be taken into more depth. In addition, documentation should be prepared to demonstrate how the system will be constructed in the most environmentally friendly way possible.
- Evaluate whether the system, as defined by the high-level design, could cause physical harm to people or property. This is especially important with robotic systems. If this is possible, ensure design features to minimise this risk and/or the amount of harm which can be done, such as safety buffers, emergency stop buttons. If the system will be able to respond to voice commands, include emergency stop vocal commands in the design.

VALUE: Accountability & Oversight

- Design the ethical governance model for the development process. This focuses on mechanisms which will enable human oversight during the development process. There are two main elements to design; a technical ethical compliance system embedded into the development architecture and a set of organisational structures and procedures. The ethical compliance architecture will need to focus on tools and processes at the developer level, such as Model Cards [16] and Datasheets for Datasets [17], but will also need mechanisms for external communication from end-users and other stakeholders during testing and evaluation. The ethical governance model will also need to include organisational structures for actually doing the governance, such as ethical review committees and/or ethical compliance officers at developer level. Ensure such mechanisms include management procedures which ensure these mechanisms are actually used, such as formal reporting and assessment by senior management. The governance model needs to address the following issues: How will governance be exercised? What is the project's version of an authority to supervise and ensure the ethical requisites are met? What powers will that authority have? How will it be selected? How can that selection process be demonstrated to be fair and inclusive? What procedures

[16] Mitchell et al., "Model Cards for Model Reporting".
[17] Gebru et al., "Datasheets for Datasets".

will be used in the case of a conflict between the ethical governance authority and developers or engineers or clients?

- Design mechanisms for human oversight and external audit once the system is deployed. This may require additional functionality inside the system solely for reporting internal activity and which has no role in the system's purpose. It may be possible to design oversight during development which can also be used once the system is deployed, but this may be more difficult than simply designing a different mechanism tuned to oversight of the deployed system. In either case, oversight after deployment will need access to the oversight work performed during development.
- Design a testing regime which can check that the system's internal operations meet the ethical requisites. This may require changes to the way functionality is achieved within the system so as to permit appropriate testing and remedial action.
- There is an increasing tendency to demand AI systems can be externally audited for ethical compliance. Even if this is not the case for your type of project now, audit requirements could arise during the lifetime of the system. Ensure that the system is designed in a manner which permits such auditing. If unsure, refer to existing ethical audit procedures, codes of conduct for ethical AI.

VALUE: Transparency

- Design mechanisms to document how data acquisition, storage and use happen. This needs to be auditable by any people who need to check for ethical compliance, including users and other stakeholders, those responsible for ensuring the data practices fulfil the ethical requisites, external ethical AI auditors, regulators (where regulations apply) and any other person who has a need to determine the ethical status of the system's data and use thereof. This consideration must cover both the development process and use once deployed.
- Design procedures and select and configure tools to document development processes to a level that humans can understand and evaluate decisions made within the design and development processes. This will be required for any people who need to check for ethical compliance within the development process. This can be anyone who has a concern the system is unethical in some way and wants to determine if this was caused within the development process, or who simply wants to understand how ethical concerns were dealt with while the system was being created. This can therefore include users and other stakeholders, those responsible for ensuring the created system meets its ethical requisites, external ethical AI auditors, regulators (where regulations apply) or managers assessing the development process in order to emulate it in other projects or to determine how to improve it. We recommend a layered approach to this documentation, so that it offers a range of technical detail, commencing with basic overviews, such as executive summaries, down to detailed schemas and other technical models. In this way people can be provided documentation appropriate to their level of expertise and their specific concerns.
- Ensure the design includes mechanisms by which the AI system will record its own decisions so that they can be subject to human review. Such review could occur through a post-deployment audit, if data subjects or end-users question system behaviour and want justification, explanation, or alteration, as part of a normal internal ethical governance review, because this information is required by developers developing other aspects of the system, or for other reasons.

- Design features and functions which will enable the capabilities and purpose of the system to be openly communicated to users and anyone else who may be affected by it.
- Ensure ethical documentation systems are sufficient to make ethical issues identifiable and their resolution traceable and explainable.
- Design mechanisms so that people will know when they are being subject to the decisions of the system. This may include operational procedures to be used once deployed. Such systems should be targeted for evaluation as part of the testing regime.
- Ensure there is no aspect of the AI system which could be mistaken for a human once the system is deployed. Bear in mind many people may not have a nuanced or educated understanding of AI operations and can innocently assume they are interacting with a person. For example, even when labelled as such, chatbots can be mistaken for humans by those who do not know what the term 'chatbot' means [18]
- Ensure processes exist, and are actively maintained, by which internal staff and third parties (e.g. suppliers, consumers, distributors/vendors) can report potential vulnerabilities, risks, or biases in the system, during the development process.

3.5 Design Phase: Data collection and preparation

General Notes

For systems that involve data processing, data must be collected, verified, cleaned, formatted and integrated. Data collection involves the collection of initial data, its description and initial analysis, and verification of quality. To integrate ethical requisites into this process, assess how different steps in the process might support or violate ethical or data protection requirements. Make necessary changes as a result. If appropriate changes are not possible, the design objectives may need to be re-evaluated. In this phase, fairness (including bias, discrimination, and diversity), privacy and data quality will be particularly important.

The processing of personal data is governed by the GDPR in the European Union and the specific national and sectorial legislative frameworks. Personal data is any information that relates to an identifiable living individual. Items of information which have the capacity to be amalgamated and then identify a particular person also constitute personal data, whether being so used or not. Online identifiers and location data also constitute personal data. Personal data that has been de-identified, encrypted or pseudonymised but can be used to re-identify a person is also personal data. Personal data that has been rendered anonymous to the degree that the individual is no longer identifiable is not personal data. However, the anonymisation must be absolutely irreversible. Special categories of data (also often called sensitive data) are a subset of personal data which is particularly sensitive and must be treated with special attention. Such data are: data concerning racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic, health related data, biometric data for the purposes of uniquely identifying a natural person, data concerning a natural

[18] Candello, Pinhanez, and Figueiredo, "Typefaces and the Perception of Humanness in Natural Language Chatbots"; Castelo, Schmitt, and Sarvary, "Robot Or Human? How Bodies and Minds Shape Consumer Reactions to Human-Like Robots".

person's sex life or sexual orientation and. Special rules may apply to the processing of data related to criminal convictions and offences.

VALUE: Privacy & Data Governance

- This document is not a definitive guide of obligations regarding data processing. For detailed information and guidance consult your data protection officer(s).
- Whenever your system is processing personal data, you must comply with the data minimisation principle. This means that you must ensure that only data which is relevant, adequate and limited to what is absolutely necessary is processed by your system;
- All personal data must be processed in lawful, transparent and fair manner. If the planned system will process personal data, you must incorporate the rights to data protection into your design. This includes how to enable individuals to withdraw consent for the use of their personal data, and what mechanisms will enable them to object to its use.
- Within the limits of current technology, the design should ensure that data controllers and data processors are able to fulfil their data protection obligations.

VALUE: Fairness

- Is it possible some of the data gathered could be biased in its representation of different groups, persons, or social entities, for example by overrepresentation of some categories, a lack of diversity in representation, or implicit stereotyping? If so, modify the criteria by which data will be selected to reduce such bias and/or plan steps to rectify the datasets once they are in the system. The requirements for transparency and oversight will demand that such rectification is documented.
- Analyse your training data and ensure that your data is representative and value-aligned.
- Undertake a formal bias assessment of the data imported into the system. Do not *assume* any data imported into the system is unbiased – test it. Assess the diversity and representativeness of users in the data, testing for specific populations or problematic use cases.
- Ensure that input, training and output data is all analysed for harmful bias (e.g., some data sets may contain harmful biases if they consist solely of the behaviour of subclasses of all people, e.g., young white men, and if the system is deployed in situations where groups other than those in the data set will be affected).
- Where it is determined that harmful bias is possible, build mechanisms to avoid or correct it.
- Make sure data from one demographics group is not used to represent another unless it is justifiably representative.
- Evaluate the potential for harmful bias being introduced during the data preparation stage (e.g., the cleaning of the data set may inadvertently remove data relating to certain minority or under-represented groups, leaving the data set as a whole biased). Take steps to mitigate any such risk.
- Ensure that, whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time *t* and, whenever possible, determine which training data was used.

VALUE: Accountability & Oversight

- Many organisations processing personal data are required to have a data protection officer or similar, so if your organisation is one of these, it is highly recommended that they are consulted on the appropriate requirements for the project.
- Build a culture of shared responsibility for the organisation's data assets and that the potential value of data assets is acknowledged. Ensure that employees understand the true cost of failing to implement a data quality culture.

- Make sure that roles and responsibilities are clear for governance and management of data assets and that all employees and stakeholders understand them.
- If using external organisations for data storage, such as cloud services, ensure these are also compliant with data protection requirements. It is not safe to assume their assurances are sufficient. The GDPR requires that you verify their practices are compliant yourself.
- Make sure you have clearly established what kind of sample you need, what kind of sample you have taken, and that you can articulate what it will be used for.

VALUE: Transparency

- Prepare a data protection document which details how the project complies with data protection requirements. This will be needed for those concerned with ensuring compliance with the ethical requisites, data protection officers and regulators, and for ethical audits. This is a mandatory requirement under data protection regulations.
- You must carry out an analysis of the ethics risks related to the data processing and produce a risk mitigation plan.
- Ensure that you can explain to others how personal data is used, shared, and stored and for how long.

3.6 Design Phase: Detailed design and development

General Notes

To a large degree this phase involves adding more detail to the ethical requisites of the system, and to designing and implementing an ethical development architecture. Just as Ethics by Design calls for ethical matters to be dealt with during the development phase, so existing development tools and processes will need adaptation to support this activity. To integrate ethical requisites into this process, ensure that ethical guidelines are communicated to all developers and engineers, and that the design is evaluated relative to these ethical guidelines by them wherever they need to make decisions regarding them. Issues that may be particularly relevant in this design are those relating to transparency, privacy and accountability.

Detailed Design and Development

VALUE: Privacy & Data Governance

- If creating new personal or sensitive data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets), further informed consent may need to be acquired. Please remember this document does not offer definitive guidance on GDPR compliance, and more authoritative information in this regard should be sought from the data protection officers or the data protection authority.
- Ensure all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
- Ensure no new personal information is, or can be, collected or created during development of the system, unless necessary. If new personal information is collected or created, then have systems in place to impose access or use limitations which will protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.

- Ensure there are processes to safeguard the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked. If in control of the quality of the external data sources used, assess to what degree the quality can be validated.
- Establish a developer culture of shared responsibility for the organisation's data assets. Make sure this culture understands the potential value of data assets. Ensure the impact and risk of data loss is continuously communicated and that employees understand the true cost of failing to implement a data quality culture.
- Make sure that roles and responsibilities are clear for governance and management of data assets and that all relevant staff understand them. Review as required.
- Be aware that once data is anonymized, it may be possible to de-anonymise it.
- Ensure there is an embedded process that allows individuals to access their data and remove it from the system and/or correct errors in the data where these occur. AI systems must support the right for someone to withdraw consent for the use of personal data or object to its use. If required by law (which it is in the EU), it should also support the right to be forgotten (from internet searches and directories). Steps must therefore be taken to guarantee a person can access their personal data, and in a manner which protects other individual's privacy.
- Make sure no new personal information is, or can be, collected or created during regular use of the system, unless necessary and in accordance with the law (e.g., for the function of the system or realisation of the business or research objectives).
- Institute both technical and organisational measures to achieve data protection by default (such as Privacy by Design methodologies), including through measures such as encryption, pseudonymisation, aggregation, anonymisation and data minimalization (especially for personal data).
- AI systems used for commercial purposes must respect data portability, meaning that a person can download their personal data and move it to a competitor. The design must therefore ensure any individual's personal data can be exported from the system and that the loss of this record will not damage the system's functionality.
- Ensure there are oversight mechanisms for data processing (including limiting access to only appropriate personnel, mechanisms for logging data access and making modifications).
- Data can be manipulated, damaged, lost or inappropriately exposed within any system. Design processes to check for on-going degradation in the ethical quality of the data (i.e.: accurate, fairness, appropriateness, security) prior to its use by the system. This should include measures to prevent external corruption, such as hacking. Ensure the data integrity systems are designed to prevent unauthorised manipulation of data and to mitigate against silent and other forms of low-level data corruption.

VALUE: Fairness

- Check for algorithmic bias during the detailed development phase. Data could be processed in a biased way, and therefore algorithms should be checked for this.
- Ensure that interface design honours principles of universal accessibility, and avoid the introduction of functional biases in the detailed development phase which could make the system unequally functional for different types of users.

VALUE: Well-being

- Follow resource-efficiency and sustainable energy usage practices. In particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities.

VALUE: Accountability & Oversight

- Create a developer culture in which it is seen as important to deal with ethical issues in a timely fashion. Do not allow a culture to develop in which dealing with ethical issues is seen as a hassle, after-thought and something to be addressed after "more important" work is completed. Most importantly, make sure a culture does not arise in which departures from the ethical requisites are treated as something to be fixed after the entire system is completed.
- Create mechanisms by which concerns raised by staff and third parties can be assessed and, if necessary, acted upon. Ensure any such steps are taken before development continues.
- Audit controls may need to be deeply embedded into the system. Ensure that audit controls are built to report performance and log the decisions made by the system.
- Build tools and mechanisms into the development architecture to trap important information relevant to ethics assessment, such as the source of datasets and the nature of models used. Ensure staff are trained and encouraged to use them.
- Refine and complete the project's ethical requisites document. This is likely to be an iterative process. As much as possible, record any decisions taken regarding how the system was made compliant with its ethical requisites.

VALUE: Transparency

- Measurements to ensure traceability to the degree needed should be established within the following methods:
    - Methods used for designing and developing systems, such as the models built, the training methods, which data was gathered and selected, and how this occurred).
    - Methods used to test and validate systems, such as the scenarios or cases used to test and validate; the data used to test and validate; outcomes of the system (outcomes of, or decisions taken by, the system); other possible decisions that would result from different cases, e.g., for other subgroups of users.
    - A series of technical methods to ensure traceability (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.
- Make sure the code is actively explained and documented within the software program (as appropriate to the language(s) and methodology) and in appropriate ancillary documentation. Make sure documentation is understandable to fellow programmers and accessible by them.
- Make sure you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model.
- Use formal methodologies and tools to ensure explainability wherever possible and if considered desirable for the particular system that is designed, such as the XAI [19] or Transparency by Design [20] approaches and programmatic documentation, such as Model Cards [21].
- Could the system present false or misleading information to people? If so, add design requirements which will minimise this risk. In some cases, the risk is more likely once the system is operational. If this is the case, add documentation, functionality, or other steps to be used once the system is deployed to minimise misinformation.

[19] Doran, Schulz, and Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives".
[20] Rossi and Lenzini, "Transparency by Design in Data-Informed Research: A Collection of Information Design Patterns".
[21] Mitchell et al., "Model Cards for Model Reporting".

- Is it unavoidable that the system will manipulate data, or make decisions based on data, which cannot be traced or understood by humans? If so, add design requirements to expose data operations to scrutiny as much as possible and/or prepare formal justification to explain why data operations cannot, and should not, be audited. Note that intellectual property concerns are not sufficient. Black box and "test track" testing regimes can be used to externally assess internal data operations [22].

3.7 Design Phase: Testing and evaluation

The following general and value-specific guidelines apply to the testing and evaluation phase.

As part of the testing and evaluation phase, perform an ethical assessment to assess how well the system meets the ethical requisites. Possible outcomes are that ethical issues have been dealt with in a satisfactory way, that further development is needed, or that specific guidance for, or restrictions on, deployment and use need to be in place to mitigate ethical issues.

Use the project's ethical requisites document to design a testing regime to check the system's compliance with its ethical requirements. While some aspects of the ethical requirements are likely to be factors in normal testing, it is highly unlikely any standard testing regime will consider all of the system's ethical requisites. The choice of testing methodology is important here. For example, metamorphic testing is popular with machine learning [23] and can easily accommodate testing against ethical requisites if suitably designed, whereas techniques such as unit testing will need significant work to be a suitable testing methodology for ethical compliance (and highly unlikely to be capable of testing all ethical requisites). Implement and evaluate this testing of the system to determine whether it meets all of its ethical requisites. Treat departures from the system's desired ethical characteristics just as seriously as any other type of bug and undertake remedial work to make the system meet its ethical requisites.

It is highly recommended that stakeholder consultation or involvement takes place during this phase in order to collect their viewpoint on whether ethical requisites have been met in a satisfactory way and to discuss what should be done when this is not the case.

VALUE: Accountability & Oversight

- Ensure practical processes exist for third parties (e.g., suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system. Ensure mechanisms exist to examine and action such reports.
- The testing process should include testing the understanding and perception of the system's functionality and behaviour by end-users and other directly affected stakeholders. Even simple items like interface messages can be misinterpreted by those without a nuanced technical understanding. It cannot be assumed others will understand the system or its output in the same way as developers. Test the understanding of users and other affected persons regarding what the purpose of the system is, who or what may benefit from it, and (most importantly) what its limits are.
- Establish processes to obtain and consider users' feedback and ensure mechanisms exist to adapt the system in response, as appropriate.

[22] Aggarwal et al., "Black Box Fairness Testing of Machine Learning Models".
[23] Xie et al., "Testing and Validating Machine Learning Classifiers by Metamorphic Testing".

- Ensure users and stakeholders are given explanations they can understand as to why the system took a certain choice resulting in a certain outcome during testing so they can assess it accurately.
- Develop and deliver training to users to help develop accountability practices (including teaching about the legal framework applicable to the system).
- Formally attempt to predict the consequences/externalities of the system's operations.

VALUE: Transparency

- Ensure audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts). Ensure mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons behind the system's outcomes.
- Test whether users understand that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision in situations where not doing so would be deceptive, misleading, or harmful to the user.
- Ensure information to stakeholders, users and other affected persons about the system's capabilities and limitations is communicated in a clear, understandable and proactive manner, and which enables realistic expectations.

4 Ethical Deployment and Use

In this section, we will present ethics guidelines for the deployment and use of AI systems. We distinguish between the development process for an AI system and its deployment and use after development, and offer separate ethics guidelines for both. We take as our principal actors the project team, while also taking into account that they will be operating in an organisational context.

Our guidelines for Ethical Deployment and Use apply to four practices we consider central to the deployment and use of AI systems in research projects: project planning and management; acquisition, deployment and implementation; monitoring.

- *Project planning and management* refers to the planning of a new research project, normally reflected in a project plan, and the management of the planned activities after the project has begun. Our ethics guidelines address what steps should be taken by project management in project planning and general project management in order to ensure proper consideration of ethical issues in the deployment and use of an AI system.
- *Acquisition* refers to process of acquiring an AI system which is to be deployed and used in the project. In some projects, the system will be acquired from an external developer or vendor. In others, it will be developed in the project itself. A combination of external acquisition and in-house development is also possible. An organisation is responsible for the ethical state of any AI system it uses, even if that system has been built by another. As a result, external acquisition imposes unique ethical tasks not required when the system is developed in-house.
- *Deployment and implementation* refers to process of deploying the AI system into a user environment, and planning and implementing required changes in the organisational context to ensure its successful implementation. It normally involves the development of an implantation plan, the preparation and training of stakeholders, the development and implementation of an

operation and use plan, the configuration of the system and its imbedding in IT infrastructure, the testing of the system in its new environment, the implementation of needed organisational changes and new policies, and post-implementation review. The manner in which a system is deployed or implemented may change the ethical characteristics of the system. For example, the system may be deployed to work with different datasets from that on which it was trained. As a result, it cannot be assumed that the ethical characteristics of the system will remain unchanged when it is deployed. Deployment and implementation therefore imposes its own tasks to ensure the system continues to meet its ethical requisites.

- *Monitoring* is the process within project of monitoring the performance of the AI system, its conformance and compliance with external requirements, and the development and implementation of plans for improving its performance. No matter how robust the testing regime, the full ethical characteristics of a system may not be apparent until the system is deployed "in the wild." The most common (but not the only) concerns are that the system may have completely unexpected (and untested) effects on users; its own internal processes may change as it learns; or the data it uses may lose ethical integrity. As a result, all AI systems require perpetual on-going ethical monitoring and, where necessary, adjustment. This is typically done through audit procedure, which is becoming an increasingly common legal requirement.

We assume that all four of these processes take place when an AI system is deployed and used in a research project, and proceed to outline ethics guidelines for each of them.

4.1 Project planning and management

- In the project plan, ensure that you budget for Ethics of Use actions and include tasks or subtasks for these actions. The budget should be sufficient to ensure proper adherence to the Ethics of Use guidelines in the project. In budgeting and planning, take into account the potential ethical issues that were revealed in the ethics self-assessment.
- In the project plan, define roles, responsibilities and procedures for implementation of the ethics guidelines and for monitoring and assessment of their implementation. This could include the institution of an AI ethics officer (with the right expertise) and the assignment of specific responsibilities to implement ethics guidelines or monitor their implementation by researchers in the project. It should not be assumed that whoever managed ethical compliance during development, even if available, is the appropriate authority for this role.
- Ensure that the objectives for which you want to use the system and the design requirements and resource choices conform to the ethical requisites provided for in the Ethics by Design objectives and requirements phases.
- Your plan should include details of the procedures for inclusion of stakeholders in decisions regarding the acquisition, deployment, implementation and monitoring of the use of the system. These procedures must ensure that stakeholders are, at a minimum, consulted regarding their values and interests with respect to the deployment and use of the system.

4.2 Acquisition

- If an AI system is externally acquired as an off-the-shelf solution, consider available options and pick the system that is most capable of meeting the ethical requisites specified in *Section 2.3: Values and Ethical Requisites of Ethics by Design*. If a system does not meet the ethical demands contained in this guidance document, consider whether adaptations can be made to the system

or focus on acquiring a different system. If the AI system is custom-built by an external developer, then give preference to a developer who uses an Ethics by Design approach or who is willing to adhere to the ethical requisites as listed in this guidance. To the degree possible, verify yourself that the system adheres to these requirements. At minimum, the vendor should be able to provide much of the required information. Since Ethics by Design calls for transparency and human oversight, it may be sufficient at first to ask them to explain the developer's ethical oversight mechanisms and show samples of their transparency documentation. If the developer cannot demonstrate these, it is unlikely they will be able to ensure the ethical requisites are being met in the system itself. Without sufficient transparency, it will not be possible to determine the ethical compliance of the system.

- If the AI system is custom-built by an external developer, then give preference to a developer who uses an Ethics by Design approach or who is otherwise willing to adhere to the ethical requisites as listed in Section 2. If possible, verify that the system adheres to these requirements. A simple way to start is to ask the developer to explain the mechanisms by which they operate Ethics by Design, such as documentation and ethical governance procedures. For example, they should be able to show how they document their datasets and models used for machine learning, including how they check for and eliminate bias.
- If in-house development is chosen, then follow the Ethics by Design methods presented earlier in this document, and verify that the resulting system adheres to the ethical requisites listed here.
- Ensure that any data that is collected and prepared for the system prior to deployment adheres to the data collection and preparation guidelines provided in *Section 3.5: Design Phase: Data collection and preparation.*
- An ethical risk assessment and impact assessment should be performed to assess specific ethical risks in the use of the system. Mitigating actions should be planned and carried out to mitigate any ethical risks detected. It may be possible to build this on top of the initial ethical assessment made when the project was first designed, which should have examined these issues.

4.3 Deployment and implementation

- Establish and implement plans and policies which support operational compliance with the ethical requisites for the system.
- Update data, access, security and risk management policies and procedures which apply to the system in order to account for the ethical requisites.
- In training for the operation and use of the system, include the new ethics policies and procedures and pay attention to ethical aspects within communication regarding the launch of the system.
- Monitor the implementation of ethics guidelines for the system throughout the implementation phase, identify issues and risks and make adjustments where needed.

4.4 Monitoring

After launch of the system, continuous or periodical monitoring is required to ensure successful ethical compliance over time:

- Verify that end-users use the system according to user policies which include ethical requisites, are vigilant about ethical issues in operation and use, and consult with senior staff on issues that are morally problematic or ambiguous.

- Ensure that monitoring goals and metrics are in place for compliance with the ethical requisites. Periodically monitor compliance and propose improvements if monitoring shows compliance to be below target.
- Ensure that stakeholders, users and subjects of the system have ethical complaint communication channels by which to alert you to their ethical concerns as they arise. Ensure that these channels are monitored regularly and concerns are processed appropriately by people with appropriate levels of seniority to ensure action if necessary. Ethical concerns should never just vanish into the system, but this requires formal management and reporting processes to avoid. In addition, ethical problems often occur because a system affects people who were never expected to be impacted by the system in the first place. Consequently, you should ensure such communication channels are open in a manner which allows unexpected groups to approach you with their concerns, and that these are handled appropriately.

References and further reading

Abrahams, Alan S, Eloise Coupey, Anuja Rajivadekar, Joshua Miller, Daniel C Snyder, and Samantha J Hayden, "Marketing to the American Entrepreneur", *Journal of Research in Marketing and Entrepreneurship*, 2012.

Aggarwal, Aniya, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha, "Black Box Fairness Testing of Machine Learning Models", *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019*, ACM Press, Tallinn, Estonia, 2019, pp. 625–635. http://dl.acm.org/citation.cfm?doid=3338906.3338937.

Avendaño, Guillermo, Pablo Fuentes, Víctor Castillo, Constanza Garcia, and Natalie Dominguez, "Reliability and Safety of Medical Equipment by Use of Calibration and Certification Instruments", *2010 11th Latin American Test Workshop*, IEEE, 2010, pp. 1–4.

Becker, Helmut, and David J Fritzsche, "Business Ethics: A Cross-Cultural Comparison of Managers' Attitudes", *Journal of Business Ethics*, Vol. 6, No. 4, 1987, pp. 289–295.

Candello, Heloisa, Claudio Pinhanez, and Flavio Figueiredo, "Typefaces and the Perception of Humanness in Natural Language Chatbots", *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, Denver Colorado USA, 2017, pp. 3476–3487. https://dl.acm.org/doi/10.1145/3025453.3025919.

Cardwell, Donald, *The Fontana History of Technology*, Fontana, London; New York, 1994.

Castelo, Noah, Bernd Schmitt, and Miklos Sarvary, "Robot Or Human? How Bodies and Minds Shape Consumer Reactions to Human-Like Robots", *ACR North American Advances*, 2019.

Cavoukian, Ann, *Privacy by Design: The 7 Foundational Principles*, Information and Privacy Commissioner of Ontario, Toronto, 2009.

Chan, Athena, "World's First AI-Powered Bar Uses Facial Recognition To Serve Customers In Proper Order", *Tech Times*, New York, 2019. https://www.techtimes.com/articles/244865/20190803/worlds-first-ai-powered-bar-uses-facial-recognition-to-serve-customers-in-proper-order.htm.

Cole, M., W. G. Lawrence, and N. Leblanc, "Effect of Installation and Maintenance on the Certification of Electrical Equipment", *2019 IEEE Petroleum and Chemical Industry Committee Conference (PCIC)*, 2019, pp. 293–302.

Dafoe, Allan, "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism", *Science, Technology, & Human Values*, Vol. 40, No. 6, November 2015, pp. 1047–1076.

Dalpé, Robert, "Effects of Government Procurement on Industrial Innovation", *Technology in Society*, Vol. 16, No. 1, January 1994, pp. 65–83.

Doran, Derek, Sarah Schulz, and Tarek R Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives", *ArXiv Preprint ArXiv:1710.00794*, 2017.

European Commission, *Coordinated Plan on Artificial Intelligence*, White Paper, European Commission, 2018. https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-795-F1-EN-MAIN-PART-1.PDF.

———, *Single Market Scoreboad*, European Commission, 2020. https://ec.europa.eu/internal_market/scoreboard/performance_per_policy_area/public_procurement/index_en.htm.

Executive Agency for Small and Medium-sized Enterprises, *Annual Report on European SMES 2018/2019*, European Commission, Brussels, 2019.

Friedman, Batya, Peter Kahn, and Alan Borning, "Value Sensitive Design: Theory and Methods", *University of Washington Technical Report*, No. 2–12, 2002.

Gall, H., “Functional Safety IEC 61508 / IEC 61511 the Impact to Certification and the User”, *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 1027–1031.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, “Datasheets for Datasets”, *ArXiv:1803.09010 [Cs]*, March 19, 2020. http://arxiv.org/abs/1803.09010.

Grosz, Barbara J., David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo, “Embedded EthiCS: Integrating Ethics across CS Education”, *Communications of the ACM*, Vol. 62, No. 8, July 24, 2019, pp. 54–61.

Gunning, David, “Explainable Artificial Intelligence (Xai)”, *Defense Advanced Research Projects Agency (DARPA), Nd Web*, Vol. 2, No. 2, 2017.

Harland, Paul, Henk Staats, and Henk A. M. Wilke, “Situational and Personality Factors as Direct or Personal Norm Mediated Predictors of Pro-Environmental Behavior: Questions Derived From Norm-Activation Theory”, *Basic and Applied Social Psychology*, Vol. 29, No. 4, November 5, 2007, pp. 323–334.

High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines For Trustworthy AI*, European Commission, Brussels, 2019.

Hollander, Rachelle, and Carol R. Arenberg, *Ethics Education and Scientific and Engineering Research: What’s Been Learned? What Should Be Done?*, National Academy of Sciences, Washington, DC, 2009. https://www.nap.edu/catalog/12695/ethics-education-and-scientific-and-engineering-research-whats-been-learned.

ISACA, *Auditing Artificial Intelligence*, Information Systems Audit and Control Association, 2018.

Joint Task Force on Computer Engineering Curricula, *Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering*, *Computing Curricula*, Association for Computing Machinery, New York, 2016.

Kission, Polen, Hong Ding, and Ahmed A Jerraya, “Structured Design Methodology for High-Level Design”, *31st Design Automation Conference*, IEEE, 1994, pp. 466–471.

Leveson, Nancy G, “The Use of Safety Cases in Certification and Regulation”, 2011.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, “Model Cards for Model Reporting”, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

Moore, James F, “Predators and Prey: A New Ecology of Competition”, *Harvard Business Review*, Vol. 71, No. 3, 1993, pp. 75–86.

Quigley, Marian, ed., *Encyclopedia of Information Ethics and Security:*, IGI Global, 2007. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59140-987-8.

Rossi, Arianna, and Gabriele Lenzini, “Transparency by Design in Data-Informed Research: A Collection of Information Design Patterns”, *Computer Law & Security Review*, Vol. 37, 2020, p. 105402.

Segun, Samuel T., “From Machine Ethics to Computational Ethics”, *AI & SOCIETY*, June 29, 2020. http://link.springer.com/10.1007/s00146-020-01010-1.

Sims, Randi L, and A Ercan Gegez, “Attitudes towards Business Ethics: A Five Nation Comparative Study”, *Journal of Business Ethics*, Vol. 50, No. 3, 2004, pp. 253–265.

Xie, Xiaoyuan, Joshua W.K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen, “Testing and Validating Machine Learning Classifiers by Metamorphic Testing”, *Journal of Systems and Software*, Vol. 84, No. 4, April 2011, pp. 544–558.

Appendix – Organisational Adoption of Ethics by Design

This section is intended for managers of development organisations, such as software development companies, engineering companies making robots, and development departments within larger enterprises. Ethics by Design requires organisational changes in the way systems are constructed, so this section summarises important organisational changes a manager will need to introduce so that their teams can engage in Ethics by Design. This Appendix may also be of use for senior management and board members who wish to understand the way in which their organisation will need to change.

Organisational Impact of Ethics by Design

Ethics by Design is intended to prevent ethical issues from arising in the first place, rather than trying to fix them after the damage has been done. This is achieved by changing how systems are built. While much of this involves changes in the way a system's functionality is determined, it also requires changes in development processes and tools. Ethics by Design will not speed development, make it easier, or reduce costs. However, the demands that AI systems act ethically will only grow, and many ethical requirements are likely to become legal requirements. The move towards Ethical AI is an unavoidable, and global, trend. Those organisations which can most quickly adapt their structures to society's demands for ethical AI will have a significant advantage over those who resist this trend.

It is important to bear in mind there is no established best practice in this area – every company will be doing this for the first time. It is therefore highly recommended that steps be taken to ensure lessons learned are documented and that procedures are modified as lessons are learned. This may mean that introducing Ethics by Design is an on-going process for some years, and that some disruption is experienced by staff. It is therefore important to maintain active and positive communication between management and developers as their work environment and culture changes around them, offering them clear advantages to supporting these changes.

The most significant changes for a development team will be:

- A large range of additional considerations during the design and development processes.
- New communication channels for ethical concerns.
- New roles.
- Additional reporting requirements.

Governance

Ethical AI requires that humans can oversee the learning, decisions, and operation of AI-driven systems. Not only does this require that developers build mechanisms into the AI product which permit this, it means organisations must put in place teams and/or individuals with oversight responsibilities. Developers could be asked to justify programming decisions to a degree of depth they have never experienced before. This will require documentation of what coding or algorithmic choices they made. It may require documenting who made these decisions and why. Resistance to explaining these matters is likely by many developers. Some will resist because they resent "outsiders" intruding into what has been their private space, while others will simply lack the skills to explain such matters.

The ethical status of a system may change as data changes, as it learns, and as usage changes. This means the organisation needs new processes for on-going monitoring of the system's ethical status, reporting channels for both staff and outsiders to register concerns, formal processes for assessing those concerns, and suitable mechanisms for implementing remedies. Those responsible for these

processes may experience resistance from developers or others because such remedies may interfere with project schedules and will certainly generate additional costs. It is therefore essential those responsible for maintaining a system's ethical status have sufficient authority to enforce their decisions. At minimum, this requires backing from appropriate senior management. Some companies have already started to give board members responsibility for this in order to ensure full compliance within the organisation.

Those who report concerns need to know that they have been taken seriously, considered and what decision was made. This is especially important with external stakeholders. Procedures for communication therefore need to be in place. These may have legal implications, so it is important that legal departments are involved in designing these processes, lest they become a barrier to effective communication. Furthermore, the lessons of ethical failure and remediation need to be retained at an organisational level. This way the organisation can avoid repeating the same errors, and can learn how to implement remedial actions more effectively.

It is likely the position of AI Ethicist will arise. There are already formal certifications for this role. The range of concerns and the skills required are wide-ranging, and so the position justifies formal training and certification. This means it may be possible to hire specialists in these governance areas. It also means it is unreasonable to expect any current staff member to take on this role without suitable training.

External Audit

It is likely many AI systems will be required to undergo auditing by external parties. This is already becoming law in some regimes for some applications. It is therefore important to understand what audit procedures a system may be subject to and ensure suitable procedures and documentation are in place to support audit. Even if audit is not currently required, it is worth preparing for the possibility new regulations will require an audit during the product's lifetime.

If the organisation has appointed a professional AI Ethicist, preparing for, and dealing with, external audit will need to be a core competency.

Culture

Ethics by Design embeds ethics into the design process. The essence of success is to build a culture which treats ethical issues with exactly the same importance as core, undeniable, values in any system, like reliability and bug-fixing. Engineers and computer scientists have traditionally treated ethics as something which happens after they have finished their work; nothing to do with them and not the type of thing which a coder or engineer should be asked to think about. They are highly unlikely to have any training or experience in these matters and so may find even thinking about them difficult or even unpleasant. Managers cannot assume their developers and engineers will rush to embrace ethical concerns, or that they have the skills to handle them. They will need both training and encouragement. Senior managers and board members must bear in mind the same may be true of their subordinate managers, or even themselves. Some degree of education is likely to be required of most staff at all levels. As with other cultural changes in development and engineering organisations, success depends on visible and motivational leadership from senior management.

Finances

Development cycles will almost certainly slow while the organisation learns how best to use Ethics by Design. It is unlikely to return to former levels because Ethics by Design requires additional tasks in

every project. This will affect costs. It is well understood that all developments work within three constraints – time, money and features. Resistance from those staff primarily concerned with finances is likely in many organisations. Those seeking to introduce Ethics by Design into the organisation may therefore need to plan for this possibility and take remedial steps in advance. This is another area where support and leadership from the most senior levels is the best path to success.

Tools

Many of the demands of the developer under Ethics by Design require new tools. It is common when altering development processes that resistance takes the form of insisting the change is impossible due to lack of suitable tools. Where this occurs it should be investigated thoroughly by someone with the suitable technical skills. The requirements of Ethics by Design are common demands of many AI projects, so suitable tools are developing very rapidly. The pace of development in this regard is so rapid it is possible a tool could appear to fulfil a need in the time between identifying the need and the point in the development process where the need must be handled. In particular, DARPA's Explainable AI (XAI) [24] is a rapidly developing set of methodologies and tools by which build effective machine learning techniques which are also explainable to humans and allow for human governance. In these and other cases, such tools are Open Source and freely available to all[25].

Final Notes

We have highlighted here the most obvious organisational changes which adopting Ethics by Design requires. There will be many others. Since AI is a new technology, and ethical AI even newer, no one knows what all the requirements are, nor what is best practice. It is therefore important that moving to Ethics by Design is understood as a significant organisational move, not just a minor change in a few development processes. One cannot expect staff who are experienced in this area, or that anyone can adapt to Ethics by Design without suitable training. In the initial stages compliance may be inconsistent, such that it requires much closer management supervision than once the organisation is used to it.

In many ways the change is similar to that experienced by organisations when web technologies emerged in the mid-1990's. Many technical characteristics of web technology rendered old programming languages obsolete and also introduced demands on IT systems, ways of constructing software, which were completely new. Many senior developers in many organisations, with decades of experience in the old paradigms, found this extremely difficult and many first-generation web applications failed because organisations did not make the required managerial and organisational changes suited to web technologies. Where organisations possess managerial or technical staff who experienced that time, these people should be to possess many lessons which can aid the move to Ethics by Design. And while that move was difficult in the 1990's, it is worth remembering that most organisations got there successfully in the end.

[24] Gunning, "Explainable Artificial Intelligence (Xai)".

[25] A working toolkit for producing Model Cards is available at https://github.com/tensorflow/model-card-toolkit. A template for Datasheets for Datasets is available at https://github.com/AudreyBeard/Datasheets-for-Datasets-Template. A python library of XAI tools is available at https://github.com/EthicalML/xai.

Industry Education and Buy-In for AI Ethics

Annex 3 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Brandt Dainow, *University of Twente* (bd@thinkmetrics.com)
Other contributors	Philip Brey, *University of Twente*

Date	February 2021
Type	Report (Annex 3 to D5.4 deliverable)
Dissemination level	PU = Public

Keywords	Ethical issues; artificial intelligence; AI; robotics; robots; ethics; software design;

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This document outlines proposals for encouraging industry to adopt ethical AI and robotics within design, development, sales, staffing and use. The central mechanisms are a set of certifications for products and people, the development of a certification business eco-system, and the use of proven market mechanisms to build customer demand for certified products and people. This document recommends that AI and robotics products should be certified as ethical upon creation (a "product certification"), and that their deployment in a working environment should also be certified as maintaining that ethical status (an "installation certification"). We further recommend the development of professional certification of staff appropriate to their roles. Recognising the existence of a thriving certification business eco-system, we recommend this certification business eco-system be encouraged to add ethical AI components within their existing professional certifications, such as COBIT. However, our strategy also anticipates, and encourages, the development of new certifying bodies. In pursuit of a self-sustaining business eco-system, we do not advocate prescriptive measures to centralise or unify certification schemes, but allow for the development of a range schemes. Finally, we discuss mechanisms by which to encourage market demand for these schemes, primarily through a product labelling scheme and the creation of awareness of need for it.

Table of Contents

Executive summary

The ultimate aim for industry in our strategy is to use proven techniques to build a self-sustaining business ecosystem which is devoted to the development and promotion of ethical AI systems. We recommend the aim of establishing a business eco-system because they are self-sustaining and also stimulate industry to promote the value of them. Thus industry becomes an active promoter of ethical AI. However, for this to work, the market must demand ethical AI products, so we discuss mechanisms for developing market demand.

The central plank of the strategy is a certification program based on similar schemes, including electrical equipment certification, industry staff certification programs (such as Microsoft's Certified Professional), the EU Energy Labelling scheme, and the established processes by which commercial industry develops certification programs, such as ISACA's COBIT. We recommend integration into existing programs rather than promoting new ones, though we allow for the rise of new certificates as well.

We recommend system certifications based on the model seen in industries where safety is important, such as electrical products, aircraft and medical equipment. Systems must first obtain an *ethical AI product certification* before they can enter the market. Increasingly, AI products sit on back-end platforms which provide the raw AI processing power. We therefore recommend an *AI platform certification* which focuses on the platform's suitability to provide its backend functions ethically. Installed systems will also need *AI installation certificates* before they can be used. Installation certifications will include assessing the wider context within which the application is operating as well as the application itself, such as management processes. A system's ethical status may change over time as it learns and acquires new data so auditing is required at regular intervals to ensure ongoing ethical status and maintain the installation certificate.

We recommend people are certified through professional training programs and exams. A range of certifications will be required, as is the case for most technical systems, according to one's role. Ethical AI requirements should be incorporated into the EU e-Competence framework because most European certification bodies use this as a standard from which to draw their requirements.

Professional associations should be encouraged to include requirements for appropriate CPD training in ethical AI. Professional associations often have codes of conduct which should be kept in line with the certification standards as they evolve, as well as relevant audit or CPD requirements. Many European professional associations draw guidance from the European e-Competence Framework, so this is another reason for updating it to include ethical AI.

Because 99.8% of all businesses in the EU are small and medium-sized enterprises (SME's), the majority of AI purchasers will be small businesses. We therefore recommend building ethical AI awareness through trade associations and business support networks. This awareness strategy must promote the value of the certification schemes. Purchase of an AI possessing an ethical AI product certification should be presented as resolving many difficulties which would otherwise fall to the SME.

The strategy for motivating industry is to make it profitable to produce ethical AI products because the market demands them. This is to be achieved by building an awareness of the need for ethical AI certification, together with confidence that certified products and staff are trustworthy. Awareness of need is to be reinforced by encouraging the insurance industry to give preferential rates to certified

products and staff and through the use of procurement procedures, such that only certified products may be submitted for government tenders.

The final, but most important, step in developing market demand is a labelling scheme based on the Energy Rating Labels, so that the ethical status of an AI system or robot can be assessed instantly and without advanced technical knowledge.

To co-ordinate all these efforts (certifications, education, labelling, and industry liaison), we recommend the development of a new central AI unit. Such a unit has been recommended by the EU Coordinated Plan on Artificial Intelligence and by the European Parliament. While we allow for a suite of certificates by different bodies, we recommend a Central Ethical AI Reference model against which certifications can be compared. We recommend this model be housed within this central AI unit.

List of figures

List of tables

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
CEARM	Central Ethical Artificial Intelligence Reference Model
CGEIT	Certified in Governance of Enterprise Information Technology
CEPIS	Council of European Professional Informatics Societies
COBIT	Control Objectives for Information and Related Technology
CPD	Continuing Professional Development
EU	European Union
GDPR	General Data Protection Regulation
IBM	International Business Machines
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISACA	Information Systems Audit and Control Association
ISO	International Organization for Standardization
IT	Information Technology
SME	Small and Medium-sized Enterprise

Table 1: List of acronyms/abbreviations

1. Introduction & Overview

This section recommends policies and strategies by which to develop a commercial industry which is committed to, and executes, the development of ethical AI and robotics systems, and which uses them in an ethical manner.[1]

We distinguish between commitment and execution because an organisation may be committed to a policy yet fail to embody that commitment in action. This report therefore offers distinct strategies by which to achieve both. Furthermore, it is possible for ethically-sound AI or robotics systems to be used in an unethical manner and for AI and robotics systems to change their ethical status as they learn or absorb new data. Consequently, this report offers strategies for ensuring the on-going ethical status of AI and robotics systems during operation. The central aim is the development of a business ecosystem[2] focused on the maintenance and expansion of ethical AI and robotics systems.

We do not attempt to offer distinct solutions, such as specific implementations or detailed policy proposals. Given the range of relevant actors already extant within commercial industry who need to be involved in this process, we consider it more effective to recommend the development of a business ecosystem, or set of self-sustaining markets, which are dedicated to the delivery of the necessary services which will achieve our aims. We do not consider it effective to expect such a range of actors will subscribe to a single solution, or even a single approach. Our aim is rather to recommend approaches which will permit these actors to implement our aims in accord with their existing operations and goals. For this reason we also recommend a "light touch" to regulation, introducing regulations and directives only to the degree necessary. Our strategy is focused on encouraging industry to adopt our recommendations by making it in their commercial interest to do so.

It is clear that industry need training[3] in ethical AI. However, we believe that training will only be sought by commercial enterprises if there is a profit to be made from it. It is therefore necessary that training provide a company with something they would not otherwise have, and that they believe this benefit leads to increased sales or reduced costs. Similarly, we have identified many features AI and robotics systems must possess if they are to be considered ethically safe. However, it is not obvious many of these will automatically increase sales or reduce costs. In many cases they will certainly increase cost of manufacture and cost of operation. We cannot therefore expect industry to automatically adopt them. Instead, we must find a way to increase the commercial value of AI and robotics products which do meet our ethical requirements, and increase the value to a degree which outweighs these costs.

We do not believe legislation alone is a sufficient strategy. Industry cannot be forced to adopt ethical AI. Firstly, the international nature of many AI producers will simply allow them to move development

[1] The authors of this report acknowledge the input of various experts and stakeholders to this text. Please see the Acknowledgement section of SIENNA D5.4 (Feb 2021) for a list of these people.

[2] Under the concept of a *business ecosystem*, "a company be viewed not as a member of a single industry but as part of a business ecosystem that crosses a variety of industries. In a business ecosystem, companies coevolve … they work cooperatively and competitively to support new products, satisfy customer needs, and eventually incorporate the next round of innovations" (Moore, 1993, p.76)

[3] We distinguish in this report between 'education' and 'training'. We use the term 'training' for education delivered commercially in industry and which does not lead towards academic degrees, while we use the term 'education' for university-delivered courses generating academic credits. This is the common usage within the private educational business sector. Thus a person delivering courses in private industry is referred to as a 'trainer', not a 'teacher' or 'lecturer' and a private educational business is referred to as a 'training company.'

to regimes which do not operate strict regulations. Secondly, no amount of legislation can convert a loss-making product into a profitable one. If regulation makes certain forms of product unprofitable, companies will simply not make them.

Our strategy is therefore focused on making it worthwhile for companies to create AI or robotics systems which meet our ethical requirements by converting these requirements from prescriptive constraints into competitive advantages. If being an ethically sound AI product increases market opportunities, while not being ethically sound harms sales, then businesses will organically adopt our ethical requirements by choice.

The centre of the strategy is a certification program. All the protocols and procedures outlined here relate to, or draw from, this certification program. This strategy does not offer radically new or untried approaches. It combines elements from a number of relevant areas which have proven successful, including certification of equipment (such as electrical and aircraft components), industry training and certification programs (such as Microsoft's Certified Professional and Singapore's Certified AI Engineer programs), the EU Energy Labelling scheme, and the established processes by which commercial industry develops certification programs, such as ISACA's COBIT and CGEIT[4]. Our strategy actively avoids suggesting the development of new initiatives, instead favouring the integration of our aims into existing programs. Just as our strategy with vendors is motivate them to adopt ethical AI certification voluntarily, our strategy with certification bodies is to build an environment which motivates them to develop and promote the required training and certification programs themselves. The ultimate aim is to see a self-sustaining business ecosystem which devoted to the development and promotion of ethical AI systems because companies profit from it.

2. The Ethical AI Certification Program

There will be three forms of certification:

- **Systems** – AI and robotics systems will be certified as meeting ethical requirements.
- **People** – People can obtain a range of Ethical AI certifications, as appropriate for their role (e.g.: developer, business manager, educator)
- **Training Programs** – Training programs leading to certification will themselves be certified as suitable for the task. Under most existing certification schemes, the training company must be certified as able to effectively deliver the training program.

These certification programs are intended for use in commercial industry, but success primarily depends on the support of ancillary activities in regulation, university education and public awareness. The aim of these ancillary activities is to create a commercial advantage for those who take on board the certification programs, primarily by generating demand amongst purchasers for certified products and staff. This commercial advantage will then generate a desire to seek certification. Since adoption of the certification system produces commercial advantage, vendors will also be motivated to maximise the value of their investment in these certifications by promoting their possession of them

[4] COBIT stands for "Control Objectives for Information and Related Technology." CGEIT stands for "Certified in the Governance of Enterprise IT". Both are training and certification frameworks created by the ISACA (Information Systems Audit and Control Association) for IT governance and management and have wide industry support.

and the value of them. The net result will be the development of a self-sustaining marketplace (a business ecosystem) focused on ethical AI and robotics products.

Certification Ecosystems

Once sufficient demand for certification arises, we anticipate the spontaneous development of a product certification ecosystem, as we have seen with initiatives such as GDPR and with the development of many technologies, such as fire alarms and Microsoft Windows. This will mean many different types of stakeholders becoming involved in the product certification process. It is to be expected that industry trade bodies will take an interest in certification of their member's products and seek to offer industry-specific certification programs. For example, the International Organization of Motor Vehicle Manufacturers[5] is active in many areas of vehicular regulation and is likely to take an interest in certification of self-driving cars. It is also likely new trade associations will arise united by specific AI functions used in multiple industries, such as a trade association for creators of facial recognition systems. We expect this because such cross-industry associations have arisen in the past. For example, the Digital Analytics Association[6] was created in 2003 to set standards and certification programs for both people and products which measure web activity, such as online advertising systems. Other less specific organisations, which are involved with the development of many standards, will become involved in certification programs. For example, IEEE is developing the Ethics Certification Program for Autonomous and Intelligent Systems[7]. We can also expect the rise of organisations dedicated to the certification process itself, similar to organisations such as the many national associations of auditors in the accountancy field. Interested charities and other civic associations are likely to develop or promote their own certification schemes. For example, the Institute for Ethical AI & Machine Learning[8] is a voluntary body which has developed frameworks for ethical assessment and procurement of AI systems. Converting these into a certification program would be a natural, and relatively straightforward, extension of their current work. Similarly, the charity ForHumanity[9] is developing audit processes and can be expected to take an interest in ensuring certification criteria match their audit criteria.

Significant providers, such as Google, IBM, Microsoft and Amazon, will develop their own organised ecosystems of adherents (developers, vendors and the final end-user organisations), just as we have seen with other technology platforms in the past, such as the Microsoft Partner Network[10] and IBM Partnerworld[11]. These companies have developed their own product certification programs previously, such as the "Certified for Windows Server" badge[12] scheme, which demonstrates that an application meets Microsoft's technical standards for performance on the Windows platform. They also create and operate their own professional certification schemes. For example, the Microsoft Certified Professional program offers over 30 discrete certifications[13] for technicians and managers of Microsoft IT systems. In many cases, training programs and trainers supporting certification programs are themselves certified under such schemes. For example, the Microsoft Certified Trainer program certifies trainers

[5] http://www.oica.net/
[6] https://www.digitalanalyticsassociation.org/
[7] https://standards.ieee.org/industry-connections/ecpais.html
[8] https://ethical.institute/
[9] https://www.forhumanity.center/
[10] https://partner.microsoft.com/en-US/
[11] https://www.ibm.com/partnerworld/public
[12] https://www.windowsservercatalog.com/
[13] https://docs.microsoft.com/en-us/learn/certifications/

as fit to teach courses preparing people for the Certified Professional exams[14]. This is paired with the Microsoft Learning Partner certification[15], which certifies training companies as suitable to teach the Certified Professional courses. Another part of the Learning Partner certification scheme certifies companies to run the certification exams. We can therefore expect technology providers such as Microsoft and IBM to develop their own range of certification schemes for their own products. In addition to the sales advantages a certification scheme offers such companies, they understand that once a partner company, such as a developer, invests in their certification schemes, it becomes more difficult for them to move to a competitor (known colloquially in the industry as "lock-in"). Company-focused certification schemes are therefore seen by large technology companies as offering both immediate and long-term benefits.

Finally, we can expect companies which specialise in private training and certification initiatives to develop their own certification programs. This process has already started and can be expected to accelerate. For example, Certnexus[16] is a private company which specialises in developing certification programs related to emerging technologies. These are developed under the guidelines laid down in ISO/IEC 17024:2012, which specifies the requirements for the development and operation of certification schemes which certify people, as opposed to products[17]. Certnexus has developed the Certified Ethical Emerging Technologist certification, together with a training program to help people prepare for the certification exams[18]. Other companies are developing training programs which do not offer formal certification, but to which certification could be appended. For example, some members of the EU's High Level Expert Group on AI have formed a commercial ethical AI consultancy, ALLAI, which includes training programs in ethical AI[19].

Our recommended approach is that this range of bodies be encouraged to develop certification schemes for their members and customers. This may lead to a single AI product being certified under a number of different schemes. Should such a practice arise, it may, in time, become appropriate to develop systems for ensuring certification schemes do not clash. For example an EU directive may be advisable for laying down minimum standards for such schemes, requiring registration or licencing of AI certifiers, or creating a central register of AI product certifications which potential purchasers or users can check. A directive may be required to specify core elements of certifications in order to ensure comparability of competing products certified under different schemes. There is also the need to avoid a "race to the bottom", in which certification schemes compete for users by lightening their requirements. Similarly, there may be a need to build a regulatory system for the "interlocking" of different certification schemes. For example, a self-driving car certification scheme may automatically accept image-processing components previously certified under an image-recognition certification scheme. There may also be the need to develop directives or similar policy initiatives which ensure nationally-based schemes are recognised across the EU or internationally. For example, Denmark is developing a Data Security and Ethics labelling scheme for AI and robotics systems[20]. We can expect

[14] https://docs.microsoft.com/en-us/learn/certifications/mct-certification
[15] https://docs.microsoft.com/en-us/learn/certifications/partners#become-a-microsoft-learning-partner
[16] https://certnexus.com/about-certnexus/
[17] https://www.iso.org/standard/52993.html
[18] https://certnexus.com/certification/ceet/
[19] https://allai.nl/allai-programs/#toggle-id-10
[20] https://eng.em.dk/news/2019/oktober/new-seal-for-it-security-and-responsible-data-use-is-in-its-way/

other countries to do the same. It is important that all such national schemes are compatible recognise each other, so that products do not need independent certification in every EU state.

We do not expect a single audit and certification program to evolve unless regulation is put in place to force this (which we do not recommend). Consequently, different certification and audit protocols are inevitable. Because industry certification and audit are commercial activities, it is likely there will be competition between them. As a result, we can expect changes in certification and audit programs over a number of years. If it is considered appropriate to regulate certification programs, we recommend the frequently-used legal approach in the case of new technologies - allow customary practice to emerge over a 10 - 20-year period and then to regulate in accordance with that.

2.1. The Ethical AI Certification Eco-system

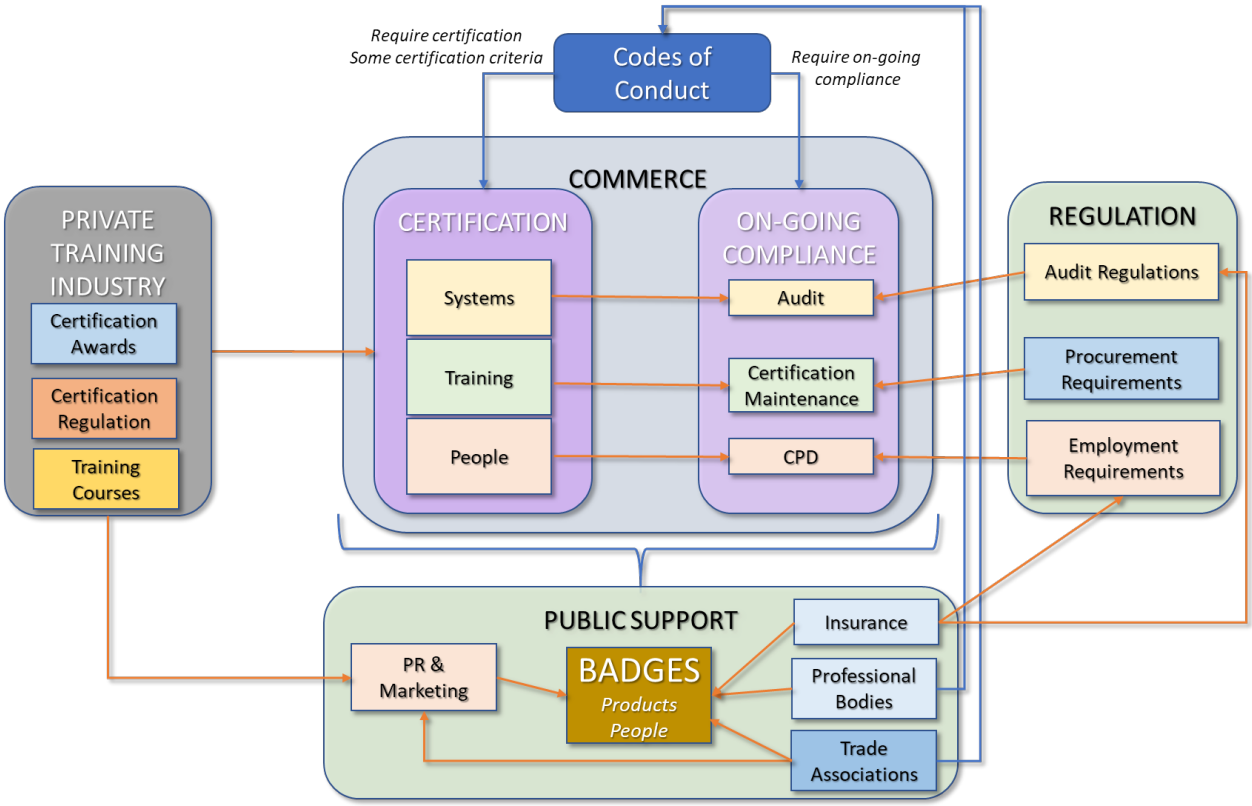

Figure 2: The Certification Ecosystem

The basis of our proposal is the development of a business ecosystem devoted to certification of AI products and people. As detailed below, we rely on the private training and certification industry to provide the certificates. People, systems and training are certified. For each there exists a corresponding process to ensure certification remains up-to-date. Government activity, such as regulations for audit and procurement requirements, enforce the need for certification. Public demand is focused on badges and other physical identifiers of the ethical status of a product. This demand is stimulated by marketing strategies and promotion by concerned sectors, such as professional member's associations.

2.2. Types of Product Certification

The model we recommend for AI and robotics product certification is based on the model seen in many industries where it is important to ensure safety, such as electrical devices. Here devices must be certified as safe before they can be sold on the market. Then the installation of those devices in a building must be certified before the building can be used. In the case of important systems, such as fire alarms, systems must be tested and certified regularly to ensure on-going compliance (Cole, Lawrence, and Leblanc, 2019). Similar models are used in many industries, such as aircraft (Leveson, 2011), medical equipment (Avendaño et al., 2010) and electronics (Gall, 2008).

Consequently, three types of certification are required for robotics and AI systems; certification of products, certification of their installation and certification of back-end AI platforms used to provide the intelligence to multiple products.

Product Certification

Systems must first obtain an *Ethical AI Product Certification* in order to enter the market. Product certification attests that the system meets the ethical requirements pertaining to a completed, but undeployed, system. This has two concerns; firstly, that the system's operation meets ethical requirements, and secondly, that the system is capable of future auditing to determine if it remains ethically compliant once in operation.

Installation Certification

Since deployment may change the ethical status of a system (e.g.: though use of different datasets) or involve elements of the system which cannot be evaluated prior to deployment, systems will need an *AI Installation Certificate* after deployment, but before operational use. For example, a facial recognition system may be deployed so that a building security system can recognise the faces of employees and open security doors for them. This system will first need an Ethical AI Product Certificate before it can be sold to anyone. Certification that it is ethically fit for sale will assess aspects such as whether the system is equally good at recognising different genders and skin tones. Once deployed in a building and trained to know who the staff are, it will then need an *AI Installation Certificate* to ensure bias as not been introduced while learning the staff faces. Organisational elements will also need to be included in an AI Installation Certificate. For example, the need for transparency requires that people know when they are being subjected to treatment by an AI. In this example, there would be a requirement that appropriate signage be displayed to offer this awareness. This would be assessed as part of the AI Installation Certification process. We can expect a similar range of stakeholders to emerge around installation certifications as with product certification, with similar concerns and calling for similar remedies.

Platform Certification

Certification of platform-based, as opposed to self-contained, AI systems is more complicated because the developer of a system which uses AI-platform capabilities cannot certify those capabilities, only their own use of them. However, these back-end platforms cannot be ignored. Where they provide learning models, pre-existing data sets and similar capabilities, their ethical status can be expected to affect, if not completely determine, the ethical status of the application calling on them. However, it is impractical to expect a platform to be investigated afresh every time a vendor wants to create a product using it. We therefore recommend that platforms be subject to certification as AI platforms. *Ethical AI Platform Certificates* will focus on certifying the system's suitability to provide its backend functions ethically. Under such a scheme a product may not be certified as ethical if it is using a back-end AI platform which does not possess an Ethical AI Platform Certification.

Multiple Product Certification Schemes

It is unlikely a standardised set of assessment criteria can be developed which is applicable to any and all AI or robotics systems. The wide variety of uses and the range of possible functions are too great, and many are yet to be invented. Some criteria can be derived from the field of application, such as medicine or self-driving vehicles. Other criteria of assessment can be derived from the functionality of the system, such as facial recognition systems or expert systems based on document analysis (such as those which underpin legal expert systems and chatbots). It is to be expected that some projects will

be so innovative that they will not easily fit into existing assessment criteria. Here some negotiation will be required between the project and the assessors. Certification schemes will need a central set of criteria used across all products in order to maintain consistency of assessment and comparability of products. However, they will also need some processes which offer the flexibility required to assess innovative or bespoke features. As indicated above, certification eco-systems for a variety of solutions in this regard are viable without compromising the aims of certification.

Ethics by Design and Certification Schemes

To some degree the certification criteria for completed, yet undeployed, systems will offer a set of formal ethical design requirements. This will allow developers to know which ethical requirements the final system must meet before designing the system. Methodologies such as Ethics by Design provide frameworks by which to achieve these requirements organically within the design process. Indeed, some of the assessment criteria for certification may require specific processes or tools during the development process, such as documenting data sources and how they were manipulated. Here Ethics by Design (or a similar methodology) will be necessary in order to integrate such requirements into the normal development process.

2.3. Auditing

A system's ethical status may change over time. For example, "on the job" learning by a working AI system may drive changes in system behaviour, as may changes in data. On-going compliance must therefore be regularly assessed through auditing. A successful audit will maintain the system's certification status. Certification programs may operate on the basis that certification only lasts for a period of time, such as one or two years. In such cases, auditing is required to obtain a new certification. Alternatively, certification may be permanent, unless the system fails audit, in which case the certification is withdrawn. The difference between these two options is not insignificant. Withdrawal of certification is an active process, whereas failure to issue a new certificate is more a lack of action. As such, a system which withdraws an otherwise permanent certification is likely to bring expectations that such withdrawal can be appealed. Thus it is likely that certification programs which withdraw otherwise permanent certifications will require a more complex and extensive organisational structure.

If a system fails an audit, a certification program has two options; call for remedial action to maintain certification or withdraw certification. Certification programs may choose to adopt a graduated system by which to rate audit failure, such that some failures merely indicate a small need for adjustment within a reasonable timeframe. In such a case certification is not fully withdrawn immediately and time is given for the system to be brought into compliance. Other failures may call for immediate loss of certification while the system is taken offline for repair.

It may be appropriate to have multiple levels of certification for a single system, especially if that system offers a range of functions or use contexts. Similarly, auditing of some systems may be most effective by offering discrete audit criteria for different functions or usages. In our earlier example, we cited a scheme by which self-driving vehicle certification accepted the certification of the image-recognition subsystem. If the vehicle changes its driving patterns as a result of experience, its driving behaviour will need regular auditing, similar to an annual vehicle fitness check. However, if it does not modify (or "teach") the image-recognition subsystem, that sub-system will not need to be audited

because its functionality will not change over time. It may be appropriate that some product or installation certificates specify the appropriate frequency of audits.

We do not recommend specific approaches in this regard, nor with regard to forms of audit process, strategies for handling audit failure or other details of the certification and audit processes. Similar certification and audit systems in other industries, such as aircraft components, operate effectively while allowing for considerable national variations in approach (Leveson, 2011). As a result we do not feel variations in certification or audit necessarily compromise the development of ethical AI and robotics systems.

2.4. People

People are to be certified through professional training programs and exams. AI Engineer certification schemes are already available, such as Singapore's Certified AI Engineer program[21] and Google's Cloud Professional Machine Learning Engineer certification[22]. While the Singapore program is not focused exclusively on the ethics of AI, ethics are included and subject to a formal exam. On the other hand, Google's certification scheme does not, as yet, include any ethical component. Purely ethically-focused certification programs are also arising, such as Certified Ethical Emerging Technologist[23] from Certnexus. There are many existing certifications which should adopt ethical AI components, such as COBIT.[24]

A range of professional certifications will be required, as is the case for most technical systems. For example, while developers need a detailed understanding of the coding decisions which can lead to ethical issues, the senior managers of an organisation using that system do not. Instead they need to understand how their organisation's way of using an AI system can affect its operational ethical status. Most professional certification programs are already organised in this way. For example, the COBIT certification program distinguishes thirty-four different roles, such as Board Member and Chief Information Officer, and sets distinct responsibilities for each. For example, Board Members have responsibility for monitoring the governance of technical systems, while the Chief Information Officer is responsible for monitoring the data quality assessment processes.

An ecosystem for professional training in ethical AI is already developing. In addition to formal certification programs like Certnexus's Certified Ethical Emerging Technologist, some organisations have developed their own ethical AI training programs. For example, the Linux Foundation, which has trained 1.7 million Linux engineers[25], offers an examinable training course "Ethics in AI and Big Data"[26] within its suite of AI/Machine Learning courses.

We do not consider our ancillary activities detailed below to be essential to the development of professional certifications because such certifications are developing organically already. However, we cannot be certain they will reach sufficient demand to change the overall direction of AI or robotics

[21] https://www.aisingapore.org/ai-certification/
[22] https://cloud.google.com/certification/machine-learning-engineer
[23] https://certnexus.com/certification/ceet/
[24] https://www.isaca.org/credentialing/cobit
[25] https://www.linuxfoundation.org/
[26] https://training.linuxfoundation.org/training/ethics-in-ai-and-big-data-lfs112/

innovation. Therefore the ancillary activities detailed below are designed to increase the value of such certifications and thus increase demand for them.

While Certnexus demonstrates that specifically ethically-focused certifications will develop, for the most part we expect that current certification programs will add ethical AI elements to existing training programs, just as they have done with GDPR and data protection requirements. This process will be facilitated if ethical AI requirements become incorporated into important guidance programs, especially the EU e-Competence framework.[27] This framework has an existing mechanism for incorporating new competencies. For example, moving from Version 2 to Version 3 of the e-Competence framework added Innovating, System Engineering, Needs Identification and Digital Marketing. Consultations with member organisations have already identified the need to add competencies relating to big data, machine learning and other aspects of AI. Furthermore, there are signs that ISACA's COBIT certification program will come to formally include ethical AI within some of its certifications. ISACA has published a white paper on auditing AI for ethical status, with particular focus on installation certification (ISACA, 2018). If ISACA follows the same processes as it did with cloud computing, we can expect to see formal components for ethical AI and robotics systems to be added to COBIT certifications. Other professional certification schemes operate in a similar fashion. Since this has been a consistent pattern as new technologies have arisen in the past, we can expect ethical AI to be incorporated in the same way in the future.

Professional Associations - CPD

Professional associations typically recognise that skills can erode over time and that working environments change as technology develops. They therefore expect, or demand, their members maintain their competencies through continuing professional development programs (CPD's). In many cases governments require these associations to oblige their members to undertake suitable CPD in order to give their profession appropriate regulatory backing. In other cases, insurance companies require CPD for matters such as professional liability insurance. There is therefore a pre-existing ecosystem for professional CPD training into which ethical AI can be incorporated.

CPD increases the range of personnel accessible for training in ethical AI and robotics systems. The majority of professionals do not obtain COBIT or similar certifications, which require considerable commitments of time and money. However, CPD training is often in the form of short courses (1 - 2 days) or self-study and provides an opportunity to educate many professionals who do not need, or will not take, more formal professional certification programs.

Professional Associations – Codes of Conduct

Professional associations often have codes of conduct. Where relevant they should be encouraged to incorporate ethical AI concerns. Codes of conduct for relevant professional associations should be kept in line with the certification standards as they evolve, as well as relevant audit or CPD requirements. This will be relatively straight-forward with professional associations directly involved in developing AI or robotics systems, such as national associations of IT professionals or engineers. A valuable channel for communication with IT professionals will be The Council of European Professional Informatics Societies (CEPIS)[28] which represents the thirty-five national IT professional associations across Europe.

[27] https://www.ecompetences.eu/
[28] https://cepis.org/

However, other professions may need to take on board concerns of a form not encountered by them before because AI has not historically formed part of their toolset. For example, the increasing use of expert systems in law may require provision in codes of conduct for members of the legal profession, as we have seen with data protection. Many of the European professional associations are actively involved in, or draw guidance from, the European e-Competence Framework[29], so this constitutes a major channel for encouraging the addition of ethical AI components into relevant professional codes of conduct.

SME's and Business Awareness

It is important to bear in mind the majority of AI purchasers will be small businesses. AI systems are not necessarily large or expensive. For example, bars are deploying facial recognition systems to determine whether someone is too young to be served, bill people automatically and monitor staff performance (Chan, 2019). Such systems cost as little as €200/month and so are easily affordable by most businesses. 99.8% of all businesses in the EU are small and medium-sized enterprises (SME's) (Executive Agency for Small and Medium-sized Enterprises, 2019). They account for 66% of all employment in the EU. Consequently, SME's must inevitably constitute the majority of purchasers and end-users of AI systems. As such they are an essential audience for increased awareness of ethical AI issues. However, professional certification schemes such as COBIT are not designed for SME's and are too expensive for most. We therefore recommend building the necessary ethical awareness through the channels which SME's already use to acquire new expertise - trade associations and business support networks.

These organisations offer short talks at networking events, such as business breakfasts, and also offer consultancy and training services, which can also constitute CPD training in some professions. Issues relating to ethical AI should be introduced into these channels. Such material should focus on general citizen awareness as subjects of AI decisions, training in ethical use of such systems in small and medium-sized enterprises, and issues relating to making ethically informed purchases. As with private training companies, there are a wide range of such organisations. The most important in this regard are those which can be leveraged to provide the most impact at a single point of contact. We identify two EU organisations as critical in this respect; the Enterprise Europe Network (EEN)[30] and the European Business and Innovation Centre Network (EBN)[31]. The EEN is the world's largest support network for small and medium-sized enterprises. It is active in more than 60 countries worldwide and brings more than 600 member organisations, reaching down to local enterprise boards. The EBN is dedicated to providing support for business support organisations, such as incubation hubs, and those receiving their services. It is operational in 29 countries and has 175 members who support over 25,000 companies. Both these organisations provide the capability to deliver a range of educational services, from short awareness briefings to more advanced training in ethical AI and robotics systems.

Other organisations should be identified which can offer comparable channels but which are not run by with government organisations or funded under EU initiatives. For example, most countries have some form of SME association, such as the Irish Small Firms Association[32], The Royal Association MKB-

[29] Some examples can be seen at https://www.ecompetences.eu/professional-bodies-trade-unions-and-sector-associations/
[30] https://een.ec.europa.eu/about/about
[31] https://ebn.lt/
[32] https://www.sfa.ie/

Nederland[33] and the German Association for Small and Medium-sized Businesses[34]. Some of these can be approached via the EU's Executive Agency for SMEs[35] or comparable EU institutions. Others will need to be approached directly. As we have seen with GDPR, we can expect the eventual development of an ecosystem of individuals and training companies specialising in delivering this material to such organisations.

SME education and awareness building should not focus on the obtaining of certification, but on general awareness of issues relevant to the purchase and operation of AI and robotics systems. Of critical importance is that this awareness strategy promotes the value of the certification schemes. Purchase of an AI possessing an Ethical AI Product Certificate should be presented as resolving many difficulties which would otherwise fall to the SME. Similarly, hiring suitably certified AI engineers and installers should be presented as a safer choice than hiring those without certification. As SME's develop awareness of the potential dangers inherent in purchasing and operating AI and robotics systems, they should come to see ethical AIR certification schemes as mitigating risk as much as possible.

2.5. Training programs

Where formal certification exists, standard industry practice is that training programs leading to certification exams must themselves be certified by the body awarding the certificate. This occurs through approval of a training program by approval of a governing body. Most, but not all, certifying bodies have some form of government backing, such as approval under an existing EU or national program or the accumulation of academic study credits. Ongoing compliance with a training programme is assured by the awarding body through ongoing maintenance and review. Where certificates already exist, we can expect such mechanisms to also exist. Accordingly, we do not need to develop ongoing compliance programs for training courses, but can rely on the certifying bodies to do this as part of their normal operating procedure. Because these are private training programs, they have been shown to be driven by student satisfaction with the quality of the training. Should the contents of a training course become obsolete due to changes in industry or regulation, we can rely on the clients to pressure the certifying body to keep the course up to date. Furthermore, if we work with a number of certifying bodies, commercial competition will further motivate them to maintain their standards in this regard.

Preparing people for formal certifications will not be the only form commercial training courses will take. As with existing schemes and industry patterns, the majority of training programs will not focus on certifications like COBIT. Instead they will be oriented to practical skills of immediate use to companies. The majority will be 1-day or 2-day courses because these can be delivered on weekends, or do not require too much time out of the office. More advanced training, especially for technicians, typically occupies a complete 5-day week, rarely longer. All such courses have a business imperative of selling themselves to commercial companies. Natural competitive processes keep these courses up-to-date as part of their appeal to potential training customers. Therefore, if demand exists for such training, no other effort is required to specifically encourage maintenance and improvement in private training material.

[33] https://www.mkb.nl/
[34] https://www.bvmw.de/
[35] https://ec.europa.eu/easme/

Similar procedures and pressures exist for the certification of trainers and for encouraging trainers to maintain their skills, and so can be relied upon to ensure trainers remain competent.

3. Commercial industry motivation

This section offers the broad outlines of a strategy by which to generate the desire for ethical certification of those who create or sell AI or robotics systems. The objective is to create a marketplace which desires certified products and staff. The rationale driving this strategy is that that suppliers will adopt certification in response to market demand. A secondary, but no less important, objective is to ensure that certification schemes are compatible and sufficient.

We make no assessment of the degree to which commercial enterprises will be motivated by ethical concerns because we do not consider it safe to base our recommendations on such judgements. There is evidence that the degree to which commercial actors feel obligated by ethical requirements varies between cultures (Becker and Fritzsche, 1987; Sims and Gegez, 2004) and individuals within those cultures. These attitudes range from "anything goes" to mild constraint. The most consistently held position across all business cultures is that the primary ethical imperative on a business is to generate a profit for the shareholders. Thus, what is universal across all commercial enterprises is the need to make a profit, whether this is merely seen as a fact of life or an ethical imperative. Our approach is therefore based on the premise all commercial enterprises are driven by the profit motive, whether other motives are present or not. Under this approach we do not seek to encourage compliance with ethical AI requirements by anything other than a desire to maintain or increase profitability. We assume industry will show an interest in ethical certification and training programs if they believe these will increase sales. However, sales will only be affected by ethical certification if potential customers are aware of the existence of such certification and actively desire products or staff which have been suitably certified in preference to those which have not.

The success of our proposals therefore depends on several strands. Firstly, building suitable beliefs and values in the marketplace, primarily amongst purchasers of AI and robotics products, but amongst the general public as well. Secondly, we recommend strategies which encourage existing training and certification companies to develop training courses and incorporate ethical AI components into their certification schemes. Thirdly, there is a need to provide a central model (or standard) of concerns against which certification schemes and training programs can be assessed. This central model will also enable guidance of the important features which should be addressed. In this sense, the central model has the capability to act as the focal point of civic debate regarding the practical requirements made of ethical AI and robotics systems.

The primary aim is to plant two key viewpoints in the market:

- An awareness of the need for ethical AI certification, accompanied by an awareness that such certification exists, and that it answers the perceived need.
- Confidence that certified products, services and staff are easily available, that selecting a certified product or person is no more difficult than selecting an uncertified one, and that certified products are just as good as uncertified ones, if not better.

While market demand can generate a desire for certification, other elements have the potential to enhance this desire further – pressure from insurers, procurement requirements and some regulation

or central organisation. Our strategy for developing industry interest in ethical certification therefore includes these elements as well.

The Central Ethical AI Reference Model (CEARM)

While we recommend allowing multiple independent paths to certification, a central ethical reference model (CEARM) is required in order to provide a common set of ethical criteria by which to measure certification programs. Such a model requires a central organisation to curate it. We believe our proposals are in line with those outlined in the EU Coordinated Plan on Artificial Intelligence (European Commission, 2018). This calls for co-ordination of efforts towards ethical AI in many dimensions. We also note recommendations the European Parliament resolution of 16 February 2017 to the Commission on Civil Law Rules on Robotics[36], which recommends the designation of a central European agency for robotics and artificial intelligence which will provide technical, ethical and regulatory expertise. Such a body would be an appropriate curator of a central model of ethical requirements which all certification programs must cover.

Such a model provides a number of benefits. Firstly, it ensures a minimum standard which certification must achieve in order to be valid. It may be determined in the future that certifications require licencing. A central agency could create and co-ordinate certification licencing if such a decision is taken, while the central model becomes the standard against which licencing is determined. A central model also ensures different national certification schemes, such as Denmark's Responsible Data Use labelling scheme, maintain minimum standards, respect the certification schemes of other member states and are compatible across the EU. A central reference model also forms a basis for international negotiations to ensure compatibility with global standards, such as those of the IEEE, and with less prescriptive ones, such as those of UNESCO.

The CEARM contains the core ethical values we want to promote through the various awareness and certification schemes outlined above. This centres on the six ethical values and the resultant ethical requisites covered in our guidance documents on AI Ethics by Design. The exact format of the CEARM can be determined later, but it would most likely take the form of recommended guidelines for certifications. This may contain a new canonical standard based on our recommendations, but we also recommend that potential certificates have the option to reference themselves against an existing standard from a recognised standards body, such as IEEE. Under this scheme, a candidate certificate would need to explain which standard (or combination thereof) it used as its basis for requirements, and how it complies with it. For example, a certificate may reference some of the IEEE's portfolio of Artificial Intelligence Systems Standards.[37] Under this approach, the proposed process for issuing an AI Product or Platform Certificate could be compared to IEEE P2840 (Standard for Responsible AI Licencing)[38], while an AI Installation Certificate could derive assessment criteria from IEEE P2863 (Recommended Practice for Organizational Governance of Artificial Intelligence)[39]. Neither standard would be sufficient alone for a certification and most certifications would need to reference a number of IEEE standards. The IEEE currently has thirty-six standards with the Artificial Intelligence Systems Standards portfolio; some are limited to specific contexts while others are very broad and applicable to most AI and robotics systems. For example, IEEE P7014 (Ethical Considerations in Emulated Empathy

[36] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051
[37] https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html
[38] https://standards.ieee.org/project/2840.html
[39] https://standards.ieee.org/project/2863.html

in Autonomous and Intelligent Systems)[40] is only relevant to systems which emulate emotion, while IEEE P7001 (Standards for Transparency of Autonomous Systems)[41] is applicable to all AI and robotics systems. ISO has a similar approach, with twenty-six relevant standards in development under the ISO/IEC JTC 1/SC 42 Artificial Intelligence Work Plan[42]. Some of these, such as ISO/IEC CD 38507 (Governance Implications of the Use of Artificial Intelligence by Organizations),[43] may be suitable as reference baselines for certificates.

Alternatively, candidates to become certificates in Ethical AI could be compared against the CEARM itself. It is likely most certification schemes would need to use the CEARM to fill gaps left by other suites of standards. Compliance need not be absolute. Similar to the Energy Rating Certificate, it would be possible to assign a set of compliance ratings, and to do so individually for each of the six values. There are six values in our recommendations - human agency, data governance, fairness, personal and social well-being, accountability and oversight, and transparency. A product or installation certificate could therefore rate each individually. For example, an AAA-AAA rating could indicate 100% compliance with all values to the utmost degree, while AAA-ADF could indicate 100% compliance with most values, but the last two letters indicate mediocre oversight mechanisms (D) and poor transparency (F). Not all values will be equally important in all situations. For example, the transparency value holds that people should know when they are being the subject of AI decisions. However, it is less important to know this when the AI is directing the traffic lights than when it is deciding whether to grant you a loan. Individual ratings for each value can be paired with variable rating requirements, which require different minimum levels for different usages. Thus, an AI certificate tells vendors and purchasers which type of situations the product may be used in. In the example provided above, such a rating could be considered sufficient for traffic management functions, but not for operations affecting individual finances. Nuanced ratings also enable developers to target particular forms of ethical status in a drive to enter particular markets and makes possessing ethical status of a particular value a positive sales feature.

We allow for the possibility the exact contents of the CEARM could require some civic debate. There are many groups who have already taken positions regarding what constitutes acceptable ethical AI. Furthermore requirements will change as AI evolves and social awareness grows. It is therefore inevitable some will have opinions about what should, or should not, be universally required. It is also inevitable new innovations will create situations which existing ethical requirements do not take into account. We therefore anticipate there will be an ongoing civic debate regarding what constitutes an ethical AI product and that this debate will focus on the CEARM. Given the nature, range and importance of ISACA certifications (especially COBIT), ISACA should also be directly involved in the development of the CEARM. ISACA support will be essential in building ethical AI into critical enterprise personnel certifications. There may be other significant organisations who should be directly involved in such a forum, ranging from charities to AI and robotics manufacturers.

While the CEARM is based on the findings of this project's research, it need not be confined to them in order to achieve the objective of developing a robust ethical AI business ecosystem. There are already a number of significant alternative sets of values and requirements extant, and it is inevitable

[40] https://standards.ieee.org/project/7014.html
[41] https://standards.ieee.org/project/7001.html
[42] https://www.iso.org/committee/6794475.html
[43] https://www.iso.org/standard/56641.html

other organisations will take positions on ethical AI over time. It would therefore be appropriate to have a mechanism for updating the CEARM periodically. This requires some form of ongoing organisation and associated review processes. We do not offer specific recommendations in this regard, merely reiterate that existing proposals and initiatives for a central AI agency provide a suitable environment for such processes. While we do not believe it is the only viable solution in this regard, we will use this below as an example of how such functions could be performed, while recognising existing organisations may be able to offer comparable capabilities.

CEARM - Summary

The CEARM functions as a reference for all initiatives to promote ethical AIRs Candidate certificates can be compared to its requirements. These may themselves receive nuanced ratings. For example, a certificate intended for a company director may require high ratings for governance activities, but limited knowledge of data bias amelioration techniques.

Major activities around the CEARM will be:

- A standard against which to assess potential certificates.
- A central civic forum through which evolving needs and requirements can be debated and determined.
- A source of reference and source information for training materials and public awareness initiatives.

3.1. Awareness of Need

Awareness of need has been shown to be an effective component in changing social norms and behaviour (Harland, Staats, and Wilke, 2007) and the processes for generating it are well understood (Abrahams et al., 2012). Awareness of the issues regarding the ethical aspects of AI drives the desire to seek remedies. Some awareness is building organically in the community already and can be expected to continue. Proven techniques for building awareness of need can enhance the speed, depth and insight of public understanding of ethical AI and robotics systems and should be deployed. The exact policies and programs which are most appropriate should be designed by those combining an understanding of these issues with expertise in the appropriate marketing and communication strategies.

As with the range of programs and audiences discussed above (*see 2.4: People, p.88*), the modes and content of communication will need to vary according to the audience and the concerns relevant to them. This requires the development of a central store of concepts and concerns, together with material ranging in detail, from basic infographics to white papers and similar in-depth material suitable to the general public, AI purchasers and operators, and for developers and vendors of AI and robotics systems. A central agency would allow for co-ordination across the EU, together with production of common elements, such as logos, source material for graphics, statistics and possibly as a source of brochures, training material and other documents.

The most important audience for the generation of an ethical AI ecosystem is those who will purchase AI and robotics systems. As we have seen, this will be mainly SME's. Consequently, the most important channels for awareness building are those which inform SME's, such as business support networks. Here there is a need to deliver the full range of business information services, from short awareness briefings to more advanced training in purchasing and managing AI and robotics systems ethically. The most important channels are the Enterprise Europe Network (EEN) and the European Business and

Innovation Centre Network (EBN). Many small business associations can be reached through the EU's Executive Agency for SMEs or comparable EU institutions, while others will need to be approached directly.

The European e-Competence Framework must be updated to include ethical AI. Other frameworks for digital competence may also need updating. Since the Council of European Professional Informatics Societies (CEPIS) draws much of its codes of conduct from the e-Competence Framework, CEPIS should also be directly involved in the development of a set of formal specifications which all codes of conduct should include.

3.2. Insurance pressure

Insurance has the potential to be an important driver of market sentiment. At some stage legal action is likely against the operator of an AI or robotics system regarding some ethical aspect of its behaviour, such as racial or gender bias. These matters therefore alter the liability of AI operators. Adoption of an AI which has been formally certified for ethical compliance significantly lowers the risk for an insurer and should therefore affect insurance premiums. The insurance industry will inevitably adapt to the rise of AI and robotics systems, but may do so in an uncoordinated fashion and may look for requirements which are not coherent with other ethical AI expectations. We therefore recommend the inclusion of the insurance industry in the development of the CEARM. We also recommend insurance companies be considered a prime channel for development of market need. This can be accomplished by the existing channels through which such concerns are discussed between the EU and insurance industry. In particular, we recommend direct involvement by Insurance Europe[44] in setting certification standards and training requirements.

3.3. Procurement

Public procurement is an effective method of influencing commercial product innovation (Dalpé, 1994). Public procurement accounts for 14% of the EU's total GDP (European Commission, 2020). This gives the government sector extremely powerful influence over vendors of systems. We recommend that the EU move towards requiring Product and Installation Certificates for all AI and robotics systems as part of its procurement requirements, and that it imposes such a requirement on subsidiary organisations, down to the level of local government bodies, to the degree that it is able. Under this policy, products may not be included in tenders unless they possess ethical AI product certifications and may not be activated once purchased until they have achieved an AI Installation Certificate. They would also be required to support appropriate auditing once deployed. This policy could also demand professional certifications, as appropriate, for personnel involved in operating government AI's. For example, all staff who can make purchase decisions for AI's could be required to obtain a certificate which qualifies them to understand the ethical issues of AI's and make informed assessments of possible purchases. The degree to which such policies can be enforced on other branches of government, especially member states, is largely a matter of political negotiation. Here we note that such policies need not be prescriptive but could simply be "highly recommended."

At the time of writing, certifications for ethical AI products are not developed to the degree that it is possible to insist on certified products. However, it is possible to announce such a requirement will be

[44] https://www.insuranceeurope.eu/. Its members are the EU's national insurance associations. Insurance Europe members represent 95% of total European insurance activity.

activated at some time in the future, possibly in 3-5 years. This will stimulate the growth of certification schemes and motivate developers to consider ethical requirements, even if they cannot actually adopt certification schemes just yet.

3.4. Certificate Badge Branding

The final element required to create a market which demands ethical AI products is widespread public awareness of certificates and what they do. Here our model is based on the EU's Energy Labelling Framework and its associated energy labels, such as the Electrical Product Labelling Scheme. We propose a similar scheme. Under this system, approved certifications would generate a standardised ethical AI label. Using similar mechanisms as have been used with the various energy labels, purchasers and users of systems can learn to read such labels. This would enable them to quickly assess the ethical status of a product at a glance. The Danish government is already developing such a badge[45]. We believe it is important to establish initiatives at the EU level as soon as possible so as to prevent the rise of competing or incompatible national schemes.

The Energy Labelling Framework has been very successful and there is good evidence it is now an active consideration when people make purchases. We do not propose to offer alternative communication, marketing or educational strategies. We believe much of what has been done with the Energy Labelling Framework could be emulated here. The framework has been running for long enough to know what works, and so we suggest simply copying that.

[45] https://investindk.com/insights/denmark-paves-the-way-for-implementation-of-trust-by-design

Glossary of terms

Term	Explanation
AI Platform	A back-end system which offers AI capabilities which other developers can use to build AI applications
AI Platform provider	A company offering AI platforms to developers, such as Clairifai, IBM and Google.
Auditability	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enabling the system's auditability.
Bias	Bias is an unfair or unjustified prejudice towards or against a person, group of people, object, or position. Bias can arise in many ways in AI systems. It does not necessarily relate to human bias or human-driven data collection. It can arise, for example, through the limited contexts in which a system in used, in which case there is no opportunity to generalise it to other contexts. Bias can be intentional or unintentional, but is a danger because it frequently causes discriminatory and/or unfair outcomes in AI systems
Business Ecosystem	A network of organizations (including suppliers, distributors, customers, and competitors) involved in the delivery of a specific product type or service through both competition and cooperation.
Ethics	Ethics is an academic discipline which is a subfield of philosophy. Applied ethics deals with real-life situations, where decisions have to be made under time pressure, and often limited rationality. AI Ethics is generally viewed as an example of applied ethics and focuses on the issues raised by the design, development, implementation and use of AI.
Ethical AI	Ethical AI refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values.
Ethics by Design	The approach of incorporating ethical considerations throughout the design, development and deployment phases of software and engineering product creation so as to avoid the product generating negative ethical effects.
SME	Small and Medium-sized Enterprise. A small enterprise has fewer than 50 employees and an annual turnover not exceeding €10m. A medium-sized enterprise has 50 - 249 employees and an annual turnover not exceeding €50m.

Table 2: Glossary of terms

References

Abrahams, A.S., E. Coupey, A. Rajivadekar, J. Miller, D.C. Snyder, and S.J. Hayden, 'Marketing to the American Entrepreneur', *Journal of Research in Marketing and Entrepreneurship*, 2012.

Avendaño, G., P. Fuentes, V. Castillo, C. Garcia, and N. Dominguez, 'Reliability and Safety of Medical Equipment by Use of Calibration and Certification Instruments', *2010 11th Latin American Test Workshop*, IEEE, 2010, pp. 1–4.

Becker, H., and D.J. Fritzsche, 'Business Ethics: A Cross-Cultural Comparison of Managers' Attitudes', *Journal of Business Ethics*, Vol. 6, No. 4, 1987, pp. 289–295.

Chan, A., 'World's First AI-Powered Bar Uses Facial Recognition To Serve Customers In Proper Order', *Tech Times*, New York, 2019.

Cole, M., W.G. Lawrence, and N. Leblanc, 'Effect of Installation and Maintenance on the Certification of Electrical Equipment', *2019 IEEE Petroleum and Chemical Industry Committee Conference (PCIC)*, 2019, pp. 293–302.

Dalpé, R., 'Effects of Government Procurement on Industrial Innovation', *Technology in Society*, Vol. 16, No. 1, January 1994, pp. 65–83.

European Commission, *Coordinated Plan on Artificial Intelligence*, White Paper, European Commission, 2018.

———, *Single Market Scoreboad*, European Commission, 2020.

Executive Agency for Small and Medium-sized Enterprises, *Annual Report on European SMES 2018/2019*, European Commission, Brussels, 2019.

Gall, H., 'Functional Safety IEC 61508 / IEC 61511 the Impact to Certification and the User', *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 1027–1031.

Harland, P., H. Staats, and H.A.M. Wilke, 'Situational and Personality Factors as Direct or Personal Norm Mediated Predictors of Pro-Environmental Behavior: Questions Derived From Norm-Activation Theory', *Basic and Applied Social Psychology*, Vol. 29, No. 4, November 5, 2007, pp. 323–334.

ISACA, *Auditing Artificial Intelligence*, Information Systems Audit and Control Association, 2018.

Leveson, N.G., 'The Use of Safety Cases in Certification and Regulation', 2011.

Moore, J.F., 'Predators and Prey: A New Ecology of Competition', *Harvard Business Review*, Vol. 71, No. 3, 1993, pp. 75–86.

Sims, R.L., and A.E. Gegez, 'Attitudes towards Business Ethics: A Five Nation Comparative Study', *Journal of Business Ethics*, Vol. 50, No. 3, 2004, pp. 253–265.

sienna.

Annex 4 – Research Ethics Guidelines for Artificial Intelligence

Research Ethics Guidelines for Artificial Intelligence

Annex 4 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Philip Brey, *University of Twente* (p.a.e.brey@utwente.nl)
Reviewers	Lisa Tambornino, EUREC Office GUG
	Declan Brady, CEPIS & the CEPIS ethics group

Due date	February 2021
Type	Report (Annex 4 to D5.4 deliverable)
Dissemination level	PU = Public

Keywords	Research ethics guidelines; guidelines; ethics; artificial intelligence; AI; robotics; robots; big data

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Table of contents

Abstract

In this report, we propose a set of research ethics guidelines for artificial intelligence (AI) and discuss their possible implementation. Our proposal includes twenty-seven research ethics guidelines that were grouped into six categories, under the headings of human agency, privacy & data governance, fairness, social and environmental well-being, accountability & oversight, and transparency, as well as "special topics" guidelines for specific techniques, products, and application domains in the AI field. We also propose using an Ethics by Design approach, which provides a comprehensive way of integrating ethical guidelines and criteria into design methodologies. After presenting these proposals, we propose how they can be used as a basis for stand-alone research ethics frameworks for AI (including robotics and big data), and next how they can be integrated into broader research ethics frameworks for computer and information sciences, and then for research ethics frameworks that span multiple fields. As far as we can see, this report contains the first comprehensive proposal for research ethics guidelines for AI, including their integration into broader research ethics frameworks. We encourage universities, companies, research funding organisations, and other organisations involved in the ethical assessment of research to utilize this report as a manual for compiling their own research ethics frameworks for AI.

1. Introduction

This report proposes how researchers, developers, and research ethics committees (RECs) and other research assessors can incorporate ethics guidelines for artificial intelligence (AI) in their research and in their research ethics frameworks and guidance documents.[1] We define AI broadly, to also include large parts of robotics and data analytics. We make proposals for three situations: (1) the inclusion of AI guidelines in research ethics frameworks that span multiple disciplines; (2) the inclusion of AI guidelines in research ethics frameworks for computer science; (3) the development of stand-alone research ethics frameworks for AI. These need to be distinguished because guidance for AI research will need to be adapted to the research ethics frameworks, if any, that are already in place for the disciplines that are covered by them.[2]

In recent years, various national and international organisations have proposed ethics guidelines for AI. These include, most prominently:

- The *Ethics Guidelines for Trustworthy AI* of the High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Union.[3]
- The *Recommendation of the Council on Artificial Intelligence* of the OECD.[4]
- *Ethically Aligned Design*, guidelines included in a publication of the Institute of Electrical and Electronics Engineers (IEEE) on the ethical development of intelligent and autonomous systems.[5]

It should be emphasized that generally, these guidelines for AI are different from the kind of ethics guidelines that are included in research ethics. Most ethics guidelines for AI are *general ethics guidelines*, for different uses in different contexts by different actors. Guidelines in research ethics are directed at one type of practice: research, and one type of actor: researchers. General ethics guidelines are directed more at prescribing desirable outcomes for society than guiding specific practices, such as those found in research. They therefore transcend particular actors, practices, or contexts, and pertain to all of them, including developers, deployers, end-users, assessors, funders, regulators and others. They specify general ethical principles, define desirable outcomes for society as a whole, and prescribe rules of conduct that all relevant actors should follow.

The OECD recommendation, for example contains guidelines like “AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle,” defining AI actors as “those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI”, as well as “actorless” guidelines like “AI systems should be robust, secure and safe throughout their entire lifecycle ...” (p. 7-8). The AI HLEG “requirements for trustworthy AI” are

[1] The authors of this report acknowledge the input of various experts and stakeholders to this text. Please see the Acknowledgement section of SIENNA D5.4 (Feb 2021) for a list of these people.

[2] The term “research ethics framework” is intended to refer to the approach, aims and methods taken by research actors to ethically assess and guide research. This includes chosen research ethics guidelines, but also methods for utilizing an implementing them in research ethics assessment and guidance.

[3] High-Level Expert Group on Artificial Intelligence (AI HLEG), *Ethics Guidelines for Trustworthy AI*, April 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

[4] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, May 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[5] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition, IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

defined as requirements that AI systems must meet throughout their lifecycle and include requirements like "Societal and environmental wellbeing" and "Diversity, non-discrimination and fairness". It is stated in the text that different groups of stakeholders have different roles in ensuring that these requirements are met, including developers, deployers, end-users, and the broader society.

Next to general ethics guidelines, there are also ethics guidelines that are directed at specific actors and/or practices. The OECD guidelines, for example, contain a section which proposes specific ethics guidelines for governments that are based on their general ethics guidelines for AI. In the EU-funded SHERPA project, a report was issued that proposed ethics guidelines for deployers of AI systems.[6] And in the SIENNA project, we developed a multi-actor strategy for ethical AI that proposed roles and responsibilities for multiple actors, including developers, deployers, educators, policy makers, standards organisations, and others.[7]

In this context, *research ethics guidelines* for AI can be understood as ethics guidelines for a specific practice, i.e., research in, and development of, AI systems. It is important to understand that they are guidelines for a *practice*, which transcends the individual *actor*. In research ethics, the desired outcome is that research is conducted in an ethical manner. Of course, this imposes an obligation on the actors that are involved in that practice (the research team) to contribute to it being carried out in an ethical manner. But this is a different requirement than requiring the actors themselves to behave ethically in general. They are only to behave ethically in as far as it contributes to the research being conducted in an ethical manner.

To address ethical conduct by individual actors, there exist actor-specific guidelines for researchers. *Professional ethics codes and guidelines* guide the behaviour of individual actors in various professional fields, including research and innovation.[8] They aim to regulate professional conduct so as to ensure it exhibits high ethical standards, professional quality, and trustworthiness. Codes of professional ethics are in place not only in professions that centre around research and innovation, but in many other fields as well (e.g., for lawyers, nurses, and journalists). In professions like computer science and engineering, in which research and innovation have an important place, professional ethics codes to some extent cover expected professional behaviours in relation to research and innovation, but much of what they cover is more general. Professionals in these fields carry out many tasks other than research and development of new technology, such as managing people, interacting with clients, teaching, writing a column for a newspaper, and sitting on a review committee in their company. A large part of professional ethics is typically devoted to general virtues and professional behaviours that define professional integrity, social responsibility, and professionalism in these fields.[9]

[6] Brey, Philip, Björn Lundgren, Kevin Macnish, and Mark Ryan, "Guidelines for the Ethical Use of AI and Big Data Systems", SHERPA project, July 2019.

[7] Brey, Philip, Philip Jansen, Jonne Maas, Björn Lundgren, and Anaïs Resseguier, "An Ethical framework for the development and use of AI and robotics technologies", Deliverable D4.7 of the SIENNA project, 2020.

[8] Martin, Clancy, Wayne Vaught, and Robert C. Solomon (eds.), *Ethics across the Professions: a Reader for Professional Ethics*, Oxford University Press, New York, 2010.

[9] Ščepanović, R., Labib, K., Buljan, I. et al. Practices for Research Integrity Promotion in Research Performing Organisations and Research Funding Organisations: A Scoping Review. *Sci Eng Ethics* 27, 4 (2021).

Research ethics guidelines, in contrast, typically do not apply to individual conduct but to research and innovation practices.[10] These practices often involve multiple researchers, and it is not their individual conduct that the guidelines are directed at, but the overall way in which the research is conducted. These guidelines are typically not only used by researchers themselves, but also by research ethics committees that assess research. Research ethics committees typically do not do this during or after the research activity, but prior to it, on the basis of a research plan or proposal. Whereas the research ethics frameworks are general and provide rules and regulations to make sure studies are conducted in an ethical manner, a research proposal outlines what steps will be taken in a specific study. Research ethics committees then assess whether the research proposal adheres to relevant ethical standards or guidelines. Researchers usually make use of a *self-assessment tool* which contains a template for carrying out an *ethics self-assessment* that is then submitted to a REC, which proceeds to do its own assessment that draws from information in the self-assessment. Research ethics guidelines may either be separately incorporated into the self-assessment tool for researchers and a separate *internal research ethics framework* for RECs, or there may be a *shared research ethics framework* that guides both the researchers and the assessors.

Fig. 1 Types of ethics guidelines for AI

The current situation, as of early 2021, is that most research ethics frameworks that are in use across the globe at research performing and funding organisations and by RECs contain little or no guidelines that pertain specifically to AI. In fact, very few proposals for research ethics guidelines for AI, if any, have been published, despite the great amount of attention given to AI ethics and general ethics

[10] Ipfhofen, Ron (ed.), *Handbook of Research Ethics and Scientific Integrity*, Springer International Publishing, 2020.

guidelines for AI in recent years. Even more so, there are very few published proposals for research ethics guidelines for the field of computer science, the research field that encompasses artificial intelligence. Codes of professional ethics have existed in computer science for a long time, and professional organisations like the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM) have championed such codes, but guidance for research ethics has not received the same amount of attention. This situation is likely to change in the near future, as more and more universities and research institutions are facilitating or requiring ethics review for computer science research generally, and more and more tech companies (e.g., Apple, Microsoft, Facebook) are also instituting research ethics committees.

So even though AI ethics has been a topic that has garnered a lot of attention and scholarship in recent years, published research ethics frameworks for AI virtually do not exist at this point, and we are doing pioneering work in this report. We will proceed as follows. In section 2, we will discuss whether and how general ethics guidelines that have been proposed for AI can constitute a basis for research ethics guidelines for AI, and we will propose core research ethics guidelines for the development of AI technology. In section 3, we will propose stand-alone ethics guidelines for R&D in AI. In section 4, we will propose how ethics guidelines for AI can be incorporated into broader ethics guidelines for computer science and information technology. In section 5, we will propose how ethics guidelines for AI can be incorporated into broad research ethics guidelines that span multiple disciplines. This will necessitate that we also develop and introduce ethics guidelines for the deployment and use of AI technology in research. Finally, in the conclusion we will discuss some further implementation issues for the guidelines and summarize our findings.

The proposed research ethics guidelines in this document are based on the ethics guidelines for AI proposed by EU High-Level Expert Group on AI, that have been endorsed widely in the European Union and beyond, as well as on widely accepted research ethics guidelines that pertain to all research fields. Our specific proposals have not yet been used and tested in actual ethics assessment procedures, but they have been subject to peer review and user review, and are based on previous proposals from the EU-funded SIENNA and SHERPA project that have been peer-reviewed and user-reviewed as well, and have also been open for public commentary.

2. From general ethics guidelines to research ethics guidelines

General ethics guidelines for a new technology normally provide a good foundation for the development of research ethics guidelines, because the practices and the actors involved in R&D will be within the scope of the general guidelines, even if they are not referred to explicitly or exclusively. For example, if a guideline states that AI systems should provide benefits to human beings, then it can reasonably be inferred that efforts should be made in R&D to develop AI systems that provide benefits to human beings. General ethics guidance does not normally provide enough specific, actionable guidance for R&D, however, because of their general nature. They fail to be specific either because they only specify desirable outcomes for society, or because they specify general types of actions that all actors should perform, without being specific about the actions to be performed by researchers. Moreover, general ethics guidelines for AI do not incorporate some ethical principles and guidelines that have already been established for research ethics more generally, such as principles of informed consent, research integrity, protection of research subjects, and animal welfare. These are principles

that have been developed to relate to research practices specifically. They would also have to be accounted for in a research ethics framework for AI.

This suggests a dual strategy for the development of research ethics guidelines for AI. Their development needs to be (1) based on general ethics guidelines for AI, which should be translated into actionable, operational guidelines for R&D, and (2) based, in addition, on general guidelines and principles already in place in research ethics, which the AI-specific guidelines should incorporate and reinforce where possible.

In what follows, we will build on the *Ethics Guidelines for Trustworthy AI* of the EU High-Level Expert Group on AI. While this implies a choice for one set of guidelines over others, it has been argued by myself and others that regarding their underlying principles, these guidelines are not substantially different from many other sets of guidelines that have been proposed by other organisations, notably the OECD and IEEE guidelines.[11] Therefore, it should be relatively easy to substitute these other guidelines for the Trustworthy AI guidelines, if one so chooses.

The AI HLEG proposes seven general guidelines, or "requirements" as they are called:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

Following our earlier report for the SHERPA project, an EU project on ethics of AI and big data, we propose to reformulate the AI HLEG requirements slightly. First, we propose to not include the requirement of technical robustness and safety in our research ethics guidelines, since these issues are normally already assessed in regular scientific assessment of research. For instance, in the new Horizon Europe research funding programme of the European Union, they are included in the regular evaluation process, and not in their ethics review procedure. There is also good reason not to include them in research ethics procedures, since ethics assessors may not have the scientific competency to assess technical robustness and safety. Second, instead of the requirement of human agency and oversight, we employ a requirement of human agency, and instead of the requirement of accountability, we have a requirement of accountability and oversight. Our reason for this is that oversight is in our assessment associated with accountability, and less so with the other notions referred to by the AI HLEG in their description of the agency and oversight requirement, which are the notions of agency, autonomy and human rights.

In our proposal for research ethics guidelines for AI, we will provide, for each of the six resulting requirements, a further elaboration of their meaning. As a next step towards research ethics, we then use the six requirements to propose more specific requirements for R&D in the field of AI. We call these requirements *ethical requisites*. The ethical requisites are high-level ethics guidelines for R&D in

[11] Ryan, Mark, Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne van der Puil, "Ethical Tensions and Social Impacts of Smart Information Systems", SHERPA project, 2019.

AI. They correspond to the top-level general ethics guidelines in AI provided by the AI HLEG. They are an attempt to define, at a high level, which actions should be performed and which states-of-affairs should be realized in R&D in order for R&D practices to make a proper contribution to the realization of more general states-of-affairs and actions prescribed by the general ethics guidelines.

To be able to propose the right set of high-level research ethics guidelines, we need an *empirical* understanding of the roles that R&D is able to have in contributing to the fulfilment of the conditions described by the general guidelines and a *normative* understanding of the moral responsibility vested in R&D for contributing to this fulfilment. Regarding the empirical understanding, a key issue is how much the activity of researchers and developers matters in bringing about conditions of privacy, autonomy, well-being, and the like. If one were to believe that technology is neutral, and its social consequences depend on deployers and users, rather than on developers, then one could conclude that R&D has little role in bringing them about, and efforts should rather focus on providing deployers and end-users with adequate ethics guidance rather than investing in research ethics. If one, at the other hand, were to believe in technological determinism, implying that new technological products necessitate certain social consequences, one might be more inclined to emphasize ethics guidelines for R&D and be less concerned with deployment and use.

The position taken here lies in the middle: researchers, developers, deployers, end-users and other actors in society contribute to the fulfilment of these requirements. For each, we need to clarify their proper role, both in terms of capabilities and professional, moral and legal, responsibilities, and formulate ethics guidelines that are commensurate to each of these roles. For most of the six top-level requirements, the choices made in R&D have significant consequences. True, choices in deployment and use influence the implications of AI systems for privacy, agency and other values, but AI systems also have a lot of autonomous capabilities that shape their context of use and help determine whether they are transparent, support agency, responsibility and fairness, and promote privacy and well-being.

The normative understanding that is needed of the role of R&D in the fulfilment of general guidelines depends on how we assign responsibility to the institution of R&D with respect to their fulfilment. Having established that researchers are able to significantly affect their fulfilment, should we also conclude that they have a responsibility to do so? Existing frameworks for both research ethics and professional ethics for researchers and developers do tend to emphasize a certain degree of moral and social responsibility for the consequences of R&D for society. It is reasonable to demand that R&D includes efforts to anticipate and mitigate for ethical issues that may result from choices in R&D. Of course, researchers and developers cannot anticipate and be held responsible for difficult to foresee applications and uses of their work, nor for social implications that are not immediately obvious. But there are many ethical issues in AI, such as many privacy issues, algorithmic bias, lack of transparency and lack of agency for users, that can be foreseen and mitigated very well by developers, and it is then reasonable to include responsibility for these issues in the research ethics framework for AI. Increasingly, moreover, design approaches are emerging that enable designers to have better foresight

and control over otherwise elusive consequences of their design such as their implications for well-being[12] and implications for the social good.[13]

For each of the six general guidelines, corresponding research ethics guidelines will now be proposed (ethical requisites).[14] It is recommended that these ethical requisites are communicated by research ethics committees to researchers as a checklist that they can use to identify, assess and mitigate ethical issues prior to and/or during the inception of research. It is moreover recommended that researchers document their compliance to them and explain whether or not the requisites raise relevant concerns for their intended research. If there are relevant concerns, then actions for mitigation should also be documented. It is moreover recommended that these guidelines are used by RECs in research ethics assessment.

The ethical requisites that we propose are based in part on the self-assessment questions that we developed for the European Commission's Ethics Review procedures for AI, and in part on the ethics guidelines proposed in the Ethics by Design framework for AI for the SIENNA project, included as an annex in this report. They have similarities to, and are partially inspired by, the checklist items contained in the *Assessment List for Trustworthy Artificial Intelligence* (ALTAI) of the High-Level Expert Group on AI.[15] That list is aimed at both AI developers and deployers, and is therefore broader in scope as for the actors and practices it covers by containing guidelines and prescribed actions for deployers. Our requisites pertain specifically to R&D.

2.1. Human agency

Human agency is defined broadly in this proposal to encapsulate the values of autonomy, freedom and dignity. These are the fundamental rights upon which the EU is founded and that are enshrined in the UN Declaration of Human Rights. Autonomy means the ability of people to decide for themselves what is right and wrong and the way they should live their life as a consequence. They should not be coerced, deceived or manipulated. Dignity means that every human being possesses an intrinsic worth which should never be compromised by others, including AI. This means they have the right not to be treated as "a means to an end". Freedom means that that people can exercise their autonomy by making their own decisions, are free to act without restraints imposed by others, including having freedom of speech and information and freedom of assembly.

The value of human agency implies a norm that AI technology should be developed so as to support, and not harm, the autonomy, freedom and dignity of end-users and other stakeholders. End-users, first of all, should be given agency, input and control. They should not be constrained by a system that

[12] Brey, Philip, "Design for the Value of Human Well-Being", in Jeroen van den Hoven, Pieter E. Vermaas, Ibo van de Poel (eds.), *Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains*, Springer, 2015, pp. 365-382.

[13] Brey, Philip, "The strategic role of technology in a good society", *Technology in Society,* Vol. 52, Feb 2018, pp. 39-45.

[14] These proposed guidelines are based in part on the self-assessment questions that we developed for the European Commission's Ethics Review procedures for AI, as part of our SIENNA activities, and in part on the ethics guidelines proposed in the Ethics by Design framework for AI and robotics for the SIENNA project, included as an annex in this report. The latter guidelines are, moreover, based in part on the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment of the High-Level Expert Group on AI (AI HLEG, 2020).

[15] AI HLEG, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*, 2020 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

thinks and decides for them, or that seduces, deceives and manipulates them, or by robotic systems that physically constrain them. Non-users should also not have their autonomy, freedom and dignity be constrained by AI. For non-users, this happens in a more indirect manner that often involves actions taken by others. In particular, AI systems sometimes process personal data of individuals that can be used by others in ways that are harmful to them or produce other types of data that is harmful to individuals, and AI systems can also act in ways that have direct or indirect negative consequences for individuals.

Ethical requisites

- AI systems should be designed to give system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system.
- It should be ensured, as much as possible, that systems that are being developed do not autonomously make decisions about vital issues that are normally decided by humans as the result of free personal choices or collective deliberations, e.g., issues affecting life, health, well-being or rights of persons, or economic, social and political decisions.
- It should be ensured, as much as possible, that end-users and others affected by the system are not deprived of abilities to make basic decisions about their own lives, have basic freedoms taken away from them, are subordinated, coerced, deceived, manipulated, objectified or dehumanized, or that attachment or addiction to the system and its operations is being stimulated. This should not happen directly, through direct operations and actions of the system, and it also should be prevented, to the extent possible, from happening indirectly, due to the system being designed to enable and support its use by others for these purposes.

2.2. Privacy & Data Governance

Privacy is the right of a person to be free from intrusion into or publicity concerning matters of a personal nature. This includes privacy with respect to one's body, one's thought, and personal space, as well as informational privacy: the right to control the processing and dissemination of one's personal information. As a value, data governance means humans must actively manage their personal data and the way the system uses it. This includes the accuracy of data, access to data, and other data rights such as ownership. Ethical issues can arise from both non-personal data (e.g. racial bias) and personal data (where the data subject's rights and freedoms must be safeguarded).

Ethical requisites

- The processing of personal data requires careful consideration of the rights and freedoms of the data subjects. These should be safeguarded at all times. For more information and guidance please see the EU's *Guidance Note on Ethics and Data Protection*.[16]
- Applications must explain how the proposed system supports the right of an individual to withdraw consent for the use of their personal data, and how they will be able to object to its use.

[16] European Commission, *Ethics and data protection*, 14.11.2018. https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf

- In the case that personal data is processed by the developed AI systems, you must demonstrate how you will ensure lawfulness, fairness and transparency of the data processing.
- Technical and organisational measures must be in place to safeguard the rights of data subjects through measures such as anonymization, pseudonymisation, encryption, and aggregation.
- Strong security measures to prevent data breaches and leakages must be in place and described in your application (such as mechanisms for logging data access and data modification).
- Data should be acquired, stored and used in a manner which can be audited by humans.
- All EU funded research must comply with relevant legislation and the highest ethics standards. This means that EU beneficiaries must apply GDPR principles.

2.3 Fairness and non-discrimination

'Fairness' is used here in a philosophical sense, not to be confused with mathematical fairness or use of the term within computational modelling. Fairness in this context has three possible meanings, depending on the context; sameness, deservedness, and compliance. Sameness means that each person is treated the same. Deservedness means ensuring an equitable distribution so that each get what they deserve. Fairness as compliance means operating in compliance with relevant rules. Fairness requires all people have the right to be treated appropriately and not on the basis of irrelevant characteristics. In particular, people should not be treated unfairly on the basis of aspects of their identity which are inalienable and cannot be taken away from them. The most important of these are gender, race, age, sexual orientation, national origin, religion, health and disability.

Ethical requisites

- *Avoidance of algorithmic bias:* AI systems should be designed to avoid bias in both input data, modelling and algorithm design. Algorithmic bias is a specific concern which needs specific mitigation techniques. Applications should specify the steps which will be taken to ensure data about people is representative and reflects their diversity. Similarly, applications should explicitly document how errors will be avoided in input data and in the algorithmic design which could cause certain groups of people to be represented incorrectly or unfairly. This needs to consider inferences drawn by the system which have the potential to unfairly exclude or in other ways disadvantage certain groups of people.
- *Universal accessibility:* Whenever possible/relevant, AI systems should be designed so that they are usable by different types of end-users with different abilities. Applications are encouraged to explain how this will be achieved, such as by compliance with relevant accessibility guidelines. Moreover, AI systems should avoid functional bias in being designed to offer the same level of functionality and benefits to end-users with different abilities, beliefs, preferences and interests, to the extent possible.
- *Fair impacts:* Applications should demonstrate that possible social impact on relevant groups has been considered and what, if any, steps will be taken to ensure the system does not cause them to be discriminated against or stigmatized, or otherwise have their interests affected in a negative way.

2.4 Social and Environmental Well-being

Something has well-being when its needs are met and it is able to function properly. People can only achieve well-being if they are able to work towards their ambitions and live whatever they consider to be a "meaningful" life.

Ethical requisites

- AI systems should take the welfare of all stakeholders into account and not reduce their well-being. It should be identified who the end-users and stakeholders will be of the application. It should then be assessed how the application could both enhance and harm their well-being, and documented choices should be made in development to support well-being and avoid harm to it.
- AI development should be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. There should be documented efforts to consider the environmental impact of the system and, where needed, steps to mitigate negative impacts. In the case of robotics systems this must include the materials used and decommissioning procedures.
- AI systems with an application towards media, communications, politics, social analytics, and online communities should be assessed for their potential to negatively impact the quality of communication, social interaction, information, democratic processes, and social relations, for example by supporting uncivil discourse, amplifying fake news, segregating people into filter bubbles and echo chambers, creating asymmetric relations of power and dependence, and enabling political manipulation of the electorate. Mitigating actions should be taken to reduce the risk to such harms.
- AI and robotics systems should not reduce safety in the workplace. If relevant, your application should demonstrate consideration of possible impact on workplace safety, and compliance with IEEE P1228 (Standard for Software Safety).

2.5 Transparency

Transparency directly enables human agency, data governance, oversight and human governance. Transparency includes *all* elements relevant to an AI system: the data, the system and the processes by which it is designed, deployed and operated. Without this level of transparency, a decision cannot be contested, or even understood. This would make it impossible to correct errors and unethical occurrences. The degree to which transparency is needed depends on the context and the severity of the consequences. However, it is important to note this is a judgement call, not a precise calculation, and others may not set boundaries or assess severities in the same manner as the researcher, so the precautionary principle dictates it is better to go too far than not far enough. This is why we recommend, if possible, that the decisions concerning the design and the use of AI systems are made by a carefully constructed group, whose composition is sufficiently diverse so as to ensure a representative range of perspectives behind these decisions. Where the formation of a formal group is not possible, it is recommended researchers take steps to ensure they understand the full range of positions others may take.

Ethical requisites

- There is a general requirement for traceability across all areas of ethical AI. When building an AI solution one should consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit.
- It must be made clear to end-users that they are interacting with an AI system – especially for systems that simulate human communication, such as chatbots.
- The purpose, capabilities, limitations, benefits and risks of the AI system and of the decisions conveyed by it must be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly. Wherever it is necessary that people can audit, query, dispute or seek to change AI or robotics activities, your application must explain how this will be possible. It is not sufficient to merely consider the structure and functionality of the system in this respect. You must explain governance and other organisational processes by which your project will receive and assess requests from third parties.
- Whenever relevant, an application should offer details about how decisions made by the system will be explainable to users. Where possible this should include the reasons why the system made a particular decision. However, with some systems this may not be possible. Nevertheless, the system (or those deploying it) should always have a mechanism by which to explain what the decision was and what data was used to make that decision. Explainability is especially a requirement for systems that make decisions and recommendations and perform actions for which accountability is required, such as decisions and actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests.
- The design and development processes will involve making decisions about ethical issues, such as how to remove bias from a dataset. The requirement for transparency means your development processes (and tools) will need components to keep records of such decisions so that it is possible to trace how these ethical obligations were met. This information may be required for audits, for disputing or resolving decisions made by the system, for correcting unexpected ethical issues which arise after system deployment and so that your own teams can learn and improve their handling of ethical issues.

2.6 Accountability and Oversight

Human oversight as a value requires humans are able to understand, supervise and control the design, development, deployment and operation of AI systems. Oversight depends on accountability because one cannot control something unless one has information about it. Accountability means being able to explain how and why a system exhibits particular characteristics.

Ethical requisites

- AI systems should allow for human oversight regarding their decision cycles and operation, unless compelling reasons can be provided which demonstrate such oversight is not required. It should be explained how humans will be able to understand the decisions made by the system and what mechanisms will exist for humans to override them.

- The application should provide details of how ethically and socially undesirable effects of the system will be detected, stopped, and prevented from reoccurring.
- *To a degree matching the type of research being proposed (from basic to precompetitive) and as appropriate*, the application should include a formal ethical risk assessment for the proposed AI system. There should be documentation for the procedures for risk assessment and mitigation after deployment.
- Whenever relevant, it should be considered how end-users, data subjects and other third parties will be able to report complaints, ethical concerns or adverse events and how they will be evaluated and actioned. The requirement for transparency means a mechanism should be included to communicate with these third parties has been done with their information.
- As a general principle, all AI systems should be auditable by independent third parties. The procedures and tools available under the XAI[17] approach support best practice in this regard. This is not limited to auditing the decisions of the system itself but will need to discuss procedures and tools used during the development process. Where relevant, the system should generate human accessible logs of the AI system's internal processes.

2.7 Additional components

These guidelines constitute core research ethics guidelines for AI. They do not, however, constitute a complete set of research ethics guidelines for these fields. As said, there are several ethics guidelines that apply to research in general, and therefore also to AI R&D. These would have to be added to the AI-specific guidelines to arrive at a complete set. In the next section, we demonstrate how this is done.

In addition, we propose three ways in which these research ethics guidelines may be operationalized and supplemented further. First, as we claimed, these are high-level guidelines. There is a gap between their somewhat abstract descriptions and the descriptions that researchers and developers use in their everyday activity. Researchers and developers may therefore find them difficult to apply in practice if no further translation is given to the scientific and technological vocabulary that they use in everyday practice. We proposed the *Ethics by Design* approach as an approach for closing this gap. Ethics by Design is a systematic way of integrating ethical considerations and guidelines in the R&D process. It uses the high-level research ethics guidelines for AI & robotics as a basis for detailed guidelines that are integrated at different steps in the development process and that provide instructions in a more technical language that are easier to apply for scientists and engineers in their everyday practice. See the framework for Ethics by Design elsewhere in this deliverable.

Second, the top-down approach of starting with general guidelines and then deriving operational guidelines has an important limitation, which is that it provides guidance of a rather general nature, the aim of which is to support rather abstract values like privacy and fairness. An alternative to this approach is one that takes as its point of departure not values and top-level guidelines in which these are prescribed, but specific technologies, techniques and artifacts in the field of AI, for which tailored ethics guidelines are then developed. This may be called a *technology-centred* approach to research ethics, as opposed to a *value-centred* approach. In a technology-centred approach, the point of departure is technologies, techniques and artifacts, like data analytics, machine learning, intelligent

[17] Adadi, Amina, and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, Vol. 6, 2018, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052. https://ieeexplore.ieee.org/abstract/document/8466590 for an overview.

agents, social robots, and natural language processing systems, it is then investigated what ethical issues are raised by them, and ethics guidelines are subsequently developed for the mitigation of these issues in R&D. While these ethics guidelines could be based in part on top-level or high-level ethics guidelines, they could also be articulations of moral intuitions with respect to these technologies, and they will typically make reference to multiple moral values and principles, including, possibly, ones that are not included in the set of top-level guidelines. For instance, they may make reference to other moral values and principles not mentioned in them, like bodily integrity, truthfulness, authenticity and virtuousness.

The benefit of a technology-centred approach is that is capable of providing dedicated ethical guidance for specific technologies in a way that is not possible in a value-centred approach. A technology-centred approach recognizes the specific character of particular technologies and the particular ethical issues that are associated with it. These ethical issues may be governed by some of the high-level and top-level guidelines that we have proposed, but they often involve a complex interplay of moral values and corresponding guidelines, as is the case, for example, in facial recognition technology, which raises intertwined issues of privacy, security, agency, identity and fairness.

We therefore recommend that the high-level research ethics guidelines of the value-centred approach are supplemented with technology-centred ethics guidelines for the most important technologies, techniques and artifacts in AI that raise specific ethical issues. We propose that these guidelines are included in a *special topics* section that is included after the high-level guidelines. We propose that this special topics section also includes ethics guidelines for the development of technologies for particular application domains, such as healthcare, defence, law enforcement and entertainment. For new technologies, it is often known that they will be used in particular application domains, or at least it can be foreseen that they will be. Ethics guidelines for technologies in relation to application domains guide development choices that for the mitigation of ethical issues that can occur in particular domains. A proposed list of special topics, with explanatory notes, can be found in the Appendix to this report.

The recommended use of the high-level and special topics guidelines is that researchers standardly apply the high-level guidelines in their R&D practices, but that they also determine whether their R&D practices also includes one or more techniques and artifacts listed in the special topics section. If so, then these technology-specific ethics guidelines should be applied as well. In (rare) cases in which there appears to be a conflict between the high-level and technology-specific guidelines, neither automatically takes priority, and a considered moral judgment will have to be made on how the guidelines should be weighed against each other.

Third, the ethics guidelines can be supplemented with supporting methods for ethical assessment, resolution of conflicts between ethical guidelines, and stakeholder engagement. Ethically guided R&D practices could be the result of the application of the proposed guidelines, but such application will not cover all ethical issues, and is not always straightforward, and a more extensive ethical assessment of the technologies that are being researched and developed could result in better inclusion of ethical considerations in R&D. In SIENNA deliverable D6.1, *Methods for ethical analysis of emerging technology fields*[18], we propose methodologies for such ethical analysis.

[18] Brey, Philip, "Research Ethics Guidelines for the Engineering Sciences and Computer and Information Sciences", in Kelly Laas, Elisabeth Hildt, and Michael Davis (eds.), *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields*, Springer, Dordrecht, Netherlands, 2021.

The resolution of conflicts between ethical guidelines is also an issue that worth addressing. Researchers sometimes face a choice between upholding different guidelines. For example, one way of designing a surveillance system may protect privacy, while another affords greater security and reduced risk to harm. How to resolve such value conflicts? This is a difficult and complex issue, but there are some methods that aid the resolution of such conflicts. Jansen et al.[19] in section 3.8 discusses methods for resolving such conflicts.

Stakeholder engagement, finally, is a recommended practice in research ethics for AI, because a consideration of stakeholders and their preferences and opinions will allow for better judgments on values and guidelines that are at play, trade-offs between values and guidelines when there is conflict, and acceptable R&D solutions. Stakeholders can be engaged in different ways. In Brey et al.[20] we discuss the engagement of stakeholders in ethical analysis, and the Ethics by Design report, which is part of this deliverable, contains proposals for the inclusion of stakeholders in R&D.

3. Stand-alone research ethics guidelines for AI

Stand-alone research ethics guidelines for AI are guidelines for the exclusive assessment of R&D projects and practices in the AI field. They are stand-alone in that they are not incorporated into previously existing research ethics framework and suffice by themselves to guide research in AI. They would be used by RECs that are specialized in assessing R&D in these fields (for example, a REC of an AI or robotics company) or would be utilized by RECs with a broader mandate, for the specific aim of assessing AI R&D.

We propose that these stand-alone ethics guidelines include the six sets of guidelines proposed in the previous section, plus any AI-specific guidelines pertaining to special topics, as discussed in section 2.7. In addition, however, more general ethics guidelines would need to be included that apply to R&D more generally, and that apply to digital technologies more generally.

In previous work[21], we have argued that five ethical principles are so central to the process of doing research that guidelines based on them should be included in any research ethics guidance framework. They are:

- Protection of and respect for human research participants
- Protection of and respect for animals used in research
- Protection of researchers and the research environment
- Protection and management of data and responsible dissemination of research results
- Social responsibility

[19] Jansen, Philip, Wessel Reijers, David Douglas, Agata Gurzawska, Alexandra Kapeller, Philip Brey, Rok Benčin, and Zuzanna Warso, “A reasoned proposal for shared approaches to ethics assessment in the European context”, SATORI Deliverable D4.3, EU FP7 Project, 2016. https://satoriproject.eu/media/D4.1_Proposal_Ethics_Assessment_Framework.pdf.

[20] Brey, Philip, op. cit., 2021.

[21] Jansen, Philip, Wessel Reijers, David Douglas, Agata Gurzawska, Alexandra Kapeller, Philip Brey, Rok Benčin, and Zuzanna Warso, “A reasoned proposal for shared approaches to ethics assessment in the European context”, 2016.

The first three protect sentient beings and valuable items that are immediately connected to the research activity. The fourth ensures that research data is managed responsibly, that personal data is protected, and that research results are disseminated in a responsible way. The fifth, finally, ensures that social consequences of the research activity are assessed and mitigated where possible. For a much more detailed statement of guidelines in these five categories, see CEN.[22]

Social responsibility is already heavily implied in the AI core guidelines, since they emphasize a wide range of social issues for which responsibility should be taken, including the broad category of social environmental well-being. The others are, however, not contained in these guidelines and should therefore supplement the core guidelines. It might be believed that the privacy and data governance guidelines for AI are contained in the guidelines for the protection and management of data, mentioned above. However, this is not exactly the case. The AI privacy and data governance guidelines prescribe how an AI system or technology should be designed to adequately protect personal data and be involved in the responsible governance of data. The protection and management of data guidelines concern the responsible management of data that is collected for and within a research project. It is possible that there is no overlap between the two: the data collected for the research project and any data processed by the system – during or after the research project ends – need not be the same. Regarding the guidelines for human research participants and animals, it could be objected that they are not important because much research in AI does not involve them, but some of it may, and for this reason one needs guidelines for their proper treatment.

In conclusion, then, we recommend that stand-alone research ethics guidelines for AI combine the AI guidelines of section 2 with the five general research ethics guidelines as presented here. This could result in a framework with eleven categories of guidelines (the six from section 2 plus the five proposed here), but some categories could be combined. In particular, the privacy and data governance category of section 2 could be merged with the protection and management of data category presented here, and the social responsibility category proposed here could be merged in part with the social and environmental well-being category of section 2, and in part with its accountability and oversight category.

4. Incorporating AI ethics guidelines into guidelines for computer science and information technology

In universities and technology companies, R&D in AI will often be included in a broader portfolio of R&D on digital technologies. In such instances, there may not be a desire to have stand-alone ethics guidelines for AI. Instead, it would be desirable to include these guidelines in a broader set of ethics guidelines for computer and information science or for information technology R&D. Interestingly, though, there is currently hardly an established tradition of research ethics for the computer and information sciences. There is a tradition of professional ethics for computer scientists, which has been around since the field was still young. The first code of ethics for computer scientists was developed in 1973 by the Association for Computing Machinery in the United States. In addition, there is a

[22] CEN, Ethics assessment for research and innovation - Part 1: Ethics committee, CEN Workshop Agreement, 17145-1:2017 E, 2017. http://satoriproject.eu/publications/cwa-part-1/

tradition of ethical reflection on computing and information, which has emerged in the mid 1980s under the name of "computer ethics" and later also "information ethics".[2324]

Research ethics guidelines and RECs for computer science have been in existence, however, only since very recently.[25] Initially, their scope has been narrow, however, with a strong focus on privacy and data protection issues, and various ethics guidelines have been developed to specifically address issues of privacy and data protection – though in most cases these are not aimed at the development of information technology but at its use.[2627] One of the very few proposals for broader research ethics guidelines are the Menlo report ethics guidelines for information and communication technology research for the U.S. Department of Homeland Security.[2829] These guidelines have a focus, however, on human subjects research only.

A broader proposal was developed in the EU-funded SATORI project, a project on approaches and methods for ethics assessment. SATORI developed, in association with the European Committee for Standardization,[30] a standard for research ethics committees, which included a proposal for ethics guidelines for the computer and information sciences.[31] The CEN proposal includes a set of guidelines for all scientific fields, followed by guidelines for specific fields. For any field, therefore, the recommended guidelines consist of the general research ethics guidelines plus the field-specific ethics guidelines proposed in the report. The set of guidelines for all scientific fields consists of the five categories proposed in the previous section. They are guidelines for protection of and respect for human research participants, protection of and respect for animals used in research, protection of researchers and the research environment, protection and management of data and responsible dissemination of research results, and social responsibility.

The additional guidelines for the computer and information sciences proposed from the CEN proposal can be summarized as follows:

Protection of privacy personal data

- Ensure that new research concepts and innovations do not pose any unjustified inherent risks to the right of individuals to control the disclosure of their personal data;
- If research concepts and innovations involve the combination of multiple data sources, carefully consider the effects on (informational) privacy;

[23] Johnson, Deborah G., *Computer Ethics,* Prentice-Hall, Englewood Cliffs NJ, 1985.
[24] Tavani, Herman T., *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, 5th ed. Wiley, 2015.
[25] Søraker, Johnny and Philip Brey, Ethics Assessment in Different Fields: Information Technology, Annex 2.b.1 to SATORI Deliverable D1.1, EU FP7 Project, 2015. http://satoriproject.eu/media/2.b.1-Information-technology.pdf
[26] European Commission, op. cit., 2018.
[27] Wright, David, "The state of the art in privacy impact assessment", *Computer Law & Security Review*, Vol. 28, No. 1, 2012, pp. 54-6.
[28] Dittrich, David and Erin Kenneally, *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*, Tech. rep., U.S. Department of Homeland Security, Aug 2012.
[29] Dittrich, David, Erin Kenneally, and Michael Bailey, Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Menlo Report, Tech. rep. U.S. Department of Homeland Security, Oct 2013. http://dx.doi.org/10.2139/ssrn.2342036
[30] CEN, op. cit., 2017.
[31] Ibid.

- If research concepts and innovations involve the development of capabilities for, or the use of, data surveillance or human subject monitoring or surveillance, then invoke the requirement for informed consent, if appropriate. Strike an appropriate balance between the need to monitor and control personal information and the right of individuals to (informational) privacy and other human rights.

Avoidance of security risks

- Ensure that new research concepts and innovations offer reasonable protection against any potential unauthorized disclosure, manipulation or deletion of information and against potential denial of service attacks, e.g. protection against hacking, cracking, cyber vandalism, software piracy, computer fraud, ransom attacks, disruption of service;
- Ensure that new research concepts and innovations, by themselves or through their use in a system, do not pose inherent direct or long-term risks of harm to public health and safety, e.g. information and communications technology (ICT) innovations used in healthcare, ICT innovations used in the monitoring and control of public infrastructure, ICT innovations that could lead to addiction;
- Do not engage in research that involves attempts to make unauthorized access to telephone systems, computer networks, databases or other forms of ICT; such research is illegal and unethical, regardless of motivation;
- Treat with extreme caution the dissemination of research involving the identification of undiscovered security weaknesses in existing systems;
- Avoid practical experiments with computer viruses or perform them in a controlled environment, and exercise extreme caution in the dissemination of the results of paper-based (theoretical) computer virus experiments;
- Carry out any experiments in breach security on designated, standalone (offline) computers or on designated isolated networks of computers.

Respect for freedom of expression

- Ensure that new research concepts and innovations do not pose unjustified inherent risks to the freedom of individuals to express themselves through the publication and dissemination of information, or to their freedom of access to information;
- If research or innovation involves the use of censorship methods, strike an appropriate balance between the need for content control and the right of individuals to express themselves freely.

Respect for intellectual property

- Ensure that new research concepts and innovations do not pose unjustified inherent risks to the intellectual property rights of individuals or organisations;
- Avoid research that could generate copyright issues, such as research involving peer-to-peer networking or file sharing and distribution.

Respect for other individual rights and liberties

- Ensure that new research concepts and innovations do not pose inherent risks to autonomy, authenticity or identity. In particular, ensure that information systems do not unnecessarily or unjustifiably take away control from users by limiting their choices or making choices for them that they would prefer to make themselves;
- Ensure that decisions made by information systems that have significant social impact take into account the rights, values and interests of stakeholders, including users, and make efforts to

ensure that the reasons for decisions made by information systems can be retrieved, so as to make the systems accountable;
- Take into account the issue of how responsibilities and liabilities are assigned between humans and machines when information systems are involved in decision-making.

Avoidance of harms to justice and equality

- Consider how new research concepts and innovations could widen or narrow social inequalities in terms of the distribution of opportunities, powers and capabilities, civil and political rights, economic resources, income, risks or hazards;
- Consider how new research concepts and innovations could harbour or counter unjust bias in terms of age, gender, sexual orientation, social class, race, ethnicity, religion or disability;
- Consider how new research concepts and innovations could harm or promote the interests of vulnerable, disadvantaged, or underrepresented groups and communities in society, including those in low income and lower-middle income countries.

Promotion of well-being and the common good

- Consider how the research or innovation activity could harm or promote the general well-being of individuals and groups in society (e.g. effects on the quality of work or quality of life);
- Consider how the research or innovation activity could harm or promote the social skills and behaviour of individuals, and how it could harm or promote the learning or exercising of important virtues, such as patience and empathy;
- Consider whether and how the research or innovation activity could harm or promote important social institutions and structures, democracy, and important aspects of culture and cultural diversity.

Promotion of environmental sustainability

- Optimize technologies for effective and cost-efficient resource use (including raw materials and energy), for resource recovery (recycling), and for lowering the production of environmentally harmful wastes and environmental pollution.

Dual use of computer and information sciences research and innovations

- Consider whether new research concepts and innovations could have military applications;
- Consider whether new research concepts and innovations could contribute to the proliferation of weapons of mass destruction;
- Consult proper authorities before publishing and adhere to relevant national and supra-national regulations if a technology has significant military applications or if it contributes significantly to the proliferation of weapons of mass destruction. Even if publication is allowed, find a proper balance between security and freedom of publication.

The privacy and data protection provisions in the CEN proposal are somewhat cursory; we now recommend more extensive guidelines such as those found in the EU Horizon 2020 research ethics framework.[32]

[32] European Commission, Horizon 2020 Programme Guidance. How to complete your ethics self-assessment, Directorate-General for Research & Innovation, Feb 2019. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

Integration of the research ethics guidelines that we have proposed for AI could proceed as follows. First, it needs to be identified which of the ethics guidelines for AI are not really specific to AI but would apply equally to other forms of information technology. When this is the case, it is proposed to include the general version of the guideline, and not a specific version for AI. For example, the AI proposal includes a guideline which states: "AI systems should take the welfare of all stakeholders into account and not reduce their well-being," This would surely apply to other information technologies as well, and in fact, this guideline closely matches the computer science guidelines on the promotion of well-being and the common good. So we propose to merge these guidelines here and remove the reference to AI.

Second, it needs to be identified which of the ethics guidelines for AI are specific to AI, but are also part of a larger category of guidelines that applies to computer science. In such cases, we recommend that the AI guideline is maintained, but is included in the larger category. For example, one of the AI-specific guidelines states: "It should be clear to people whether they are interacting with an AI system. They should be informed about the system's abilities." This guideline could be included in the category of computer science guidelines on "Respect for other individual rights and liberties", maintaining its explicit reference to AI systems. Third and finally, it needs to be identified if the AI ethics guidelines introduce new categories of ethics guidelines in the computer science guidance framework. Candidates are the categories of agency, transparency, and accountability and oversight.

Our recommendation here is to include categories of transparency and accountability and oversight in the computer science framework, and to combine the category of agency with the category of "respect for other rights and liberties" into a category of "agency, autonomy and other rights". Conceivably, all of these categories will contain both guidelines that are AI-specific and ones that apply to information technology as a whole.

A resulting set of ethics guidelines for computer science / information technology could therefore contain the following categories of guidelines, combining general research ethics guidelines, ethics guidelines for computer and information sciences such as SATORI, and ethics guidelines for AI:

- Protection of and respect for human research participants
- Protection of and respect for animals used in research
- Protection of researchers and the research environment
- Privacy and data management
- Responsible dissemination of research results
- Social responsibility (general research ethics guidelines)
- Avoidance of security risks
- Respect for freedom of expression
- Respect for intellectual property
- Agency, autonomy and other rights
- Avoidance of harms to justice and equality
- Promotion of well-being and the common good
- Transparency
- Promotion of environmental sustainability
- Dual use of computer and information sciences research and innovations
- Accountability and oversight

Finally, a word about robotics. Robotics is, historically, part of the engineering sciences, being rooted in electrical and mechanical engineering in particular. Increasingly, however, robotics also includes concepts and methods from computer science, including AI. Research ethics for robotics could therefore be integrated in research ethics guidelines for engineering science as well as in computer science guidelines. In case the former is the preferred option, we recommend the inclusion of the AI guidelines in broader research ethics guidelines for engineering, using a method similar to the one here proposed for the computer and information sciences. For a proposal for research ethics guidelines for the engineering sciences, see the SATORI – CEN standard document[33] and the summary and discussion in Brey.[34]

5. Incorporating AI ethics guidelines into research ethics guidelines that span multiple disciplines

Research performing organisations sometimes have a central research ethics committee that covers multiple disciplines. For example, a university may have a single central REC that assesses research in in all science and engineering fields, and in social sciences and humanities. It may employ a set of research ethics guidelines for this purpose that contain both guidelines that apply to all fields as well as ones that apply to specific fields. In such instances, we recommend integration of the AI guidelines in a manner similar to the one proposed in the previous section for the integration of these guidelines with guidelines for computer science.

One novelty in this setting, however, is that research involving AI will often not be focused on their development, but on their deployment and use. Research projects in social science may for example use AI-driven processing of large data sets to generate research results, they may study the use of an AI system to regulate traffic, they may study the use of robots in therapy for children with autism. Similarly, projects in engineering and in medicine may include the deployment and use of AI technology. It should be added that the same applies to projects in computer science; some of them also focus on the use of information technology rather than its development.

The research ethics guidelines that were proposed for R&D in AI in section 3 focus on technology development, and do not necessarily apply to deployment and use. Generally speaking, the guidelines of section 3 prescribe features that AI systems should possess and benefits that they should provide, as well as actions that should be taken by developers. For deployment and use, one generally wants these same features and benefits to be present in the systems that one selects, but the actions that should be taken will be different: instead of actions directed at responsible development, the required actions should be directed at responsible deployment and use. Moreover, specific applications of AI may require special features and benefits to be present in AI systems that are deployed. For example, the use of AI systems in criminal justice may require enhanced qualities of transparency, explainability and accountability so as to protect the rights of defendants.

We therefore recommend that ethics guidelines for AI that are incorporated into multidisciplinary research ethics frameworks include guidelines for both development and for deployment and use. The ethics guidelines for deployment and use will find a partial basis in the guidelines for development in sections 2 and 3. Wherever it says there that an AI system should be designed to have feature *X* or

[33] CEN, op. cit., 2017.
[34] Brey, Philip, op. cit., 2021.

provide benefit *Y*, a corresponding guideline can be developed which states that AI systems that are deployed should have feature *X* or provide benefit *Y*. In addition, specific features or benefits may be required for specific applications, e.g., applications in healthcare, education, research, and defence, amongst others. Also, guidance may be provided for deployers and users on actions that they should take to deploy and use the technology in a responsible way.

In the EU-funded SHERPA project, we proposed ethics guidelines for the deployment and use of AI technologies that focus on actions that are to be performed by different actors in the deployment and use process.[35] These guidelines have been adapted for a research ethics context in another annex to this document, *Ethics by Design and Ethics of Use approaches for artificial intelligence and robotics applications*. Our guidelines are divided up into four categories: project planning and management, acquisition, deployment and implementation, and monitoring. These guidelines advocate for a planned approach for the inclusion of ethical considerations in the deployment and use of AI systems in research, which starts with and assessment of and planning for potential ethical issues raised by the deployment and use of these systems, followed by ethically guided acquisition and procurement deployment and implementation processes, in which different actors all take up some of the responsibilities that this involves. Continuous monitoring for compliance and for new ethical issues is recommended in addition.

6. Conclusion

In this report, we proposed a set of research ethics guidelines for artificial intelligence (AI) and discussed how these can serve as a basis for stand-alone guidance for AI R&D and how they can be incorporated into broader research ethics frameworks for computer and information science, and for frameworks that span multiple disciplines. We based our proposal for guidelines for AI on the general ethics guidelines for AI proposed by the EU High-Level Expert Group on AI. Development of research ethics guidelines based on general ethics guidelines is not a straightforward process and requires an empirical analysis of the R&D setting as well as a normative analysis of the responsibilities of scientists and engineers for ethical aspects of their R&D practices. Our proposal included 27 research ethics guidelines (which we called ethical requisites) that were grouped into six categories, under the headings of Human agency, privacy & data governance, fairness, social and environmental well-being, accountability & oversight, and transparency. We also proposed, next to these guidelines, the inclusion of “special topics” guidelines, which are research ethics guidelines for specific techniques, products, and application domains in the AI field. We also proposed some other research ethics tools for AI, in particular an Ethics by Design approach, which provides a comprehensive way of integrating ethical guidelines and criteria into design methodologies.

Next, we discussed stand-alone research ethics guidelines for AI, arguing that these should be a combination of the proposed ethics guidelines for AI and general ethics guidelines for research. We then discussed how the ethics guidelines for AI can be integrated into broader research ethics frameworks for computer and information sciences, presenting a research ethics framework for this field and then proposing how integration could succeed. We did the same for research ethics frameworks that span multiple fields and that contain research in which AI technology is not developed but deployed and used. We argued that special guidelines are needed for the deployment and use of

[35] Brey, Philip, Björn Lundgren, Kevin Macnish, and Mark Ryan, op. cit., 2019.

these technologies in research, although these can be based in part on the ethics guidelines for development that we proposed in section 2.

We think that this report may contain the first comprehensive proposal for research ethics guidelines for AI, including their integration into broader research ethics frameworks. We encourage universities, companies, research funding organisations, and other organisations involved in the ethical assessment of research to utilize this report as a manual for compiling their own research ethics frameworks for AI.

Appendix: Special Topics

The following is a list of special topics in AI for which we recommend the inclusion of dedicated research ethics guidelines. For a full statement of such guidelines, see the guidance document on special topics that will be included in the Ethics Review framework of the new EU Horizon Europe programme, in the section on AI. Please note that this listing is necessarily incomplete, and other relevant techniques, products and applications may be added.

1. Ethical guidance for AI techniques and methods
    - 1.1 Algorithms
    - 1.2 Knowledge representation and reasoning techniques
    - 1.3 Automated planning and scheduling
    - 1.4 Machine learning
    - 1.5 Machine ethics
    - 1.6 Robot sensing, actuation and control
    - 1.7 Data analytics
2. Ethical guidance for types of products and systems
    - 2.1 Intelligent agents
    - 2.2 Knowledge-based systems
    - 2.3 Decision support systems
    - 2.4 Computer vision systems
    - 2.5 Natural language processing systems
    - 2.6 Affective computing systems
    - 2.7 Big Data analytics systems
    - 2.8 Embedded AI and Internet of Things
    - 2.9 Autonomous intelligent systems
    - 2.10 Humanoid robots
    - 2.11 Social robots
    - 2.12 Robotic exoskeletons
    - 2.13 Robots - other
    - 2.14 Tracking, behaviour analytics, facial recognition, biometrics and surveillance
    - 2.15 Covert and deceptive AI and big data systems
3. Ethical guidance for different application domains
    - 3.1 Infrastructure & cities
    - 3.2 Healthcare
    - 3.3 Finance and insurance
    - 3.4 Transportation
    - 3.5 Defence

Annex 5 – AI Ethics Education, Training and Awareness Raising

AI Ethics Education, Training And Awareness Raising

Annex 5 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Philip Brey, *University of Twente* (p.a.e.brey@utwente.nl)
Other contributors	Brandt Dainow, *University of Twente*

Due date	February 2021
Type	Report (Annex 5 to D5.4 deliverable)
Dissemination level	PU = Public

Keywords	Ethical issues; artificial intelligence; AI; robotics; robots; ethics; software design; ethics education; training

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This document outlines proposals for developing awareness and skills regarding ethical AI and robotics systems within society. It covers the various needs, and proposes solutions, in Higher Education curricula, industry training, product certification and the means by which to raise awareness within the general public. We have sought to minimise risk and increase the chances of success, by limiting our proposals to established methodologies and proven approaches. With regard to Higher Education, we outline the case for integration of ethical education within Computer Science and Engineering, as well covering the needs of other disciplines, such as law and business, to understand relevant ethical AI concerns within their speciality. Finally, we provide a best practice case study of a module taught at Harvard University which exemplifies the ideal format we recommend. Moving to commercial industry, we briefly outline policy proposals for the development of commercial certification schemes for products and people. We then discuss methods by which these schemes can obtain popular support, such that it becomes profitable for AI developers to pursue such certifications. This document concludes with a brief discussion of methods by which to promote general public awareness of ethical AI and its issues.

Table of contents

Executive summary

This document outlines in broad terms strategies for increasing awareness of, and competency to handle, ethical issues of Artificial Intelligence systems and robots. The document discusses the needs of higher education, commercial industry and the general public.

We discuss the needs for improving ethical awareness and the ability to resolve ethical issues in Computer Science and Engineering, where the majority of AI and robotics developers will come from. Here we recommend a deep integration of ethical training within the standard curricula within these fields and not just in specialised degrees tuned towards AI or robotics. We use as a case study a module from Harvard University's Computer Science program which we believe illustrates a practical, proven and best-practice approach. This case study demonstrates the degree to which it is possible to integrate ethical education by philosophers deeply into highly technical Computer Science courses. We also discuss the need for some degree of ethical AI awareness in many other disciplines, such as law and business. Here graduates can be expected to assess, purchase and use AI and robotics systems and so will need a good understanding of the ethical issues AI or robotics systems can generate in their professional lives.

Our recommended strategy for industry is to encourage the development of certification schemes for products and people. Here our aim is the development of a self-sustaining business ecosystem providing training towards certification, certification exams and product labelling schemes. We do not propose the development of new initiatives, instead recommending encouragement of existing schemes to expand into this area, as has occurred, for example, with data protection. We also recommend the development of a product labelling scheme modelled on the Energy Rating Labelling scheme. We recommend strategies to generate market demand for these certifications on the premise that industry will adopt ethical AI if the market demands it.

Finally, we discuss raising public awareness of ethical AI through existing public communication channels and civic bodies, and the addition of relevant ethical AI considerations into codes of professional through bodies such as the Council of European Professional Informatics Societies (CEPIS).

List of tables

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
CEARM	Central Ethical Artificial Intelligence Reference Model
CEPIS	Council of European Professional Informatics Societies
COBIT	Control Objectives for Information and Related Technology
IEEE	Institute of Electrical and Electronics Engineers
MOOC	Massive Open Online Course
OECD	Organisation for Economic Co-operation and Development

Table 1: List of acronyms/abbreviations

1. Introduction

There has been a growing awareness in recent years that artificial intelligence (AI) raises significant ethical issues.[1] Current developments in AI include such techniques as deep learning and genetic algorithms, and products such intelligent agents, computer vision systems, natural language processing systems, and big data analytics systems. In various applications contexts, these developments may raise issues, which range from issues of safety and security to issues of responsibility and accountability, and from issues of privacy to issues of justice and fairness.[2]

It is increasingly recognized that if AI technology is to be beneficial to society, ethical issues must be recognized and addressed. Various methods and instruments are being developed to address ethical issues, such as detailed ethical guidelines for the development and use of AI, Ethics by Design methodologies to address ethical issues in the development of AI, research ethics frameworks, ethics standards and certification for AI, and ethics-inspired regulations and policies.

What is increasingly also being recognized is that in order to promote ethics for AI, and for professionals to utilize the various methods and instruments that are being developed, education and training are needed. Professionals need education and training so they can recognise and decide on ethical issues relating to AI in their work, and to utilize methods and instruments to deal with them. This document discusses the education and training needs which currently exist and how these can be properly addressed. It also discusses the more general objective of awareness-raising: how can we ensure that the general public is aware of, and knowledgeable about, the ethical issues relating to AI?

2. AI Ethics In Higher Education

2.1. For which students?

It is increasingly accepted that AI ethics should be an important component in the curricula of students who pursue degrees in AI. However, many AI developers come from less directly-related degree programs in engineering or computer science, such as computational thinking, data analytics or even just general programming. Moreover, in general degree programs which have some relevance to AI, such as IT law, business or innovation management, AI is becoming increasingly central, which also warrants having AI ethics covered in the curriculum.

Furthermore, AI is an enabling technology which will deeply affect all sectors of society, especially healthcare, education, government, retail and media. Increasingly, students whose curricula focus on such sectors cannot avoid learning about AI and its impacts. As part of any course which covers the impact of AI on their sector, it would be appropriate to include ethical issues relating to AI. This would affect study programs in many areas, especially in applied social science such as governance studies, management science, health sciences, architecture, communication studies, media studies and educational sciences. In traditional social science fields such as sociology, law, politics and psychology, studies of the impact of AI on their subject matter, including ethical issues, would seem relevant.

[1] The authors of this report acknowledge the input of various experts and stakeholders to this text. Please see the Acknowledgement section of SIENNA D5.4 (Feb 2021) for a list of these people.

[2] Please see SIENNA D4.4 for a detailed account of ethical issues in AI. Philip Jansen, Brey, P, Fox, A., Maas, J., Hillas, H., Wagner, N., Smith, P, Ouoch, I., Lamers, L., van Gein, H., Resseguier, A., Rodrigues, R., Wright, D., and Douglas, D., SIENNA D4.4: Ethical Analysis of AI and Robotics Technologies, 2020, https://doi.org/10.5281/zenodo.4068082.

Furthermore, given the increasing importance of AI applications in science, engineering and medicine, consideration of ethical issues relating to the use of AI in these fields is also warranted.

A special place can be assigned to the discipline of philosophy, which includes the field of ethics. In applied ethics programs, a course on AI ethics appears relevant giving the increasing importance of AI ethics as a field of applied ethics. It is also conceivable, in the future, that specialized master programs in ethics of AI, or multidisciplinary programs in AI ethics, law and governance or AI and society will be developed.

Because of its vast implications for all sectors of society, it is becoming apparent that AI is a multi-disciplinary field, with people specializing not just in AI technology, but also in the application of AI in different social sectors, and in areas such as AI law, AI ethics and AI governance. The multidisciplinary nature of research in AI also suggests the importance of multidisciplinary courses in which students from different disciplines participate to learn from other perspectives and how to combine them. A course in AI ethics which is open to diverse disciplines could have this role of bringing different disciplines together. Alternatively, a course in AI and Society, which studies the interaction between AI and society and the various ways to govern this interaction, and in which ethical issues are covered, could have a similar role of bringing disciplinary perspectives together.

Courses which teach the interaction between AI and society are of particular importance. There is a tendency towards technological determinism[3] when evaluating the impact an AI system will have on society, under which it is assumed the social impact can be predicted purely from the functional characteristics of the AI system. This naïve and mistaken view fails to understand that the social impact of any technology largely depends on how society uses it. It is therefore impossible to predict the social effects of AI without understanding the social context in which it will be deployed. Because fields such as engineering and computer science do not typically teach anything about society and how it functions, it is therefore vital these fields introduce such courses.

2.2. Objectives of AI ethics teaching units

We will now cover the objectives of AI ethics teaching for the different student groups identified above. First, what should be the objectives of AI ethics teaching for students in AI and in broader computer science and computer engineering fields? We can distinguish two overall objectives:

- *Applied ethics of AI:* To develop a care for, and understanding of, ethical issues at the intersection of AI and society, including an ability to recognize, analyse, reflect on and discuss such issues, and to propose possible solutions.
- *Professional ethics of AI:* To develop an understanding of professional ethical issues regarding AI and the responsibilities of AI and computer science professionals in relation to these issues, and to develop skills in handling ethical issues relating to AI in one's professional activity.

For students in applied social sciences, who are active in specific domains like government, healthcare, or education, the applied ethics objective also appears relevant, next to a modified professional ethical objective:

- *Professional ethics of AI (for applied social sciences):* To develop a care for, and understanding of, professional ethical issues regarding the application of AI in one's field and one's

[3] Dafoe, "On Technological Determinism".

professional responsibilities in relation to these issues, and to develop skills in handling ethical issues relating to the application of AI in one's professional activity.

Analogous objectives also apply to students in science, engineering and medicine.

Moreover, the applied ethics of AI objective could be focused, for these fields, on those applied ethics issues which occur within the domain parameters of one's field, rather than those of society as a whole. For instance, for those working in healthcare, the applied ethics of AI objective could be the following:

- Applied ethics of AI in healthcare: To develop an understanding of ethical issues at the intersection of AI and healthcare, including an ability to recognize, analyse, reflect on and discuss such issues, and to propose possible solutions.

We furthermore recommend that ethics teaching does not consist solely of the formulaic teaching of principles and their application, and of methods for addressing ethical issues in one's profession. The development of new AI systems will inevitably lead to the rise of unanticipated issues. It is therefore essential workers are able to detect and resolve these themselves. AI ethics in their degree courses should therefore include the teaching of critical thinking and the reasoning skills required to critically assess ethical issues, engage in critical debate with others, and to arrive at meaningful and well-balanced solutions to ethical issues. This requires more than merely understanding the impact of AI on their field. In order to determine appropriate strategies and solutions, one must understand what AI is capable of, and also what its limits are (significant ethical problems have occurred in AI because of an over-estimation of its capabilities). This requires a basic understanding of how AI systems are built and how they work.

Curriculum Content

AI ethics courses or modules can be structured around the applied ethics and professional ethics objectives which apply to the type of programme that they are part of. We will first discuss course content related to applied ethics objectives and then move to professional ethics objectives.

2.3. Applied ethics objectives

The *applied ethics of AI in society objective* can be met through several paths. Firstly, it is recommended that students are acquainted with a number of paradigmatic, real-world, cases which have raised important ethical issues of AI in society. Ideally, these cases would cover a range of key ethical issues associated with AI, including issues relating to fairness, autonomy, accountability, privacy, transparency and well-being. For instance, these could be prominent algorithmic bias cases, prominent examples of AI systems violating privacy at a large scale, and prominent cases of autonomous AI systems which raise issues of accountability.

Secondly, teaching is needed regarding how these cases can be approached through ethical analysis. This includes an exploration of fundamental ethical concepts and principles in AI, such as those mentioned above and of methods of ethical analysis. Instruction could initially focus on ethical concepts and principles and how they relate to AI. This would be the appropriate point at which to introduce existing sets of ethical guidelines in AI, such as those proposed by the High-Level Expert Group on AI (AI HLEG)[4] of the EC, by IEEE[5] and OECD[6]. These are founded on a central set of ethical

[4] See https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[5] See https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[6] See https://www.oecd.org/going-digital/ai/principles/

concepts and principles, so it is appropriate to teach these and how they have been used in applied ethics. What, for example, is autonomy and why is it important? What is privacy? What kinds of privacy are there, and what justifications are given for its importance?

These concepts and principles will then need to be discussed in the context of AI. This could include a review of how they relate to AI. How, for example, do ethical issues of fairness show up in AI? This could include a coverage of the more general issues of algorithmic bias and of digital divides and it could include fairness issues in various types of applications of AI, for example, in the justice system, in healthcare, in policing and in education. Various additional examples can be presented of cases in which such ethical issues played a role.

Next, after training in the recognition and identification of ethical issues in AI, training in ethical analysis is needed, as well as in the development of solutions to ethical issues. As part of this training, moral debate should be practiced. This step can rest, on the one hand, on standard approaches to ethical analysis in applied ethics, as well as specific analyses and tools of analysis that have been developed in AI ethics.

Finally, students can be taught about the various methods for supporting ethics for AI in practice, such as Ethics by Design methods, research ethics frameworks, ethics standards and certification and others, as well as the responsibilities and roles of various actors.

Applied ethics of AI in particular domains objectives can be met in a similar way, starting with exemplary cases in the domain in question, followed by coverage of ethical concepts and principles (including, possibly, ethical guidelines for AI that have been developed for that particular domain, if any), their application to ethical issues in AI in the domain, and then on to practicing ethical analysis and being introduced to relevant methods and actors.

> *"Professionalism and ethics should be the cornerstone of any curriculum in computer engineering. The focus on design and development makes social context paramount to one's studies in the field."*
>
> - [7]

2.4. Professional ethics objectives

Professional ethics of AI objective for students in AI, computer science and related fields, can be met in the following way. In all forms of applied ethics education, it is often a good idea to start off with exemplary cases. The purpose of the cases here is not to present particular societal ethical issues, but rather to present moral dilemmas that AI professionals may encounter in their work. This can be cases in which decisions have to be made about whether or not to go forward with projects which appear to raise ethical issues. A common method which has emerged for all forms of ethical education in this regard is to ask students to design an AI system and then explore the ethical issues their own design generates. We provide an example of how this is done at Harvard University below.

Next, it will be useful to cover existing professional ethical codes in the field of computer science and AI, and more generally discuss the responsibilities of AI scientists, engineers and computer scientists.

[7] Joint Task Force on Computer Engineering Curricula, *Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering*, 43.

This discussion of professional responsibilities can then also be related to the ethical guidelines for AI that were discussed under the Applied Ethics of AI objective.

A third step is to relate these conceptions of professional responsibility to professional activity in the AI field. What, first of all, are one's professional responsibilities in relation to particular ethical principles (e.g., fairness, privacy, accountability)? What should one do, for example, in the development of AI systems to ensure that they are fair and free from bias? These responsibilities can also be investigated in relation to particular types of techniques and systems (e.g., machine learning, affective computing, drones, decision support systems) and application domains (e.g., healthcare, manufacturing, defence). Cases can be used to help illustrate ethical issues in relation to these principles, techniques, systems and application domains. The objective of these explorations is not just to identify ethical issues that should be faced, but also to discuss what professional responsibility dictates and how professionals should act. How to balance particular values and interests in this process is a particular challenge that should be discussed and practiced.

Some of the recommended topics for ethics courses for AI, computer science and related fields are: ethics of algorithms; Ethics by Design approaches (both learning about them and practicing them in, or in conjunction with, other courses); coding for fairness (including the prevention and mitigation of algorithmic biases and mitigation of digital divides, and with special consideration of risks of discrimination in relation to social categories such as gender, race, and class); privacy issues in AI; user autonomy and freedom and the maintenance of meaningful control in AI; the ethics of transparency and explainability and what standards to use in AI development; issues of responsibility, accountability and liability in relation to AI systems and how to approach them; social and environmental impacts of AI and how to take them into account in development; supporting diversity in AI development and in the AI profession; ethical issues relating to particular types of techniques such as machine learning, automated planning and scheduling, computational ethics, intelligent agents, computer vision, embedded AI, social robots, big data, and unmanned aerial vehicles; ethical issues in different application domains such as healthcare, finance, law enforcement and education.

It is important not to be too prescriptive regarding specific topics at this time. Many of the demands of ethical AI are new, such as full-development process auditability. In such cases there are no established methods by which the developer community can meet these demands. This is not to say they are impossible, merely that best practice, or even common practice, has yet to be established. Similarly, there are no standardised tools by which to meet them. In many cases tools will be bespoke to individual teaching institutions. In some cases the field has yet to achieve agreement over even basic terminology. For example, 'computational ethics' is used to describe computer simulations of individuals with varying ethical values interacting in order to predict social attitudes [8], but also as an alternative term for Ethics by Design [9]. Any curriculum design therefore needs to allow for variations in approaches to individual topics, significant change in course contents over time, and frequent additions or removals as AI systems become more sophisticated and their impact on society becomes better understood.

It is therefore necessary to recognise that individual institutions will vary considerably in their approach to teaching AI ethics, and to encourage this variation. Over time best practice will emerge.

[8] Quigley, *Encyclopedia of Information Ethics and Security*.
[9] Segun, "From Machine Ethics to Computational Ethics".

Similarly, following a universal pattern in the history of technology [10], standardised tools and methodologies for meeting the needs of ethical AI will also emerge. Courses in computer science and engineering will come to incorporate these tools and methodologies as they come into common usage. After a period of years it will become possible to set common core curricula which can be reasonably expected of all such courses, but this is unlikely in less than five to ten years.

The *professional ethics of AI objective for students in other fields* can be met in an analogous way. In these other fields, the emphasis is on the application of AI technology, rather than its development and maintenance. Here, it is also good to start off with exemplary cases that present moral dilemmas in the application of AI that students professionals may encounter in their future profession, such as, moral dilemmas regarding the introduction and use of legal expert systems in the profession of law. Secondly, discussion of professional ethics in the field at issue should take place and should be related to existing ethical guidelines for AI. Finally, the conceptions of professional responsibility should be applied to particular applications of AI that are relevant to one's field of study, with the intent to analyse and discuss how professionals should act in relation to these applications so as to act morally and professionally.

2.5. Teaching staff and teaching methods

AI ethics courses for computer science students should be taught by faculty with significant training in ethics as well as a good understanding of AI technology. Faculty should therefore be computer scientists who have received training in ethics, ethicists who have familiarized themselves thoroughly with AI, or teams of computer scientists and ethicists. In computer science curricula, it is reasonable to expect that ethics of AI will be integrated both into general AI courses intended for computer science students and stand-alone AI ethics courses. An increasing number of students are pursuing PhD's in this area by combining Computer Science and Philosophy. Such students are positioned either in Philosophy departments or in Computer Science departments, with no detectable preference for one or the other, but rather roughly equal numbers in both. This suggests either path is suitable for developing specialists in ethical AI and that such graduates will become an increasingly important cohort of educators for the future should be encouraged.

Regarding teaching methods, it is recommended that a combination of active, passive, individual, and group tasks and assignments are used to bridge the gap between theoretical discussion and concrete cases. On top of attending lectures and seminars, students may be asked to engage critically with extensive (non-fiction and science fiction) readings, actively participate in group discussions, carry out research papers and presentations, analyse (custom) case studies, engage in practical exercises, propose projects and policy recommendations, as well as algorithmic solutions, depending on their background. This combination of methodologies should allow students to achieve a better understanding and retention of the issues at stake, and ultimately incorporate AI ethics into their future careers.

2.6. Case Study – Harvard University: Embedded EthiCS[11]

There is a developing consensus that ethical education needs to be embedded as modules inside existing courses rather than taught as independent streams within a degree course. While there is no mechanism for determining majority sentiment in this respect, it is noteworthy that no organisation

[10] Cardwell, *The Fontana History of Technology*.
[11] https://embeddedethics.seas.harvard.edu/

has disagreed with this position nor advocated anything to the contrary. In particular, this is the position of the US National Academy of Engineering [12] and the ACM/IEEE Joint Task Force on Computing Curricula [13] and the European AI Alliance.

An example of this is offered by Harvard University, which operates a program of ethical education modules within the computer science degree programs using a "distributed pedagogy" approach, in which philosophers are embedded within Computer Science education. The primary aim of the program is to teach students how to think through the ethical and social implications of their work. Harvard's "Embedded EthiCS" pedagogical methodology treats ethical reasoning as integral to computer science education by embedding ethical education inside standard computer science courses across the curriculum. It works by using philosophers to teach ethical modules within courses which explore the ethical issues raised by that particular course's contents. For example, in a data systems course, a philosopher might explore issues of privacy in large, distributed systems. In a machine learning class, the philosopher might explore how solving problems using machine learning can lead to inadvertent discrimination. The modules emphasize "active learning" assignments which teach students to apply the philosophical ideas they learn to real-world ethical problems.

The program is a joint initiative of Harvard University's Computer Science and Philosophy departments. Commencing in 2017 with four modules, it now offers twelve.

The goal of Embedded EthiCS is to teach students to think ethically. This requires an understanding of the social context in which systems operate as well as the ethical issues within a particular system. The Embedded EthiCS program aims to teach students to:

- Identify and anticipate ethical and social issues in their work.
- Think clearly about those issues, both alone and collaboratively.
- Communicate their understanding of those issues effectively.
- Design systems that take into account ethical and social concerns.

The creators of this program hold that embedding ethical education within standard courses achieves better results than independent parallel streams of ethical training, for the following reasons:

- It shows students the extent to which ethical and social issues permeate virtually all areas of computer science.
- It familiarizes students with the wide range of ethical and social issues arising across the field.
- It provides students with repeated practice reasoning through those issues, communicating their positions, and designing systems that take into account what they've learned.[14]

While this program is intended to cover all aspects of ethics relevant to computer science, some modules directly address issues of ethics in AI and can be used to illustrate the way in which a module can be embedded into a traditional computer science curriculum.

[12] Hollander and Arenberg, *Ethics Education and Scientific and Engineering Research: What's Been Learned? What Should Be Done?*

[13] Joint Task Force on Computer Engineering Curricula, *Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering*.

[14] Grosz et al., "Embedded EthiCS".

Module: Discrimination and Machine Learning

This module is taught within *CS 181: Machine Learning*, which is described as an "advanced undergraduate" course. Students are expected to be proficient programmers familiar with probability theory, calculus and linear algebra. The aim of the course is to develop the ability to create real-world AI applications.

The ethical module within this course focuses on the concept of bias and discrimination, teaching the distinction between treatment and impact and the ways in which bias within AI systems can result in discrimination in society. Specific techniques for detecting and removing bias are evaluated, such as the limitations of using statistical comparisons to determine the presence of bias. Such issues are contextualised through teaching the relationship between technical feature and social context, such as the distinction between mathematical and social bias.

The module works through "real-world" examples, in which students design theoretical systems and explore the ethical issues they raise. Assignments are typically short and require a combination of programmatic and philosophical input in which they must explain how their algorithms can be tuned to avoid bias, accompanied by two or three paragraphs of text justifying their approach in terms of ethical and social values. The emphasis is on reasoning skills, with little to no requirement for research.

The authors of the module claim the response has been very positive because

> "students are able to see immediately how the moral issues raised in the module were relevant to concrete, socially important applications of machine learning, as well as how current machine learning researchers are addressing issues of discrimination in current research."[15]

This module is thus an excellent example of how to teach our recommended approach of Ethics by Design. Here ethical considerations are taught as an integral part of the programmer's education. The idea that ethics are something to be done after a system is finished never arises because ethics are taught as one of the criteria by which systems are designed.

However, this module is tightly integrated with the main topic of the course and requires extensive knowledge of the main course material. Developing such highly integrated modules is likely to be unattainable in many universities, especially considering the degree of computational and mathematical knowledge required of the philosophers and pressure on time for course development. However, it may be regarded as an ideal type for integration of ethics into AI courses in Computer Science. By contrast, other modules take a more general philosophical approach. For example, the module "Robots and Work", taught within *CS 189 - Autonomous Robot Systems*, teaches Rawls' theory of justice so as to connect ethical aspects of the workplace with the societal impacts of automation.

Conclusions

The success of Harvard University's Embedded EthiCS program demonstrates the feasibility of integrating ethical education into any higher education course and the positive outcomes which can be expected. We have taken as our example the most extreme integration of philosophical education with another course possible. Modules in law, business, sociology, medicine and other fields not directly concerned with constructing AI do not need such deep integration and we can therefore be

[15] https://embeddedethics.seas.harvard.edu/module-CS181.html

confident they can be introduced much more easily. Even the majority of Computer Science and Engineering courses do not need such deep integration in order to achieve positive benefits. Consequently, Harvard University's Embedded EthiCS program demonstrates the feasibility of integrating education in ethical AI within any existing higher education courses.

3. AI Ethics Education And Training In Industry

Given the range of relevant actors already extant within commercial industry who need to be involved in ethical AI, we do not think a single solution is viable. Our strategy is focused on encouraging industry to adopt our recommendations by making it in their commercial interest to do so. The ultimate aim is to see a self-sustaining business ecosystem which devoted to the development and promotion of ethical AI systems because companies profit from it.

Our strategy here uses only techniques proven to work in other areas. The centre of the strategy is a certification program based on proven policies from relevant areas, such as certification of equipment (such as electrical and aircraft components), industry training and certification programs (such as Microsoft's Certified Professional programs), the EU Energy Labelling scheme, and the established processes by which commercial industry develops certification programs, such as ISACA's COBIT and CGEIT. We recommend integration of our aims into existing programs rather than promoting new ones. Our strategy with certification bodies is to build an environment which motivates them to develop and promote the required training and certification programs themselves.

3.1. The Ethical AI Certification Program

There will be three categories of certification:

- **Systems** – AI and robotics systems will be certified as meeting ethical requirements.
- **People** – People can obtain a range of Ethical AI certifications, as appropriate to their role (e.g.: developer, business manager, educator)
- **Training Programs** – Ethical AI training programs leading to certification will themselves be certified as suitable for the task. Under many certification schemes, the training company must be certified as able to effectively deliver the training program.

For this policy to work, there needs to be a commercial advantage for those who certify their products and staff. This commercial advantage will then generate a desire to seek certification. Since adoption of the certification system produces commercial advantage, AI vendors will then be motivated to maximise the value of their investment in these certifications by promoting their possession of them and the value of them. In other words, industry will have a stake in promoting ethical AI to the public.

Certification Ecosystems

The development of a product certification ecosystem means many different types of stakeholders becoming involved in the product certification process. This will include existing trade bodies, new trade associations united by specific AI functions used in multiple industries, such as a trade association for creators of facial recognition systems, and standards bodies such as IEEE (which is developing the Ethics Certification Program for Autonomous and Intelligent Systems[16]). Significant AI platform providers, such as Google, Microsoft and IBM, will develop their own organised ecosystems of adherents (developers, vendors and the final end-user organisations). We can expect such companies

[16] https://standards.ieee.org/industry-connections/ecpais.html

to develop their own range of certification schemes for ethical AI just as they have done for other technologies. Finally, private certification organisations have already started to develop ethical AI certifications. For example, Certnexus[17] has developed the Certified Ethical Emerging Technologist certification, together with a training program[18].

AI Products

We recommend system certifications based on the model seen in industries where safety is important, such as electrical products, aircraft and medical equipment.[19] Systems must first a *product certification* before they can enter the market. This certifies that the system meets the ethical requirements for an undeployed AI system. Increasingly, AI products sit on back-end platforms which provide the raw AI processing power. The ethical status of such back-end AI platforms will affect, if not determine, the ethical status of application using them. We therefore recommend an *AI platform certification* which focuses on the system's suitability to provide its backend functions ethically. An AI product may not obtain a product certification if it is using a back-end AI platform which does not possess an AI platform certification.

Installed systems will also need *AI installation certificates* before they can be used. This will consider the wider context within which the application is operating as well as the application itself. For example, a facial recognition system may be deployed to control access to a building. The installation certification would include checks that that appropriate signage is displayed to inform people that they are being subject to AI surveillance.

Auditing

A system's ethical status may change over time as it learns and acquires new data. On-going compliance must therefore be regularly assessed through auditing. Audit procedures will need to be in accord with the product's certification. Auditing will need to be regular. We do not recommend specific approaches to audit. Other industries, such as aircraft components, operate effectively while allowing for considerable national variations in approach [20].

People

People are to be certified through professional training programs and exams. Existing AI engineer certificates are already available, but need ethical components added. Purely ethically-focused certification programs are also emerging, such as Certified Ethical Emerging Technologist[21] from Certnexus. There are many existing certifications which should incorporate ethical AI components, such as COBIT.[22]

A range of certifications will be required, as is the case for most technical systems. For example, while developers need a detailed understanding of the coding decisions which can lead to ethical issues, the

[17] https://certnexus.com/about-certnexus/
[18] https://certnexus.com/certification/ceet/
[19] The Danish government is in the process of developing a "responsible data use" seal for AI product certification, but this program is in early stages and no details are available for us to model our proposals on. See https://dataethics.eu/danish-companies-behind-seal-for-digital-responsibility/
[20] Leveson, "The Use of Safety Cases in Certification and Regulation".
[21] https://certnexus.com/certification/ceet/
[22] https://www.isaca.org/credentialing/cobit

senior managers of an organisation using that system need to understand how their organisation's way of using it can affect its ethical status once operational.

Ethical AI requirements should be incorporated into the EU e-Competence framework[23] because most European certification bodies use this as a standard from which to draw requirements.

Professional Associations and CPD

Professional associations typically require continuing professional development programs (CPDs). They should be encouraged to include requirements for appropriate CPD training in ethical AI. Professional associations often have codes of conduct which should be kept in line with the certification standards as they evolve, as well as relevant audit or CPD requirements. A valuable channel for communication with IT professionals is the Council of European Professional Informatics Societies (CEPIS)[24] which represents the thirty-five IT professional associations across Europe. Many European professional associations draw guidance from the European e-Competence Framework[25], so this is another reason to update it to include ethical AI.

SME's and Business Awareness

It is important to bear in mind the majority of AI purchasers will be small businesses. AI systems are not necessarily large or expensive. 99.8% of all businesses in the EU are small and medium-sized enterprises (SME's) [26]. We therefore recommend building ethical AI awareness through the channels which SME's already use to acquire new expertise - trade associations and business support networks. SME education should focus on deployment certification and issues relevant to the purchase and operation of AI systems. This awareness strategy must promote the value of the certification schemes. Purchase of an AI possessing an ethical AI product certification should be presented as resolving many difficulties which would otherwise fall to the SME.

3.2. Commercial industry motivation

The strategy for motivating industry is to make it profitable to produce ethical AI products by building market demand for them.

The primary aim is to plant two key perspectives in the market:

- An awareness of the *need* for ethical AI certification. This needs to be accompanied by awareness that such certification exists, and that it answers the perceived need.
- *Confidence* that certified products, services and staff are easily available, that selecting a certified product or person is no more difficult than selecting an uncertified one, and that certified products are just as good as uncertified ones, if not better.

A central ethical reference model is required in order to provide a common set of ethical criteria by which to ensure a minimum standard which all certifications must achieve in order to be valid. This will need to be updated periodically. This requires some form of ongoing organisation and associated review processes.

[23] https://www.ecompetences.eu/
[24] https://cepis.org/
[25] Some examples can be seen at https://www.ecompetences.eu/professional-bodies-trade-unions-and-sector-associations/
[26] Executive Agency for Small and Medium-sized Enterprises, *Annual Report on European SMEs 2018/2019*.

Insurance pressure

At some stage legal action is likely against the operator of an AI regarding some ethical aspect of its behaviour, such as racial or gender bias. These matters therefore alter the liability of AI operators. Adoption of an AI which has been formally certified for compliance significantly lowers the risk for an insurer and should therefore affect insurance premiums. We therefore recommend the inclusion of the insurance industry in the development of central standards. We also recommend insurance companies be considered a prime channel for development of market need. In particular, we recommend direct involvement by Insurance Europe[27] in setting ethical certification standards and training requirements for AI.

Procurement

Public procurement is an effective method of influencing commercial product innovation [28]. We recommend that the EU move towards requiring product and installation certifications for all AI's as part of its procurement requirements, such that products may not be included in tenders unless they possess ethical AI product certifications and may not be activated once purchased until they have achieved an installation certificate.

3.3. Certificate Badge Branding

The final element required to create a market which demands ethical AI's is widespread public awareness of certificates and what they offer. In this respect, our model is based on the EU's Energy Labelling Framework and its associated energy labels, such as the Electrical Product Labelling Scheme. We propose a similar scheme. Under this system, approved certifications would generate a standardised ethical AI label. Using similar mechanisms as have been used with the various energy labels, purchasers and users of systems can learn to read such labels. This would enable them to quickly assess the ethical status of a product at a glance. The Danish government is already developing such a badge[29]. We believe it is important this work be conducted at EU level so as to avoid the rise of competing, possibly incompatible, national standards.

The Energy Labelling Framework has been very successful and there is good evidence it is now an active consideration when people make purchases. We believe much of what has been done with the Energy Labelling Framework could be emulated here. The framework has been running for long enough to know what works, and so we suggest simply copying that.

4. Public Awareness Raising About Ethical issues in AI

While more and more industries are investing in AI, popular culture promotes an extremely limited understanding of the issues connected with the deployment of AI technologies, both from a technical perspective and regarding their social and ethical implications. This has resulted in the general public taking little to no part in discussions surrounding AI technologies, which are becoming integral to an ever-increasing number of societal functions.

While AI ethics research and teaching has been centred around universities for the most part, as well as in the technology sector, the general public should also become familiar with social and ethical

[27] https://www.insuranceeurope.eu/. Its members are the EU's national insurance associations and it represents 95% of total European insurance activity.
[28] Dalpé, "Effects of Government Procurement on Industrial Innovation".
[29] https://investindk.com/insights/denmark-paves-the-way-for-implementation-of-trust-by-design

issues in AI, as its role in and impact on society is increasingly significant. A failure of the public to comprehend and engage with social and ethical issues in AI could result, on the one hand, in distrust in AI and therefore a lack of uptake or even active resistance. This could have negative economic effects and a loss of some of the benefits that AI could provide to society. On the other hand, it could lead to misplaced trust in AI, and social and moral harms that result from ignorance in the deployment and overuse of AI.

So how can we attain public awareness about the social and ethical issues in AI, public awareness that makes members of the public educated citizens that allows them to hold informed viewpoints and make informed (political) decisions regarding AI and its role in society? Which actors are responsible for generating such awareness and how can they do it?

Some of the potential actors include the media, civil society organisations, AI industry, universities, and governments. Each will have their part to play. The media obviously have a central role in informing the public and in organizing events and debates in which people can participate. Yet, journalists may not always be properly informed themselves for them to take on this role. How can they be better supported in this task?

Civil society organisations also should play a role, especially civil organisations which are focused on issues related to AI, such as professional organisations in AI and computer science, and in other areas that are heavily impacted by AI, as well as human rights organisations, labour organisations and consumer organisations. Universities have a significant role because they are at the forefront of knowledge, both in AI and in ethics of AI. Finally, governments also need to recognize the importance of an educated public and stimulate public awareness raising campaigns where possible.

People with non-technical or non-academic backgrounds should also be given the chance to explore key issues in AI ethics through accessible in-person and distance-learning courses which combine theory with practical exercises by looking at real-world controversies on the topic. Universities and educational institutions could consider making massive open online courses available to the general public which provide accessible learning opportunities.

The European e-Competence Framework needs to be updated to include ethical AI systems. Other frameworks for digital competence may also need updating. Since the Council of European Professional Informatics Societies (CEPIS) draws much of its codes of conduct from the e-Competence Framework, they should also be directly involved in the development of a set of formal specifications which all codes of conduct should include.

5. Glossary of terms

Term	Explanation
AI Platform	A back-end system which offers AI capabilities which other developers can use to build AI applications
AI Platform provider	A company offering AI platforms to developers, such as Clairifai, IBM and Google.
Accountability	Accountability applies to both individuals and institutions. It means taking responsibility for your actions rather than trying to shift responsibility (or blame) elsewhere. This involves being able to explain the reasons behind your actions when necessary, and being prepared to discuss your actions and their consequences. It implies a willingness to accept and act on criticism of your actions where that is justified.
Auditability	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enabling the system's auditability.
Autonomy	Autonomy is the ability to decide courses of action independently of a ruling body. In AI, a machine or vehicle is referred to as autonomous if it doesn't require input from a human operator to function properly. However, ethical AI is more concerned with human autonomy, of which there are three types. Moral autonomy refers to the innate capacity of humans to determine for themselves what is morally good and bad. Political autonomy refers to the capacity of human beings to form their own political opinions. Personal autonomy refers to the innate capacity of human beings to decide how they should live, especially by what values they should make their decisions.
Bias	Bias is an unfair or unjustified prejudice towards or against a person, group of people, object, or position. Bias can arise in many ways in AI systems. It does not necessarily relate to human bias or human-driven data collection. It can arise, for example, through the limited contexts in which a system in used, in which case there is no opportunity to generalise it to other contexts. Bias can be intentional or unintentional, but is a danger because it frequently causes discriminatory and/or unfair outcomes in AI systems
Discrimination	The act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong. Principles of non-discrimination state that in regard to human rights, there should not be any differentiation that is based on inalienable parts of one's identity, including gender, race, age, sexual orientation, national origin, religion, income, property, health, disability and opinions.
Diversity	Diversity is the inclusion of different types of people, based on identity markers like gender, race, age, cultural heritage, ability, educational background, cognitive style and the like. The principle of respect for diversity goes beyond the principle of non-discrimination to include positive valuation of individual differences, recognition of differences in individual need and support for the diverse composition of organisations and communities. Applied in an AI context, respecting diversity accounting

	for designing for diversity in the composition of data sets that represent people, in user-centred design, in the inclusion of stakeholders and stakeholder perspectives, and in the composition of design teams.
Ethics	Ethics is an academic discipline which is a subfield of philosophy. Applied ethics deals with real-life situations, where decisions have to be made under time pressure, and often limited rationality. AI Ethics is generally viewed as an example of applied ethics and focuses on the issues raised by the design, development, implementation and use of AI.
Ethics assessment	The assessment, evaluation, review, appraisal or valuation of plans, practices, products and uses of research and innovation that makes use of ethical principles or criteria.
Ethical AI	Ethical AI refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values.
Ethical impact assessment	An approach for judging the ethical impacts of research and innovation activities, outcomes and technologies that incorporates both the means for a contextual identification and evaluation of these ethical impacts and the development of a set of guidelines or recommendations for remedial actions aimed at mitigating ethical risks and enhancing ethical benefits, typically in consultation with stakeholders.
Ethical requisite	An ethical requisite is a requirement relating to ethical aspects of the system and the development thereof. Ethical requisites must be met in order to be compliant with the demands for responsible, trustworthy, ethical AI.
Ethics by Design	The approach of incorporating ethical considerations throughout the design, development and deployment phases of software and engineering product creation so as to avoid the product generating negative ethical effects.
Explainability	Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
Informed consent	Permission freely given and granted in full knowledge of the possible consequences. Informed consent must be appropriately documented, based on written or otherwise documentable records stemming from a person capable of giving consent or, where the person is not capable of giving consent, by his or her legal representative.
Oversight	The ability to oversee, supervise, and watch carefully over something – in this context, to oversee the functionality and output of AI systems.
Personal data	Information relating to an identified or identifiable natural person, directly or indirectly, by reference to one or more elements specific to that person. Among these, special categories of data within the meaning of the General Data Protection Regulation concern personal data relating to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, as well as genetic data, biometric data, data concerning health or concerning sex life or sexual orientation.
Personal data processing	Any operation or set of operations performed or not using automated processes and applied to personal data or sets of data, such as collection, recording, organisation, structuring, storage, adaptation or modification, retrieval, consultation, use, communication by transmission,

	dissemination or any other form of making available, linking or interconnection, limitation, erasure or destruction.
Privacy by design	Privacy by Design is an approach taken when creating new technologies and systems. Privacy by Design encompasses IT systems, business practices and physical design. The approach is characterized by proactive anticipation of privacy invasive events so as to prevent them from occurring, rather than fixing them afterwards.[30]
Profiling	According to Article 4(4) of the GDPR, 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements
Pseudonymisation	According to Article 4 of GDPR, 'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person
Reproducibility	Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions.
Stakeholders	All those that research develop, design, deploy or use AI, as well as those that are (directly or indirectly) affected by AI – including but not limited to companies, organisations, researchers, public services, institutions, civil society organisations, governments, regulators, social partners, individuals, citizens, workers and consumers.
Traceability	Traceability of an AI system refers to the capability to keep track of the system's data, development and deployment processes, typically by means of documented recorded identification.

Table 2: Glossary of terms

[30] Cavoukian, *Privacy by Design: The 7 Foundational Principles*.

6. References

Cardwell, Donald, *The Fontana History of Technology*, Fontana, London; New York, 1994.

Cavoukian, Ann, *Privacy by Design: The 7 Foundational Principles*, Information and Privacy Commissioner of Ontario, Toronto, 2009.

Dafoe, Allan, 'On Technological Determinism: A Typology, Scope Conditions, and a Mechanism', *Science, Technology, & Human Values*, Vol. 40, No. 6, November 2015, pp. 1047–1076.

Dalpé, Robert, 'Effects of Government Procurement on Industrial Innovation', *Technology in Society*, Vol. 16, No. 1, January 1994, pp. 65–83.

Executive Agency for Small and Medium-sized Enterprises, *Annual Report on European SMES 2018/2019*, European Commission, Brussels, 2019.

Grosz, Barbara J., David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo, "Embedded EthiCS: Integrating Ethics across CS Education", *Communications of the ACM*, Vol. 62, No. 8, July 24, 2019, pp. 54–61.

Hollander, R., and C.R. Arenberg, *Ethics Education and Scientific and Engineering Research: What's Been Learned? What Should Be Done?*, National Academy of Sciences, Washington, DC, 2009.

Joint Task Force on Computer Engineering Curricula, *Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering*, *Computing Curricula*, Association for Computing Machinery, New York, 2016.

Leveson, Nancy G, "The Use of Safety Cases in Certification and Regulation", 2011.

Quigley, Marian, ed., *Encyclopedia of Information Ethics and Security:*, IGI Global, 2007. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59140-987-8.

Segun, Samuel T., "From Machine Ethics to Computational Ethics", *AI & SOCIETY*, June 29, 2020. http://link.springer.com/10.1007/s00146-020-01010-1.

Annex 6 – Ethics as Attention to Context: Recommendations for AI Ethics

Ethics as Attention to Context: Recommendations for AI Ethics

Annex 6 to D5.4: Multi-Stakeholder Strategy and Tools for Ethical AI and Robotics

[WP5 – The consortium's proposals]

Lead contributor	Anais Resseguier, Trilateral Research
	anais.resseguier@trilateralresearch.com
Other contributors	Rowena Rodrigues, Trilateral Research Nicole Santiago, Trilateral Research
Reviewer	Stearns Broadhead, University of Graz Zachary Goldberg, Trilateral Research

Date	February 2021
Type	Report (Annex 6 to D5.4 deliverable)
Dissemination level	PU = Public

Keywords	AI ethics; ethics as attention; practical ethics; ethics of machine learning; impacts of AI; multi-stakeholder strategy; ethical tools; robot ethics

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This document shows that current AI ethics guidance and initiatives tend to be dominated by a principled approach to ethics. Although this brings value to the field, it also entails some risks, especially in relation to the abstraction of this form of ethics that makes it poorly equipped to engage with and address deep socio-political issues and practical impacts. As such, this document seeks to complement the existing principled approach to ethics with an approach to ethics as attention to context and relations. It introduces this proposal and makes practical recommendations to promote ethical AI by drawing from an approach to ethics as attention to context.

Table of contents

1. Introduction

The ethics of Artificial Intelligence (AI) has generated high interest over the last few years.[1] The numerous ethics guidelines and other forms of ethics initiatives produced provide ample evidence of this.[2] Although the field has seen major developments in various arenas (including in academia, industry, and policy), it has also come under intense criticisms for being too abstract and high-level, and therefore, unable to properly guide technological development, deployment and use.[3] Critics have highlighted that some of these initiatives lead to ethics washing[4] and/or contribute to reproducing structural inequalities in the society.[5]

Although this document recognises the value of what the ethics of AI has produced over the last few years, it also acknowledges that efforts in this area are still needed, especially *to move beyond the abstraction of current AI ethics initiatives*. More recently, there have been numerous efforts at making ethics guidelines more fit for purpose by operationalising them to specific sectors of application or development contexts.[6] These are assuredly necessary and praiseworthy efforts.

Several experts on AI have critiqued AI ethics for failing to consider contextual elements, especially the socio-political context, which, as they note, is essential to address ethical challenges posed by this technology.[7] The present document shows that this identified gap in AI ethics finds its root in the very nature of the currently dominant approach to AI ethics, i.e., a principled approach, *a view on ethics that considers it as a softer version of the law*. It points to the need to complement this approach and

[1] The authors of this report acknowledge the input of various experts and stakeholders to this text. Please see the Acknowledgement section of SIENNA D5.4 (Feb 2021) for a list of these people. Special thanks to Zachary Goldberg (Trilateral Research), Ana Valdivia (King's College London), and Charalampia (Xaroula) Kerasidou (Lancaster University). This text also includes in the concept of Artificial Intelligence systems that have a physical component, i.e., AI-powered robotics.

[2] The following references give a listing of these: Jobin, Anna, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines", *Nature Machine Intelligence* 1, no. 9, 2019, pp. 389–99; Hagendorff, Thilo. "The Ethics of AI Ethics. An Evaluation of Guidelines", *Minds and Machines*, no. 30, 2019, pp. 99–120; Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, and Madhulika Srikumar, "Principled Artificial Intellience: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI", Cambridge, MA: Berkman Klein Center for Internet & Society at Harvard University, January 2020, https://cyber.harvard.edu/publication/2020/principled-ai. The organisation Algorithm Watch maintains a Global Inventory of AI Ethics Guidelines and had 160 items as of April 2020: https://inventory.algorithmwatch.org.

[3] Mittelstadt, Brent, "Principles Alone Cannot Guarantee Ethical AI", *Nature Machine Intelligence 1*, Nov 2019, pp. 501-507.

[4] Ethics washing corresponds to the use of ethics to avoid strict legal regulation. See in particular Wagner, B., "Ethics as an escape from regulation: From ethics-washing to ethics-shopping" in Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, Amsterdam University Press, Amsterdam, 2019, pp. 84–89.

[5] D'Ignazio, Catherine, and Klein, Lauren F., *Data Feminism*, MIT Press, Cambridge, MA; London, England, 2020.

[6] See for instance Annex 2 D5.4 proposing an "Ethics by Design and Ethics of Use in AI and Robotcs". See also Brey, Philip, Björn Lundgren, Kevin Macnish, and Mark Ryan, "Guidelines for the Ethical Use of AI and Big Data Systems", SHERPA project, July 2019 and the High-Level Expert Group on AI, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", European Commission, Brussels, July 2020, https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[7] Gebru, Timnit, "Race and Gender", in Markus D. Dubber, Frank Pasquale, and Sunit Das, *The Oxford Handbook of Ethics of AI*, Oxford, Oxford University Press, 2020.

makes a series of practical recommendations. As such, it calls for *a shift of attention in AI ethics: away from high-level abstract principles to concrete practice, context and social, political, environmental materialities.* It draws extensively from critical approaches to AI and ethics, especially those emerging from feminist perspectives (including the ethics of care). It also derives from consultation with a number of stakeholders as part of SIENNA engagement.[8]

This document first undertakes a theoretical detour to better understand the nature of the ethics approach at stake in current AI ethics (section 2) and how to best complement it (section 3). This detour is necessary to build the conceptional groundwork for the practical recommendations developed in section 4, which contains the actual proposals to promote ethical AI. This document, and the guidance it offers, are addressed to policymakers in government, AI ethicists, engineers (software engineers, data scientists, etc.), organisations developing AI, and, more generally, anyone concerned by the development, deployment and use of AI and the potential social and ethical impacts of this technology.

2. Current AI ethics and the risk of abstraction

2.1 The principled approach to ethics in AI ethics

The numerous AI ethics initiatives, documents and guidelines produced over the past few years have primarily taken the shape of high-level abstract and prescriptive principles, such as "Ethics guidelines for trustworthy AI" of the High-Level Expert Group on AI (AI HLEG) set up by the European Commission, the "Recommendation of the Council on Artificial Intelligence" by the OECD, or on the industry side, the Google AI principles.[9] However, as Resseguier and Rodrigues have argued, these AI ethics initiatives are dominated by a "law-conception of ethics", which is a view on ethics that considers it as a *replica* of the law, i.e., a softer version of the law.[10] Resseguier and Rodrigues point to the risk of misusing ethics as a replacement for legal regulation, and in particular the risk of ethics washing. While legal regulation might come with hard lines that could restrict innovation, regulation through ethical principles and guidelines offer more flexibility and leeway; hence, they are favoured by industry actors.[11] An issue with this approach to ethics is that it leads to ethics being used as a weaker form of regulation by actors with interest in avoiding hard lines.[12] Additionally, this approach to ethics comes

[8] In particular, an earlier version of this piece was presented at the SIENNA online Workshop on Multi-Stakeholder Strategies for Ethical AI on 9 September 2020. It was entitled "Critical Insights from the Social Sciences and Humanities for the Ethics of AI". This piece includes feedback received by stakeholders on this occasion as well as during the EUREC/SIENNA online workshop on guidance documents for Research Ethics Committees on 26 and 27 October 2020.

[9] Respectively: High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI", European Commission, Brussels, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai; OECD, "Recommendation of the Council on Artificial Intelligence", adopted on 22 May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; Google, Artificial Intelligence at Google: Our Principles, https://ai.google/principles/. For listings of AI ethics guidelines and other guidance documents, see references in footnote 1.

[10] Anscombe, Gertrude, E., M., "Modern moral philosophy", *Philosophy*, Vol. 33, Issue 124, 1958, pp. 1-19; Resseguier, Anais, and Rodrigues, Rowena, "*AI ethics should not remain toothless?* A call to bring back the teeth of ethics", *Big Data & Society*, July-Dec 2020, pp. 1-5.

[11] Benkler, Yochai, "Don't Let Industry Write the Rules for AI", *Nature* 569, 2019, p. 161.

[12] In addition, Mittelstadt points to issues of ineffectiveness the approach of ethics through codes and lists of principles in Mittelstadt, Brent, "Principles Alone Cannot Guarantee Ethical AI", *op. cit.*, p. 504

with another pitfall that needs to be addressed: *its abstraction and the risk of disconnection from social, political and environmental materialities*.

The ethical theory of the "ethics of care" or "care ethics" that emerged in the 1980s with the work of Carol Gilligan has identified a fundamental gap in this "law conception of ethics", a conception that the ethics of care has called "ethics of justice", "principlism", or "principled approach": its neglect of context and actual practices, i.e., its abstraction.[13] This neglect is not simply a side-effect of this approach; *it is one of its constitutive features*. Indeed, as the ethics of care shows, the principled approach is primarily and fundamentally characterised by its gesture of abstraction from concrete situations. It is through this abstraction that it can develop the general and impartial principles that it relies on and seeks to promote. This is particularly clear in the "veil of ignorance" promoted in the ethics of John Rawls. This veil aims at hiding any elements that constitute a person's specific socio-political situation in the world (including race, gender, socio-economic status, nationality, etc.) in order to formulate judgements from an unbiased and impartial standpoint, what Thomas Nagel has called the "view from nowhere".[14] Although this approach to ethics has its value, its abstraction also brings with it several blind spots that are particularly problematical in the context of AI, even more so when it is applied to a field such as AI which presents itself as supposedly immaterial.

2.2 AI immaterial narrative and AI ethics

There is a dominant narrative surrounding digital technologies that presents these technologies as intangible. The supposed dematerialisation of technology is a narrative that has existed before digitalisation; however, it has become particularly pervasive with digitalisation, and especially AI.[15] A telling example of this is the concept of the "cloud" that makes data storage appear as deprived of a physical existence and, therefore, of practical impacts on the society and the environment. However, as we know, this is far from being true. The "cloud" relies on giant data centres that consume massive amount of energy. Similarly, studies have shown that the training of AI systems requires a high level of energy consumption as well as resource extraction that have significant environmental costs.[16]

The situation of micro-workers in the AI ecosystem today provides another clear case that undermines the AI "immaterial narrative". Micro-workers are people who work for platforms, such as Amazon's Mechanical Turk, to prepare, label and verify the data that go into AI systems.[17] The AI industry tends to hide the situation of these workers, pretending that all the hard work is fully automatised. However, this is fallacious: AI systems rely on the work of people around the world that are extremely poorly

[13] Gilligan, Carol, *In a Different Voice: Psychological Theory and Women's Development*, Harvard University Press, Cambridge, MA, 1982.
[14] Nagel, Thomas, *The View from Nowhere*, Oxford, Oxford University Process, 1989.
[15] Izoard, Celia, "Les Réalités Occultées Du 'progrès' Technique : Inégalités et Désastres Socio-Écologiques", *Ritimo*, no. 21, May 2020, pp. 27–33. See also Campolo, Alexender, and Kate Crawford, "Enchanted Determinism: Power without Responsibility in Artificial Intelligence", *Engaging Science, Technology, and Society*, Vol. 6, 2020, pp. 1–19. In this article, Campolo and Crawford point to a discourse surrounding deep learning systems as one of "exceptional, enchanted, otherworldly and superhuman intelligence" (p. 9).
[16] On the environmental impacts of AI: Strubell, Emma, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP", *arxiv.org*, 2019. https://arxiv.org/abs/1906.02243; Crawford, Kate, and Joler, Vladan, "Anatomy of an AI System: The Amazon Echo as An Anatomical Map of Human Labor, Data and Planetary Resources", AI Now Institute, September 2018.
[17] https://www.mturk.com

paid and are working with no social security protection.[18] Here as well, AI "immaterial narrative" hides highly problematical impacts on the society, in this case those related to working conditions and employment rights.

Hence, this "immaterial narrative" that serves to pretend that AI, because intangible, is deprived of questionable social or environmental impacts, is highly problematic. However, because of its neglect of situatedness and materialities to reach high-level abstract principles, the approach dominant in AI ethics today renders it ineffective at properly addressing these negative material implications of AI. In other words, the disconnected nature of AI ethics today does not allow an appropriate response to some of the key critical issues that AI raises for society and individuals. It is therefore necessary to complement it with methods to ensure AI ethics turns to concrete practices, considers the socio-political context and materialities, and addresses impacts.

2.3 AI ethics and the neglect of structural inequalities

This need for AI ethics to engage with the socio-political reality is even more necessary to address a major issue of AI: the risk of this technology further entrenching already existing structural inequalities. As Klein and D'Ignazio have shown in *Data Feminism,* AI ethics at present appears inadequate to properly respond to this challenge. According to them, data ethics is a field that relies on "concepts that secure power", and that, as such, "maintain the current structure of power".[19] This is highly problematic as it renders AI ethics unable to address the root causes of one of the most significant ethical issues of AI: biases and social inequalities, especially those pertaining to race and gender. Studies have shown that the most vulnerable populations are those who face the most problematic impacts of AI (and often who are subject to the deployment of AI technology without choice or ability to influence its design and development), while, on the contrary, those who benefit the most from AI are those who hold positions of power in relation to this technology.[20] For instance, the study "Gender Shades" shows this clearly, demonstrating the problematic consequences of producing facial recognition systems trained with white males as the norm.[21] The consequence is a technology that works best for white males, but poorly for black females, white females and black males lying in between. Here as well, the "point of view of nowhere" that characterises AI and AI ethics fails to offer a proper response to key issues in the field of AI, particularly those related to structural inequalities and injustice.[22] Hence, AI ethics needs to be complemented in a way that enables it to address and respond to these issues.

[18] Tubaro, Paola, Antonio Casilli, and Marion Coville, "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence", *Big Data & Society*, Jan-June 2020, pp. 1–12.
[19] D'Ignazio, C. and Klein, L., op. cit., 2019, p. 60 and p. 61.
[20] See for instance Jansen, Philip, Philip Brey, Alice Fox, Jonne Maas, Bradley Hillas, Nils Wagner, Patrick Smith, Isaac Oluoch, Laura Lamers, Hero van Gein, Anais Resseguier, Rowena Rodrigues, David Wright, David Douglas, Ethical Analysis of AI and Robotics Technologies, SIENNA D4.4, Aug 2019.
[21] Buolamwini, Joy, and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", in *Conference on Fairness, Accountability, and Transparency*, 81, New York: PLMR, 2018, pp. 1-15. See also: Birhane, Abeba, and Cummins, Fred, "Algorithmic Injustices: Towards a Relational Ethics", arXiv: 1912.07376, 2019.
[22] In "Race and Gender", Gebru condemns this "point of view of nowhere" in AI and its negative impacts on the most vulnerable and marginalised populations. Gebru, T., op. cit., 2020.

There is an interesting parallel to be drawn between (1) the criticism toward AI ethics as being elaborated from a perspective of the privileged members of the society and (2) the criticism formulated by the ethics of care toward what it identified as the dominant form of ethics (the ethics of justice). On the side of the critique toward AI ethics, Klein and D'Ignazio point to the "privilege hazard" which they define as "the phenomenon that makes those who occupy the most privileged positions among us—those with good educations, respected credentials, and professional accolades—so poorly equipped to recognize instances of oppression in the world", i.e., an "ignorance of being on top".[23] The ethics of care highlights a similar situation characterised by the "indifference" of those who are privileged, i.e., those who occupy positions of power.[24]

Thus, the dominant form of ethics in current AI ethics is *primarily a principled one characterised by abstraction*. This approach to ethics makes AI ethics *poorly equipped to respond to some of the key issues of AI related to social, political and environmental materialities*. Therefore, the ethics of AI as it has been developed over the past few years *needs to be complemented by other means to ensure proper consideration of actual practices, situations and socio-political context*.

3. For an ethics of attention to context, situatedness and materialities

The second section pointed to one of the key risks of the dominant understanding of ethics in current AI ethics discourses and initiatives: its abstraction leading to a potential neglect of problematical social, political and environmental impacts of AI. The third section presents how this approach can be best complemented to respond to the identified limitation. It does so by promoting a different approach to ethics, one that *invites to a sharp attention to context, situatedness and materialities*. This section also responds to feedback from SIENNA's stakeholders and other AI experts who have repeatedly called for making AI ethics more practical and usable for organisations and developers.[25]

3.1 Ethics as attention in situation

This section argues for the *need for AI ethics to shift attention away from high level abstract and prescriptive principles to practical contexts and relations*. This shift is one of the lessons learned in bioethics, a field that was at first heavily dominated by high level principles, such as in the famous *Principles of Bioethics* by Beauchamp and Childress that derive answers to ethics challenges from four basic principles (autonomy, beneficence, non-maleficence, and justice). This approach has been challenged for being too abstract, top-down and insufficiently attentive to particulars.[26] Aren't we reproducing the same issue in AI ethics today as bioethics in the 1980s with the dominating principled approach (or "principlism")? As indicated by a participant to the SIENNA Workshop on Multi-

[23] D'Ignazio, C. and Klein, L., op. cit., 2019, p. 29 and p. 28 (respectively).
[24] Molinier, Pascale, "De la civilisation du travail à la société du Care", *Vie Sociale* 14, no. 2, 2016, p. 138.
[25] Workshop on the analysis of present and future ethical issues in AI and robotics, Uppsala (Sweden), 13-14 June 2019 and Online workshop on Multi-stakeholder Strategies for Ethical AI, 8-9 September 2020.
[26] Arras, John, D., "The Way we Reason Now: Reflective Equilibrium in Bioethics", in Bonnie Steinbock, *The Oxford Handbook of Bioethics*, Oxford University Press, Oxford, 2009.

stakeholder strategies for ethical AI (Sept 2020), AI ethics should draw from lessons learned in bioethics.[27]

The ethics of care has developed resources to move beyond the principled approach to ethics. It has called for ethics to "modify its field: from an enquiry into general concepts to the study of particular situations, individual's moral configurations."[28] It moves away from the "view from nowhere" that characterises the principled approach and invites to a sharp attention to concrete practical reality, situations and relations.

This attention to specific situations is more generally a key contribution from feminist theory. As Klein and D'Ignazio put it: "one of the central tenets of feminist thinking is that all knowledge is situated".[29] The recognition of the situatedness of knowledge is a particularly essential recognition for the fields of AI and data science, fields that are often "framed as an abstract and technical pursuit" and that, as such, leave aside the "social context, ethics, values, or politics of data."[30] Data science has the tendency to pretend it is always the same regardless of its context of application, whether it is astrophysics, criminal justice or carbon emission.[31] This is deceptive and the ethics of AI needs to be able to challenge and address this fallacious claim. To do so, it needs to advocate for attention to context and social, political, and environmental materialities.

For instance, when dealing with criminal data, it is essential for AI ethics to recall the historical structures of injustices and inequalities through which these data have been produced.[32] Similarly, when handling health data, one should have in mind that white males are significantly over-represented in such sets of data, leading to the risk of poor quality healthcare for women and non-white population.[33] The case of the machine learning technique of words embeddings is another telling example of the need to consider the social background, as Bolukbasi et al. show clearly in their article "Man is to Computer Programmer as Woman is to Homemaker?"[34]

Ethics defined as attention means attention to socio-political realities and materialities at, at least, two different levels: (1) that of the development of an AI system (both from the production of the algorithm and the datasets that were used for the training of the algorithm) and (2) that of the context of

[27] For an excellent comparative analysis of AI ethics versus medical ethics, see Mittelstadt, Brent, op. cit., 2019.
[28] Molinier, Pascale, Sandra Laugier, and Patricia Paperman, *Qu'est-ce que le care? Souci des autres, sensibilité, responsabilité*, Payot & Rivages, Paris, 2009, p. 23. Authors' translation.
[29] D'Ignazio, C. and Klein, L., op. cit., 2020, p. 152.
[30] *Ibid.*, p. 66.
[31] *Ibid.* Similarly, Campolo and Crawford have pointed to the "epistemological 'flattening' of complex social contexts into clean 'signal' for the purposes of prediction" that machine learning brings about. Campolo and Crawford, op. cit., 2020, p. 10.
[32] For a famous study of how AI systems used in the criminal systems have led to further entrenching already existing structures of inequalities in the US context, see: Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias", *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[33] Hart, Robert David, "If You're Not a White Male, Artificial Intelligence's Use in Healthcare Could Be Dangerous," *QZ*, July 10, 2017. https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use- in-healthcare-could-be-dangerous/. For a general study on the poor representation of women in data see: Criado Perez, Caroline, *Invisible Women. Exposing Data Bias in a World Designed for Men*, London, Chatto & Windus, 2019.
[34] Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings", ArXiv.Org, 2016. Word embedding is a machine learning technique that represents text data as vectors.

application and intended use. The proposal formulated by Gebru et al. documenting the "motivation, composition, collection process, recommended uses, and so on" of a database used for machine learning would be particularly useful in that regard.[35] Another expression of ethics as attention to context is developed by Asaro in an article in which he calls for an ethics of care approach to AI ethics in predicting policing. He highlights the need for AI ethics "to seek likely ways in which political, economic, or social pressures may have influenced historical datasets, to consider how it [sic] may be shaping current data collection practices, and to be sensitive to the ways in which new data practices may transform social practices and how that relates to the communities and individuals a system aims to care for."[36] To do so, he recommends the inclusion of "domain experts as well as critical social scientists as members of design teams" and the recognition of the "necessity of their expertise in shaping ultimate system design." [37]

3.2 Ethics as attention to relations of power

Another crucial insight from the ethics of care that AI ethics can draw on is the need to pay closer attention to *relations of power and inequalities*. When the ethics of care argues that ethics is a matter of attention to situations, this includes relations. It is important to note that these relations do not only refer to interpersonal relationships, although these are at stake as well. It also refers to broader relations in the society, including relations of power, i.e., power asymmetries between different social groups. This is one of the key contributions of the ethics of care to ethical theory. It is also a call for realism. As Laugier puts it: "Care takes us back to this requirement of realism in the sense of the need to see what is in front of our eyes: the reality of inequality before the idealness of principles."[38] This is particularly essential in the context of AI considering the high economic interests involved and how these come to shape policy agenda on the regulation of AI.

Several experts have shown that the dominant form of ethics has failed to pay sufficient attention to the power imbalance at stake in the discussions on the regulation of AI, especially with regards to the concentration of power in the hands of a few big tech companies. For instance, Wagner shows how ethics has been used "as an escape from regulation".[39] Article 19, a non-governmental organisation, argues that ethics initiatives have often "proven to be a strategy of simply buying time to profit from and experiment on societies and people, in dangerous and irreversible ways."[40]

---

[35] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford, "Datasheets for Datasets" *ArXiv.Org*, 19 March 2020. https://arxiv.org/abs/1803.09010.

[36] Asaro, Peter M, "AI Ethics in Predictive Policing. From Models of Threat to an Ethics of Care" *IEEE Technology and Society Magazine*, June 2019, p. 50.

[37] *Ibid*.

[38] The author's translation. Original French language: "Le care nous ramène a cette exigence de réalisme, au sens de la nécessité de voir ce qui est sous nos yeux: la réalité de l'inégalité, avant l'idéalité des principes." Laugier, Sandra, "Le care comme critique et comme féminisme", *Travail, genre et sociétés*, vol. 26, no. 2, 2011, p. 185-186.

[39] Wagner, B., op. cit., 2019.

[40] Article 19, "Governance with Teeth: How Human Rights Can Strengthen FAT and Ethics Initiatives on Artificial Intelligence", London, Article 19, April 2019, p. 11. https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

In the face of this, it is essential for ethicists to (1) avoid giving tools and resources that may serve this form of misuse of ethics, and (2) combat these misuses. To do so, ethicists need to be clear on the power relations and economic interests at stake. For instance, with these interests at stake, they need to consider when self-regulation is a meaningful and potentially effective response and when it is not. For instance, Access Now, another non-governmental organisation, has demonstrated in a recent report, "taking an 'ethics'-based approach to facial recognition and other dangerous applications of AI would leave millions exposed to potential human rights violations, and with little to no recourse."[41] As a SIENNA stakeholder put it: we need to be clear on what we can realistically expect from ethics. This is even more important bearing in mind the power of the big technology companies, especially the GAFAM.[42]

Paying attention to relations of power in the AI field also entails considering the severe diversity issue in the AI community. The 2019 report by the AI Now Institute "Discriminating Systems. Gender, Race, and Power in AI" powerfully points to this issue and calls the AI industry to "acknowledge the gravity of its diversity problem".[43] This "diversity problem" is not only an employment issue; it has direct implications on the type of AI systems produced. As West et al. write, "these patterns of discrimination and exclusion reverberate well beyond the workplace into the wider world", i.e., they lead to the development of technological products that are more beneficial to males than females or non-binary, to white rather than non-white.[44] Once again, this is essential to take into account for an AI ethics that directly engages with its situatedness and develops adequate tools to ensure AI systems are assessed with their social, political, and environmental impacts in consideration, so that the negative ones may be properly addressed and mitigated.

3.3 Enabling ethical agency

Hence, this piece argues that ethics must entail a sharp attention to specific situations and relations, accounting for the different levels of the personal, the interpersonal, the organisational, up to broader social, political, and environmental configurations. But a question then arises: how can there be any ethics guidance, if ethics is primarily a matter of attention to specific situations? In other words, isn't this approach to ethics as attention rendering any recommendations inadequate as they would take away from the specificities of the context? Although this document argues that ethics is primarily a matter of attention to specific situations, guidance is still needed. However, the status of such ethics guidance needs to be specified (the guidance formulated in this piece is precisely of this nature).

[41] Access Now, "For a truly 'Trustworthy AI,' EU must protect Rights and deliver benefits", 7 Dec 2020, https://www.accessnow.org/eu-trustworthy-ai-strategy-report/

[42] As we write this document (Dec 2020), there are ongoing policy initiatives in the EU to regulate big technology companies. These companies are fighting hard to have a seat at the table of discussions in order to limit the extent of these regulations as much as possible. Satariano, Adam and Stevis-Gridness, Matina, "Big Tech Turns Its Lobbyists Loose on Europe, Alarming Regulators", *The New York Times*, 14 Dec 2020. https://www.nytimes.com/2020/12/14/technology/big-tech-lobbying-europe.html

[43] West, Sarah Myers, Whittaker, Meredith, and Crawford, Kate. 'Discriminating Systems. Gender, Race, and Power in AI'. AI Now Institute, 2019, p. 3. https://ainowinstitute.org/ discriminatingsystems.html. Hagendorff has analysed AI ethics guidelines and notes that only 37,1% of these have female authors, this number comes down to 7.7% for the guidelines developed in the FAT ML community. Hagendorff, T., op. cit., 2019.

[44] West, S. et al., 2019,

The recommendations formulated here do not aim to say what one should do or should not do, but rather aim to *promote the conditions of possibility of doing ethics, i.e., of being able to identify the right course of action from within a particular situation*. In that sense, the form of ethics that is promoted here does not determine the 'right' or the 'good' from a distanced position; it is not prescriptive or imperative. Rather, it aims to provide tools and resources to ensure actors, in their situated position – such as a developer in an organisation, within her own role, position, organisation, socio-political-environmental context, and confronted to a particular engineering challenge – can make ethical choices. Rather than determining from a distanced position what the ethical thing to do is, it should be ensured *actors are in a position to determine what is the right thing to do*. To use the terminology proposed by Canguilhem, it is not a matter of determining norms, i.e., the right thing to do, but to developing, actors' "*normative capacity*", which we can also define as ethical agency.[45]

This approach avoids the risk of ethics guidance being patronising. This is one issue that was raised by a SIENNA stakeholder from a big technology company: the risk of the ethicist "coming on high horse" with predetermined ideas of what engineers should do. As Miller and Coldicutt show, technology workers are sensitive to the impacts of their products: "79% agree it's important to consider potential consequences for people and society when designing new technologies".[46] They have called this capacity "personal moral compass".[47] For AI ethics, it is essential to recognise this already existing sensitivity to issues and ability to respond to ethical challenges. However, as participants in the SIENNA workshop on Strategies of ethical AI highlighted, this is not sufficient. This ethical sensitivity and ability need to be further enhanced, promoted and protected. This is precisely the objective of the practical recommendations below. Promotion and protection of ethical sensitivity and ability can take various shapes. One of these is by conducting *impact assessment* of AI technologies and products. Miller and Coldicutt point to the fact that 81% of people in AI "would like more opportunities to assess the potential impacts of their products".[48] These can also take the shape of operationalised guidance documents, such as those developed in SIENNA with the ethics by design methodology, or research ethics guidance documents.

It is essential to clarify that these are only tools to promote ethical sensitivity and ability to take the right decision. Human expertise and sense of the specificity of the situation at stake (e.g., *this* AI system developed in such and such context with such and such objective) remains essential. In other words, high-level principles, norms, guidance, although they can help provide the broad lines, cannot fully dictate what should be done and what should not be in a specific situation. There is always, necessarily, the need to pay attention to the specificities of a particular situation. Therefore, for instance, in addition to research ethics guidelines, research ethics also needs human expertise in order to ensure the guidelines (necessarily general) are properly applied and interpreted with a specific situation (necessarily particular). The EUREC/SIENNA online workshop on 26-27 October 2020 that brought together members of European research ethics committees and SIENNA consortium partners on the

[45] Canguilhem, Georges, *The Normal and the Pathological*, Zone Books, 1991.
[46] Miller, Catherine, and Rachel Coldicutt, "People, Power and Technology: The Tech Workers' View," London, Doteveryone, 2019, p. 16. https://doteveryone.org.uk/report/workersview.
[47] *Ibid.*
[48] *Ibid.*, p. 19

topic of research ethics guidelines made this clear: expertise in interpreting and applying ethics guidelines to specific research cases is essential to research ethics.[49]

Another way of ensuring AI products/technologies are developed with care for their socio-politico-environmental impacts, is through the *protection of whistle-blowers and unions*. The need to protect whistle-blowers was mentioned by a participant in the Multi-Stakeholder Strategies for Ethical AI workshop on 8-9 September 2020. The protests by Google employees against the Maven project, in which Google was developing AI technology for US military drone programme, is a strong example of this.[50] Google employees saw issues with a powerful technology company such as Google getting into "the business of war".[51] Doing ethics implies asking hard questions, and those who have the courage to do so should be protected. In other words, if we want workers "to be vectors of change",[52] they need to be able to raise concerns when they see something that is problematic in their organisation and be protected appropriately from the ensuring fall out.

4. Conclusion

To conclude, this document has shown that current AI ethics guidance and initiatives tend to be dominated by a principled approach to ethics. Although this brings value to the field, it also entails some risks, especially in relation to the abstraction of this form of ethics that makes it poorly equipped to engage with and address deep socio-political issues and practical impacts. As such this document has sought to complement the existing principled approach to ethics with an approach to ethics as attention to context and relations. Below are some more practical recommendations to promote ethical AI by drawing from an approach to ethics as attention.[53]

5. Practical recommendations for AI ethics

AI ethicists should:

- → **Engage with social scientists and their research on social impacts of AI** in the short, medium and long term
- → **Engage with data scientists and software engineers** to better understand the way AI systems are developed and the data collected, cleaned, processed and interpreted

[49] The ethics guidelines the workshop focused on included guidelines identified as the most important ones used by research ethics committees in Europe.

[50] Gibbs, Samuel, "Google's AI is being used by US military drone programme", *The Guardian*, 7 March 2018. https://www.theguardian.com/technology/2018/mar/07/google-ai-us-department-of-defense-military-drone-project-maven-tensorflow

[51] Wakabayashi, Daisuke, and Shane, Scott, "Google Will Not Renew Pentagon Contract That Upset Employees", The New York Times, 1 June 2018. https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html. More recently, the AI ethics researcher Timnit Gebru has been forced out by her employer at Google for raising a number of ethical issues related to AI technology used by the company. Hao, Karen, "We read the paper that forced Timnit Gebru out of Google. Here's what it says", *MIT Technology Review*, 4 Dec 2020. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

[52] Jobin, Anna, "Why Dr. Timnit Gebru Is Important for All of US", *Medium*, 8 Dec 2020. https://annajobin.medium.com/why-dr-timnit-gebru-is-important-for-all-of-us-5c12d9d08c12

[53] Please note that several of these recommendations are relevant to diverse groups and are therefore repeated.

- → **Draw from ethics theories beyond principlism**, especially the ethics of care, virtue ethics, or Spinozist ethics.
- → **Ensure diversity** in the composition of AI ethics team (especially encouraging inclusion of non-white females and non-binary).
- → **Pay attention to the "privilege hazard"**, i.e., the risk of people in position of privilege failing to notice instances of oppression and injustices perpetuated by AI technologies
- → **Recognise, build upon and further develop AI developers' ethical sensitivity** (rather than impose guidance that may be top-down and disconnected from practice)
- → **Develop/share use cases on ethical and social impacts of AI** (especially negative ones) to make impacts of AI more concrete and understandable
- → **Engage more with the impacted communities, especially the most vulnerable among them,** and consider social and ethical impacts from their perspective
- → Take into consideration **environmental impacts of AI** (throughout the whole life cycle, from resource extraction to end-of-life disposal)
- → When carrying out impact assessments, consider **impacts on labour**, especially the conditions of "micro-workers" and other precarious workers in the AI industry

The AI industry should:

- → Engage with **social sciences studies** on the social and material impacts of AI
- → Promote **inclusion of social scientists** in AI research and development projects
- → Recognise that a technological product is **never "neutral"** and always reflects specific worldviews and intentions
- → Conduct/require **assessments of ethical, social, environmental, and human rights impacts of AI** before, during and after deployment of AI systems
- → **Encourage prudence and honesty about the capabilities of the AI system being sold and do not gloss over the limits**
- → **Ensure diversity** in the composition of the team researching, developing, using, or assessing AI (especially encouraging inclusion of non-white females and non-binary).
- → **Pay attention to the "privilege hazard"**, i.e., the risk of people in position of privilege failing to notice instances of oppression and injustices perpetuated by AI technologies
- → **Provide a safe environment for whistle-blowers and union members** in the AI industry
- → Develop and ensure respect for **research ethics in the private R&D AI sector.**
- → Promote and create r**esearch ethics committees/AI ethics officers** for AI R&D in the private sector
- → **Engage more with the impacted communities, especially the most vulnerable among them,** and consider social and ethical impacts from their perspective
- → **Recognise that not all social problems can be solved with technology**, i.e., be wary of technosolutionism

AI researchers and developers should:

- → Engage with **social sciences studies** on the social and material impacts of AI
- → Promote the **inclusion of social scientists** in AI projects

- → Recognise that a technological product is **never "neutral"** and always reflects a specific worldviews and intentions
- → Conduct/require **assessments of ethical, social, environmental, and human rights impacts of AI** before, during and after deployment
- → **Consider the possibility of *not* developing an AI system** when ethical and social issues identified are too severe and difficult to mitigate. Also account for the remaining uncertainties on the ethical and social impacts of AI.
- → **Ensure diversity** in the composition of the team developing, using or assessing AI (especially encouraging inclusion of non-white females and non-binary).
- → **Pay attention to the "privilege hazard"**, i.e., the risk for people in position of privilege to fail to notice instances of oppression and injustices perpetuated by AI technologies
- → **Report on societally and human rights harmful development of AI technologies** and report incidents to AI incident databases, to regulators or civil society organisations, particularly where no impact assessment of such technologies has taken place.
- → **Engage more with the impacted communities, especially the most vulnerable among them,** and consider social and ethical impacts from their perspective
- → **Recognise that not all social problems can be solved with technology**, i.e., be wary of technosolutionism

Policy-makers engaged in AI regulation and governance should:

- → **Do not let the AI hype hide the ethical and social issues this technology brings about and the difficulty to solve them** (e.g., exacerbation of structural inequalities, privacy risks and surveillance) as well as the remaining uncertainties on the ethical and social impacts of AI
- → **Protect whistle-blowers and unions** in the AI industry and develop adequate protection mechanisms and safeguards where this is missing.
- → **Encourage research ethics in private-sector AI R&D**
- → **Encourage the creation/use of research ethics Committees** for AI R&D in the private sector
- → Require/encourage/mandate the conduct of **assessments of ethical, social, environmental, and human rights impacts of AI** before, during and after deployment.
- → **Engage more with the impacted communities, especially the most vulnerable among them,** and consider social and ethical impacts from their perspective
- → **Recognise and promote awareness of the fact that not all social problems can be solved with technology**, i.e., be wary of technosolutionism
- → **Be wary of the AI hype**

AI public research funding organisations should:

- → **Fund cutting-edge social science studies** on AI and its impacts in the short, medium, and long term.
- → Require/encourage the conduct of **assessments of ethical, social, environmental, and human rights impacts of AI** research and development
- → **Engage more with the impacted communities, especially the most vulnerable among them,** and consider social and ethical impacts from their perspective

→ **Recognise and promote awareness of the fact that not all social problems can be solved with technology**, i.e., be wary of technosolutionism
→ **Be wary of the AI hype**

AI STEM (science, technology, engineering, mathematics) educators should:

→ **Train AI researchers and developers** on non-neutrality of AI and familiarise them about AI ethical and social impacts and impacts on human rights
→ **Develop/share use cases on ethical and social impacts of AI** (especially negative ones) to make impacts of AI more concrete and understandable
→ **Recognise and promote awareness of the fact that not all social problems can be solved with technology**, i.e., be wary of technosolutionism

6. References

Access Now, "For a truly 'Trustworthy AI,' EU must protect Rights and deliver benefits", 7 Dec 2020, https://www.accessnow.org/eu-trustworthy-ai-strategy-report/

Anscombe, Gertrude, E., M., "Modern moral philosophy", *Philosophy*, Vol. 33, Issue 124, 1958, pp. 1-19

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias", *ProPublica*, May 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arras, John, D., "The Way we Reason Now: Reflective Equilibrium in Bioethics", in Bonnie Steinbock, *The Oxford Handbook of Bioethics*, Oxford University Press, Oxford, 2009.

Article 19, "Governance with Teeth: How Human Rights Can Strengthen FAT and Ethics Initiatives on Artificial Intelligence", London, Article 19, April 2019. https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

Asaro, Peter M, "AI Ethics in Predictive Policing. From Models of Threat to an Ethics of Care" *IEEE Technology and Society Magazine*, June 2019, p. 50.

Benkler, Yochai, "Don't Let Industry Write the Rules for AI", *Nature* 569, 2019, p. 161.

Birhane, Abeba, and Cummins, Fred, "Algorithmic Injustices: Towards a Relational Ethics", arXiv: 1912.07376, 2019.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings", ArXiv.Org, 2016.

Buolamwini, Joy, and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", in *Conference on Fairness, Accountability, and Transparency*, 81, New York: PLMR, 2018, pp. 1-15.

Campolo, Alexender, and Kate Crawford, "Enchanted Determinism: Power without Responsibility in Artificial Intelligence", *Engaging Science, Technology, and Society*, Vol. 6, 2020, pp. 1–19.

Canguilhem, Georges, *The Normal and the Pathological*, Zone Books, 1991.

Crawford, Kate, and Joler, Vladan, "Anatomy of an AI System: The Amazon Echo as An Anatomical Map of Human Labor, Data and Planetary Resources", AI Now Institute, September 2018.

Criado Perez, Caroline, *Invisible Women. Exposing Data Bias in a World Designed for Men*, London, Chatto & Windus, 2019.

D'Ignazio, Catherine, and Klein, Lauren F., *Data Feminism*, MIT Press, Cambridge, MA; London, England, 2020.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, and Madhulika Srikumar, "Principled Artificial Intellience: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI", Cambridge, MA: Berkman Klein Center for Internet & Society at Harvard University, January 2020, https://cyber.harvard.edu/publication/2020/principled-ai.

Gebru, Timnit, "Race and Gender", in Markus D. Dubber, Frank Pasquale, and Sunit Das, *The Oxford Handbook of Ethics of AI*, Oxford, Oxford University Press, 2020.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford, "Datasheets for Datasets" *ArXiv.Org*, 19 March 2020. https://arxiv.org/abs/1803.09010.

Gibbs, Samuel, "Google's AI is being used by US military drone programme", *The Guardian*, 7 March 2018. https://www.theguardian.com/technology/2018/mar/07/google-ai-us-department-of-defense-military-drone-project-maven-tensorflow

Gilligan, Carol, *In a Different Voice: Psychological Theory and Women's Development*, Harvard University Press, Cambridge, MA, 1982.

Google, Artificial Intelligence at Google: Our Principles, https://ai.google/principles/.

Hagendorff, Thilo. "The Ethics of AI Ethics. An Evaluation of Guidelines", *Minds and Machines*, no. 30, 2019, pp. 99–120;

Hao, Karen, "We read the paper that forced Timnit Gebru out of Google. Here's what it says", *MIT Technology Review*, 4 Dec 2020. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Hart, Robert David, "If You're Not a White Male, Artificial Intelligence's Use in Healthcare Could Be Dangerous," *QZ*, July 10, 2017. https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use- in-healthcare-could-be-dangerous/.

High-Level Expert Group on Artificial Intelligence, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", European Commission, Brussels, July 2020, https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI", European Commission, Brussels, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai;

Izoard, Celia, "Les Réalités Occultées Du 'progrès' Technique : Inégalités et Désastres Socio-Écologiques", *Ritimo*, no. 21, May 2020, pp. 27–33.

Jansen, Philip, Philip Brey, Alice Fox, Jonne Maas, Bradley Hillas, Nils Wagner, Patrick Smith, Isaac Oluoch, Laura Lamers, Hero van Gein, Anais Resseguier, Rowena Rodrigues, David Wright, David Douglas, Ethical Analysis of AI and Robotics Technologies, SIENNA D4.4, Aug 2019.

Jobin, Anna, "Why Dr. Timnit Gebru Is Important for All of US", *Medium*, 8 Dec 2020. https://annajobin.medium.com/why-dr-timnit-gebru-is-important-for-all-of-us-5c12d9d08c12

Jobin, Anna, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines", *Nature Machine Intelligence* 1, no. 9, 2019, pp. 389–99

Laugier, Sandra, "Le care comme critique et comme féminisme", *Travail, genre et sociétés*, vol. 26, no. 2, 2011.

Miller, Catherine, and Rachel Coldicutt, "People, Power and Technology: The Tech Workers' View," London, Doteveryone, 2019. https://doteveryone.org.uk/report/workersview.

Mittelstadt, Brent, "Principles Alone Cannot Guarantee Ethical AI", *Nature Machine Intelligence 1*, Nov 2019, pp. 501-507.

Molinier, Pascale, "De la civilisation du travail à la société du Care", *Vie Sociale* 14, no. 2, 2016.

Molinier, Pascale, Sandra Laugier, and Patricia Paperman, *Qu'est-ce que le care? Souci des autres, sensibilité, responsabilité*, Payot & Rivages, Paris, 2009

Nagel, Thomas, *The View from Nowhere*, Oxford, Oxford University Process, 1989.

OECD, "Recommendation of the Council on Artificial Intelligence", adopted on 22 May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449;

Resseguier, Anais and Rowena Rodrigues, "*AI ethics should not remain toothless!* A call to bring back the teeth of ethics", *Big Data & Society*, July-Dec 2020, pp. 1-5.

Satariano, Adam and Matina Stevis-Gridness, "Big Tech Turns Its Lobbyists Loose on Europe, Alarming Regulators", *The New York Times*, 14 Dec 2020. https://www.nytimes.com/2020/12/14/technology/big-tech-lobbying-europe.html

Strubell, Emma, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP", *arxiv.org*, 2019. https://arxiv.org/abs/1906.02243;

Tubaro, Paola, Antonio Casilli, and Marion Coville, "The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence", *Big Data & Society*, Jan-June 2020, pp. 1–12.

Wagner, Ben, "Ethics as an escape from regulation: From ethics-washing to ethics-shopping" in Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, *Being Profiled: Cogitas Ergo*

Sum: 10 Years of Profiling the European Citizen, Amsterdam University Press, Amsterdam, 2019, pp. 84–89.

Wakabayashi, Daisuke, and Shane, Scott, "Google Will Not Renew Pentagon Contract That Upset Employees", The New York Times, 1 June 2018. https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.

West, Sarah Myers, Meredith Whittaker and Kate Crawford, "Discriminating Systems. Gender, Race, and Power in AI", AI Now Institute, 2019. https://ainowinstitute.org/ discriminatingsystems.html.