Ethics by design and research ethics for artificial intelligence

Jansen, Philip; Henschke, A.H.; Erden, Y. J.; Marchiori, S.; Brey, Philip; Hoefsloot, Marit

DOI
10.21253/DMU.16912345.v1

Publication date
2021

Document Version
Final published version

Citation (APA)
Jansen, P., Henschke, A. H., Erden, Y. J., Marchiori, S., Brey, P., & Hoefsloot, M. (2021). *Ethics by design and research ethics for artificial intelligence*. SHERPA Project. https://doi.org/10.21253/DMU.16912345.v1

SHERPA

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

Deliverable No. 5.7

Ethics by Design and Research Ethics for Artificial Intelligence

[28 October 2021]

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

Document Control

Deliverable	D5.7: Ethics by Design and Research Ethics for Artificial Intelligence
WP/Task Related	WP5
Delivery Date	28 October 2021
Dissemination Level	Public
Lead Partner	University of Twente
Contributors	Philip Jansen, Adam Henschke, Yasemin J. Erden, Samuela Marchiori, Philip Brey, Marit Hoefsloot
Reviewers	Doris Schroeder, Bernd Carsten Stahl
Abstract	In this report, we outline the Ethics by Design approach, especially as developed in the SIENNA project. We then offer some approaches to teaching this methodology so as to support researchers in their practice. The report can also be used to embed Ethics by Design, and as a toolkit for researchers and research ethics assessors, including for research ethics for artificial intelligence. The report begins with a review of Ethics by Design, followed by an outline of some challenges and proposed solutions. Then we offer a foundation for applying this content in an educational context, first in academia and then as training materials that can be used in a business or company setting. In the annexes, we provide sample syllabi and course outlines. These materials have been piloted in a number of contexts, including as delivered in training sessions for researchers and employees within the European Commission's H2020 funding programme, and as offered to those who intend to bid for Horizon Europe funding. The report builds on existing SHERPA work to develop guidelines for developers of AI as well as guidelines for users of AI. In these ways the task feeds into the EC's guidance for ethics review of Horizon Europe projects.
Key Words	Ethics by Design; Research Ethics; Artificial Intelligence

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
1.4	08/10/21	Philip Jansen, Adam Henschke, Yasemin J. Erden, Samuela Marchiori, Philip Brey	UCLan Cyprus, Bernd Carsten Stahl	First Draft
2.0	28/10/21	Philip Jansen, Adam Henschke, Yasemin J. Erden, Samuela Marchiori, Philip Brey		Final Draft

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

Table of Contents

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

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

Executive Summary

In this report, we outline the Ethics by Design approach developed in the SIENNA project[1] which built on earlier work within SHERPA,[2] and offer approaches to teaching this methodology so as to help researchers in their practice. The report can also be used to embed Ethics by Design, and as a toolkit that can used by researchers and research ethics assessors, including for research ethics for artificial intelligence (AI).

In chapter one we introduce Ethics by Design and research ethics, and outline why these approaches are important to consider in the context of AI ethics. In chapter two we review Ethics by Design approaches. This includes the SIENNA Ethics by Design approach and their research ethics guidelines for AI. We present the Horizon Europe ethics appraisal procedure for AI and discuss both the Institute of Electrical and Electronics Engineers (IEEE) Standard for Value-Based Engineering as well as the World Economic Forum's (WEF) proposals for Ethics by Design for AI.

Then in chapter three we outline some challenges in Ethics by Design and propose some solutions. We cover issues related to skills, integration, structures and tools, management, and present ways to manage resistance and promote stakeholder communication.

Chapters four and five offer content that can be used in teaching and training materials respectively. In chapter four we present a foundation for applying this content in a traditional educational context, such as within academia, and then in chapter five we outline what training materials for a business or company setting should include. In the annexes we provide course outlines and syllabi which can be used as templates, edited, and further developed and adapted as required.

These materials have been piloted in a number of contexts, including as delivered in training sessions for researchers and employees within the European Commission (EC) H2020 funding programme, and as offered to those who intend to bid for Horizon Europe funding.[3] The report builds on existing SHERPA work to develop guidelines for developers of AI as well as guidelines for users of AI. In these ways the report feeds into the EC's guidance for ethics review of Horizon Europe projects, and within a broader approach to ethics in a 'research ethics framework'. Recently these approaches encourage researchers to use an Ethics by Design approach to ensure that ethical issues are systematically accounted for, including within ethics self-assessment processes, such as within Horizon Europe funding programmes.

The training materials offered in this report can be used beyond the lifetime of the SHERPA project, including as a basis for training that can be offered to interested audiences such as EC National Contact Points (NCPs) and industry. SHERPA further intends to develop an Ethics by Design training programme that will be offered as a commercially available resource through the non-profit spin-out company ORBIT.[4]

---

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

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

[3] The most recent of which were two SHERPA training courses on 06/10/21 and 26/10/21. For the first course there were 72 registrations and 39 participants on the day, for the second 68 people registered and 31 attended. The training was provided by SHERPA partners from De Montfort University (DMU) and University of Twente (UT). Cf. https://www.project-sherpa.eu/ethics-by-design-course-registration/

[4] Cf. www.orbit-rri.org

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

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
EC	European Commission
EU	European Union
EbD	Ethics by Design
IEEE	Institute of Electrical and Electronics Engineers
SHERPA	Shaping the Ethical Dimensions of Smart Information Systems (EU project)
SIENNA	Stakeholder-Informed Ethics for New techNologies with high socio-economic and humAn rights impact (EU project)
WEF	World Economic Forum

Table 1: List of acronyms/abbreviations

Glossary of terms

Term	Explanation
Artificial Intelligence	The science and engineering of machines with capabilities that are considered intelligent (i.e., intelligent by the standard of *human* intelligence).[5]
Ethics	Ethics is a branch of philosophy dealing with the moral principles that govern an individual's behaviour. Applied ethics deals with the use of moral principles in real-life situations. Ethics of AI is an example of applied ethics focused on the ethical issues raised by AI.
Ethics by Design	Ethics by Design is an approach that aims to incorporate ethical considerations into every stage of a technology's life cycle, from its design to its development and implementation, in order to mitigate possible negative ethical consequences produced by the technology.

[5] Jansen, P., Brey, P., Fox, A., Maas, J., Hillas, B., Wagner, N. et al. (2020). SIENNA D4.4: Ethical Analysis of AI and Robotics Technologies (Version V1.1). Zenodo.

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

Research Ethics	Research ethics is a branch of ethics that deals with the practice of conducting research in accordance with ethical principles that apply to the research practice.
Ethical requisite	A requirement relating to ethical aspects of the AI systems and their development. In order to be compliant with the demands for responsible, trustworthy, ethical AI, ethical requisites must be adhered to. The term is used extensively in previous work on Ethics by Design in the SIENNA project.

Table 2: Glossary of terms

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

1. Introduction

Developing and using technologies like artificial intelligence (AI) raises many ethical issues, but how and why those ethical issues arise, and how to manage or resolve them is not always easy to agree. There are some who would claim that technologies are themselves relatively neutral, and that ethical issues arise because of the ways that they are used.[6] Yet this approach fails to recognise that design and use are intimately connected, as we'll show in this chapter.

Others argue that technologies, and in particular the way that they are designed are value-laden, whether this is the intention or not.[7] On this account, human choices can never be value-neutral since they demonstrate not only a person's values, but also their tendencies, preferences, beliefs, and biases (whether negative or otherwise). Choices and decisions are also made within a context of opportunity and constraint (existing or emergent), as well as social, political, and economic structures. All of which influence and underpin the perspectives and decisions that are made during the design of scientific and technological products and processes. Accordingly, social and ethical concerns need to be considered at all stages of the design and development of a technology, and not left until after the technology has been built. Especially given that 'scientific knowledge, technological invention, and corporate profit reinforce each other in deeply entrenched patterns that bear the unmistakable stamp of political and economic power'.[8]

The question is then how we select and apply key values and ethical principles[9] to the design of technologies, and in this case to AI? To do so means that that we need to consider how to resolve complex ethical issues**,** many of which defy easy answers. These include difficulties associated with striking a balance between benefits and risks within contexts where there may be many unknowns or uncertainties. It can also be difficult to assess the impacts of technologies on societies, including in terms of individual health (physical, emotional, psychological) and in terms of impacts related to accessibility and to socio-economic outcomes. There are of course legal implications to consider, with associated difficulties in regulating new and emerging technologies, but the legal cannot be conflated with the ethical. Legal measures are rarely sufficient to address and manage ethical and social issues where technology is concerned.

The precautionary approach, originally devised as a response to environmental risk, is one method to reduce the possibility of harm. It states that where 'there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.'[10] In this way, epistemological uncertainty cannot be used as a justification for ignorance; lack of certainty is not sufficient for denying the *plausibility* of potential for harm. While the precautionary approach is a useful tool to govern or regulate potentially harmful

---

[6] Cf. Pitt, J. C. (2014). "Guns Don't Kill, People Kill"; Values in and/or Around Technologies. In Kroes, P., & Verbeek, P. P. (eds) *The Moral Status of Technical Artefacts: Philosophy of Engineering and Technology*, 17, pp. 89-101. Springer, Dordrecht.

[7] For further information on the debate regarding value and technology, cf. van de Poel I. and Kroes P. (2014). Can Technology Embody Values?. In: Kroes P. and Verbeek P. P. (2014) Op cit.

[8] Winner, L. (1980). Do artifacts have politics?. *Daedalus*, 109: 1, pp. 121-136 (p. 126).

[9] Values and principles are related but different concepts. Values relate to things that are deemed to be important, like well-being, whereas principles are ways of protecting, ensuring, or respecting things deemed to be important, such as a direction that we ought to maximise well-being, or a rule that we ought not to kill.

[10] Rio Declaration (1992). Rio declaration on environment and development. Report of the United Nations conference on environment and development, Rio de Janeiro, 3–14 June 1992. Annex I.

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

technologies, it offers limited guidance for the successful application of ethical principles to technologies that are not so obviously harmful.

For these kinds of reasons, several approaches and methodologies have been proposed that seek to offer practical steps to increase ethical outcomes. Some of these focus on innovation, including responsible innovation, and responsible research and innovation initiatives. Others focus on design, including value-sensitive design, universal design, participatory design, sustainable design, design for values (or well-being etc), and '…by design' approaches.[11] This deliverable focuses on the Ethics by Design approach.

The Ethics by Design approach aims to incorporate ethical considerations into every stage of the technology's life cycle, from its design to its development and implementation, in order to mitigate the negative ethical consequences produced by the technology. Ethics by Design is a very recent approach that started to gain traction around 2018, and limited academic research currently exists on the topic.[12] The term is associated with both the capability of agents to reason about the ethical aspects of their decisions (e.g., Dignum et al.[13]), and with the integration of ethical principles into the design and development processes of AI-based systems (e.g., d'Aquin et al.[14]). In this deliverable, the latter understanding of the term is used.

The European Commission has recently made Ethics by Design a priority topic for research in general, and AI specifically. However, not a lot of practice has been built up with the approach, in terms of actual use cases. For the approach to gain widespread acceptance, its rollout beyond Horizon Europe is necessary, as is the development of an Ethics by Design training infrastructure. As such, this report offers an introduction to some key principles to be applied in teaching and training courses on Ethics by Design.

2. Research Ethics and Ethics by Design for AI – A review

This chapter introduces and reviews the current state of the art of Ethics by Design and research ethics approaches for the field of artificial intelligence (AI). In doing so, it largely focuses on the approaches recently proposed by the SIENNA project.[15] These include SIENNA's Ethics by Design approach for AI, SIENNA's research ethics guidelines for AI, and SIENNA's recommendations for the European Commission's Horizon Europe Ethics Appraisal Procedure for AI. Additionally, this chapter points to the

[11] Cf. Davis, J., and Nathan, L. P. (2015). Value sensitive design: Applications, adaptations, and critiques. In Hoven, M. J., Vermaas, P. E., and Poel, I. R. (eds). (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*, pp. 11-40.

[12] Ethics by Design has emerged from a group of methods and approaches that seek to ensure ethical thinking permeates science and technology at every stage of design and development (see also value-sensitive design, responsible innovation etc.).

[13] Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L. et al (2018, December). Ethics by design: Necessity or curse?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 60-66.

[14] d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., and Holden, L. (2018, December). Towards an "ethics by design" methodology for AI research projects. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 54-59.

[15] Resseguier, A., Brey, P., Dainow, B., Drozdzewska, A., Santiago, N., and Wright, D. SIENNA D5.4: Multi-stakeholder strategy and practical tools for ethical AI and robotics (forthcoming).

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

Institute of Electrical and Electronics Engineers (IEEE) and World Economic Forum's (WEF) proposals for Ethics by Design for AI.

The proposals discussed here are the best developed existing proposals that were found during our literature search. Each of them has a significant focus on AI development and provides a practical set of guidelines to developers with the aim of preventing or reducing the occurrence of problematic ethical issues during the design and development process. To our knowledge, as of writing, no other full-blown Ethics by Design approaches have yet been published that are specific to AI.

2.1 SIENNA Ethics by Design approach for AI

Within the SIENNA project, an Ethics by Design approach was developed that offers a way by which to incorporate ethical principles into the design and development processes of AI-based systems. This approach builds on the SHERPA Guidelines for the development and use of Smart Information Systems.[16] It is aimed at preventing or reducing the occurrence of problematic ethical issues in relation to newly developed computational systems. Starting with a number of important ethical values, it derives from these a set of ethically-focused activities and tools that are intended for use throughout the design, development and deployment phases of a given AI project.

SIENNA's Ethics by Design approach can be described in terms of a five-level model. In this model, higher levels are more abstract, while specificity increases going down the levels. At the **first level** (from the top, hence more abstract), there are the Ethics by Design values, which are the primary ethical values by which the development of an AI system should be guided. A system violating these values can be considered unethical.

The approach lists six main values that are important in designing AI-based systems (which are ultimately based on those presented by the EU's High-Level Expert Group on AI (HLEG-AI) in its *Ethics Guidelines for Trustworthy AI*[17]):

- **Human agency**: This value encompasses the values of autonomy, dignity, and freedom. Respecting autonomy means allowing people scope to decide for themselves what is right and wrong, and the way they should live their lives. Respecting dignity means every human being possesses an intrinsic worth, which should be protected. And respecting freedom means leaving people free to exercise their autonomy and live with dignity.
- **Privacy and data governance**: As a value, data governance means that personal data and the manner in which a system uses it must be proactively managed. 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., bias related to gender or ethnicity) and personal data (where the data subject's rights and freedoms must be safeguarded).
- **Fairness***:* This value means that all people have the right to be treated fairly and not on the basis of irrelevant characteristics. In this sense, non-discrimination is the application of fairness in the context of human characteristics as they are used to determine burdens and benefits. In particular, people should not be treated unfairly on the basis of aspects of their personal identity, including gender, race, age, sexual orientation, ethnicity, nationality, religion, health and disability.

[16] Warso, Z. & Gaskell, S. (2019). SIENNA D3.2: Analysis of the legal and human rights requirements for Human Enhancement Technologies in and outside the EU (Version V2.0). Zenodo.

[17] 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

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

- **Well-being***:* This value covers a range of properties. Something has well-being when its needs are met, and it is able to flourish. The values of autonomy and freedom mean that people can achieve well-being if they are able to work towards their ambitions and live what they consider to be *a meaningful life*.
- **Accountability and oversight***:* Human oversight as a value requires that 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 and knowledge about it. Accountability means there are mechanisms to explain how, and why, a system exhibits particular characteristics.
- **Transparency***:* This value contributes to human agency, data governance and oversight. Transparency includes all elements relevant to an ethical AI system: the data, the system and the processes by which it is designed, deployed and operated should all be transparent. Without this level of transparency, a decision cannot be contested, or even understood. This would often make it impossible to correct errors and unethical occurrences.

At the **second level** of the five-layer model, there are "ethical requisites". These are the conditions that a solution or application must meet in order to achieve its goals in an ethical fashion. They are instantiations of the values that AI systems should reinforce. 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 negative or unfair bias (for instance on the basis of race or gender).

For each of the six main Ethics by Design values for AI, the SIENNA approach details a number of general ethical requisites. The main ethical requisites for AI 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.'[18]

At the **third level** of the model, there are the ethical guidelines. Whereas ethical requisites are concerned with the final characteristics of AI systems, ethical guidelines focus on the steps by which AI systems are created, with a view to ensuring that each step is ethical. Ethics by Design for AI works on the basis that there are steps in the development process which are common to all AI design methodologies. The Ethics by Design approach offers a generic description of these steps in the development process (see below) and maps the ethical requisites onto these phases. This leads to specific guidelines (usually formulated as tasks) at each phase, which ensures that the final system instantiates the ethical requisites and therefore does not violate any of the six ethical values the system embraces. For example, the guidelines state that

[18] Ibid., pp. 36-37.

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

during the data gathering stage, data should be screened for fairness and that any identified discriminatory biases should be corrected.

The approach's generic design model for AI is as follows:

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 the 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.* This is the process of collecting, verifying, cleaning, formatting and integrating relevant data.
5. *Detailed design and development.* This involves the actual construction of a complete 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 the system, including an evaluation against the original objectives and requirements.

At the **fourth level** of the model, there are the AI methodologies. A variety of methodologies are used in AI projects. Examples are Agile, CRISP-DM and V-model. They are, at least partially, distinguished by the manner in which they organise the development process. Each methodology offers its own steps and sequence. If a project uses a particular methodology, a developer can refer to the generic model and map the steps in the generic development process to their own methodology thereby finding relevant guidelines.

Finally, at the **fifth level** of the model, there are the AI tools and methods. This level accommodates specific programmatic artefacts and processes deployed within the development process (e.g., Datasheets for Datasets) to undertake Ethics by Design. Those which are used in the development methodology in question are modified to support the ethical requisites.

2.2 SIENNA research ethics guidelines for AI

Next to an Ethics by Design approach for AI, the SIENNA project has developed a set of research ethics guidelines for AI.[19] The project discussed how these could form a foundation for stand-alone guidance for AI research and development, and how they could be incorporated into broader research ethics frameworks for computer and information science, and for frameworks covering multiple disciplines. SIENNA's proposal was based on the general ethics guidelines for AI put forth by the EU's High-Level Expert Group on AI (HLEG-AI) in its *Ethics Guidelines for Trustworthy AI.*

The SIENNA proposal includes 27 research ethics guidelines grouped into six categories:

- Human agency;
- Privacy and data governance;
- Fairness;
- Social and environmental well-being;

[19] SIENNA D5.4, op cit.

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

- Accountability and oversight; and
- Transparency.

When the guidelines are supposed to be standalone, SIENNA argued these are to be supplemented by general research ethics guidelines covering the following areas:

- 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.

In addition to these guidelines, SIENNA proposed a number of "special topics" guidelines, which are research ethics guidelines for particular techniques, products, and application domains within AI.

2.3 Horizon Europe Ethics Appraisal Procedure for AI

The European Commission, in collaboration with the SIENNA project, has recently developed an ethics appraisal procedure for the self-assessment of Horizon Europe applicants for projects involving the development and/or use of AI-based systems or techniques.[20] The procedure consists of a list of questions that applicants for such projects must answer.

The questions are premised on a number of key prerequisites for ethically sound AI systems, which have been identified by the HLEG-AI on AI set up by the European Commission and published in the *Ethics Guidelines for Trustworthy AI*.[21] These prerequisites include:

- *'Human agency and oversight* — AI systems must support human autonomy and decision-making, enabling users to make informed autonomous decisions regarding the AI systems.
- *Privacy and data governance* — AI systems must guarantee privacy and data protection throughout the system's lifecycle. The principles of privacy by design and by default must be taken into account in the process of designing, developing, selecting and using AI. The quality, integrity and security of data should be rigorously checked and adequately managed.
- *Fairness, diversity and non-discrimination* — Best possible efforts should be made to avoid unfair bias (e.g. stemming from the used data sets or the ways the AI is developed).
- *Accountability* — Appropriate mechanisms should be set in place to ensure auditability and accountability of the AI solutions and their outcomes, both before and after their development, deployment and use. Potential negative impacts should be identified and addressed at early stages.
- *Transparency* — All data sets and processes associated with AI decisions must be well communicated and appropriately documented. The principle of transparency is closely linked to the principles of tractability and explicability and facilitates the implementation of human agency, data governance and human oversight. It includes all elements relevant to an AI system (e.g., the data, the system and the processes by which it is designed, deployed and operated).

---

[20] European Commission (2021). EU Grants. How to complete your ethics self-assessment: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf

[21] 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

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

- *Societal and environmental well-being* — The impact of the developed and/or used AI system/technique on the individual, society and environment must be carefully evaluated and any possible risk of harm must be avoided. Increased vigilance is needed for solutions that may potentially have significant negative social or environmental impact.'[22]

Based on these prerequisites, the appraisal procedure for Horizon Europe asks applicants the following main questions:

- 'Could the AI based system/technique potentially stigmatise or discriminate against people (e.g. based on sex, race, ethnic or social origin, age, genetic features, disability, sexual orientation, language, religion or belief, membership to a political group, or membership to a national minority)?
- Does the AI system/technique interact, replace or influence human decision-making processes (e.g. issues affecting human life, health, well-being or human rights, or economic, social or political decisions)?
- Does the AI system/technique have the potential to lead to negative social (e.g. on democracy, media, labour market, freedoms, educational choices, mass surveillance) and/or environmental impacts either through intended applications or plausible alternative uses?
- Does the AI to be developed/used in the project raise any other ethical issues not covered by the questions above (e.g., subliminal, covert or deceptive AI, AI that is used to stimulate addictive behaviours, life-like humanoid robots, etc.)?
- Does this activity involve the use of AI in a weapon system?
- Does the AI to be developed/used in the project raise any other ethical issues not covered by the questions above (e.g., subliminal, covert or deceptive AI, AI that is used to stimulate addictive behaviours, lifelike humanoid robots, etc.)?'[23]

In the case that any of these questions are answered in the positive, applicants may need to provide detailed information on mitigating measures, and important documents may need be provided and/or kept on file.

2.4 IEEE Standard for Value-Based Engineering

The Institute of Electrical and Electronics Engineers (IEEE) and the IEEE Standards Association have produced a standard to address ethical concerns during the design of artificial intelligence (AI) and other technical systems.[24] The *IEEE 7000™-2021 - IEEE Standard Model Process for Addressing Ethical Concerns During System Design* offers a methodology to analyse human and social values relevant for ethical system development.[25]

The standard provides:

---

[22] European Commission (2021). op cit.
[23] Ibid.
[24] It should be noted that there are other standards that deal with *responsible innovation* (for example, the BSI PAS440 https://pages.bsigroup.com/l/35972/2020-03-17/2cgcnc1 and *ethics assessment* (for example, the CEN CWA 17145-2:2017, https://www.nen.nl/en/cwa-17145-2-2017-en-235742). However, unlike the IEEE's, these standards are less focused on AI, and on the design process and the specific steps that need to be taken here by developers to ensure adherence to ethical requirements.
[25] Cf. https://ieeexplore.ieee.org/browse/standards/reading-room/page/viewer?id=9536679

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

a) a system engineering standard approach integrating human and social values into traditional systems engineering and design;
b) processes for engineers to translate values and ethical considerations into system requirements and design practices; and
c) a systematic, transparent, and traceable approach to address ethically-oriented regulatory obligations in the design of autonomous intelligent systems.

As IEEE explains, 'the standards could help organisations that design, develop, or operate AI or other technical systems to directly address ethical concerns upfront, leading to more trust among the end-users and increased market acceptance of their products, services, or systems.'[26]

Adhering to the standard means that companies have to run through five processes:

- The first one is a *Context & Concept Exploration* process, which invites engineers to investigate the ethical feasibility of a project, to understand who the relevant stakeholders are, to explore the context of the system and to identify suitable partners and data processors that are willing to commit to a value-based service delivery in an accountable way.
- The second process is an *Ethical Value Elicitation* process, which channels the creativity of the innovation teams and stakeholders to think about ethically relevant values, harms and benefits, and personal virtues and maxims impacted by the future technology. It also provides companies with a way to prioritise the value mission, enhance their corporate value proposition, and embeds corporate principles and human rights into their value priorities.
- The third process is an *Ethical Value Requirements Identification* process, which translates value clusters into Ethical Value Requirements. These are technical and organisational requirements that the innovation teams derive as action goals from their "value book" (i.e., the ethically relevant values they have identified).
- The fourth process is a *Risk-Based Design* process, that runs ethically relevant values through a threat control analysis, at the end of which concrete system features, needs and design elements are identified. These features must then find entry in a functional product roadmap. Their effects must also be validated and monitored throughout the later system deployment to potentially improve and refine the value commitment.
- Finally, accompanying these four processes of value based engineering which IEEE 7000 foresees, is a fifth process which is a *Transparency* process that keeps a record of all the process stages and creates traceability on how the values (through EVRs) led to system features.

2.5 WEF proposal for Ethics by Design

The World Economic Forum (WEF) has published an organisational approach to the responsible use of technology that uses important insights from behavioural economics and psychology and combines them with findings from interviews with leaders of international organisations and market research.[27] The approach is aimed at helping to shape organisational design decisions, which the WEF suggests will result in better and more routine ethical behaviours. It suggests focusing on context and especially on creating the environments that can encourage people to engage in ethical behaviour. The report offers recommendations for organisational design that it says 'have proven to be more effective than conventional approaches such as compliance training and financial compensation' (p. 5).

[26] Cf. https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html
[27] WEF. (2020). *Ethics by Design: An organizational approach to responsible use of technology* (p. 37) [White Paper]. World Economic Forum. http://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf

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

The WEF's recommendations are as follows:

1. 'Invest in each of the dimensions of Epley's framework[28] – attention, construal and motivation – and consider how those investments work together. Companies should activate a mix of cultural elements to increase the likelihood that they will get employees' attention, help them use ethics and provide incentives to do so. Executives can activate aspects of culture that range from structure provided by the organizational chart, to what is said, what is done and the beliefs that are shared within the company supporting these choices. The matter of which dimension to address first is less important because most tools and tactics will likely touch on aspects of each dimension.
2. Use assessments to gain an understanding of how mature the motivation level in the organization is presently. These should be used as a benchmark to track the progression of people acting in ways that are simplistic – to get what they want or avoid getting in trouble, as opposed to being more consistent with the organization's values and purpose. Ultimately, employees should move from following rules to having a level of comfort in creating them.
3. Implement some form of regular organizational introspection. Companies are using mixes of surveys, focus groups and assessments that examine ethical cultures both broadly and with a specific focus on the use of certain technologies.
4. Check for the conditions that encourage ethical behaviour to take root: an organizational sense of responsibility to society as an actor that both influences and is influenced by it; the distribution of moral autonomy throughout the company to create a climate of trust; and the use of ethical deliberation practices.
5. Use ethical deliberation practices wherever possible by using data and information, involving those affected by decisions, considering the downstream effects and publicly sharing motivations behind decisions. Engage a diverse set of stakeholders, both internal to and external to the company.
6. Develop a rubric for ethical decision-making. Whether this is a set of principles or guidelines unique to a specific technology, such as AI, or used more comprehensively when the company makes decisions, its existence activates each aspect of Epley's recommended approach. Make sure the rubric is consistent with the organization's mission and values.
7. Teach these practices to members of any centralized ethical deliberation bodies, and train them on the organization's rubric as well.
8. Look for new opportunities to introduce ethics and do not shy away from opportunities to use challenging moments to do so. Organizations are primed for ethics in such moments.
9. As the challenges pass, capture them in the organization's memory, and return to them to reinforce the learning and create conditions that promote innovation. This can foster more routine ethical construal.
10. Institutionalize your commitment to ethics in hiring, orientation, training and evaluation protocols' (p. 26).

[28] Epley, N. and Tannenbaum, D. (2017). Treating Ethics as a Design Problem, *Behavioural Science & Policy*, 3: 2, pp. 72–84, https://behavioralpolicy.org/articles/treating-ethics-as-a-design-problem/

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

3. Implementing Ethics by Design: Challenges and solutions

This chapter discusses remaining challenges for the successful implementation of Ethics by Design and some possible solutions. These challenges were identified through discussion among the authors of this report and analysis of previous work on Ethics by Design. They include that the approaches require significant education and training for users, integration with preferred design approaches, structures and tools, organisational management, overcoming potential resistance to implementation among developers, and adequate practices of stakeholder communication. Some of these challenges are at the level of individuals (development of skills; positive attitudes towards ethics), whereas others are at the level of organisational processes and structures (adjustments to design approaches; organisational management; stakeholder practices).

3.1 Development of skills

Successful implementation of Ethics by Design depends on proper skills among developers in carrying out the approach. And such skills, in turn, depend on adequate training of developers. Training may take significant time and financial resources, and the result may depend on the level of affinity with and interest in ethics among developers. Some developers may dislike having to deal with 'muddy' ethical topics or may not think their job relates to ethical matters. This may dissuade organisations from investing in Ethics by Design.

Information materials for Ethics by Design training courses must make a convincing argument that such training is relevant, necessary and, in many cases, financially beneficial for organisations. In the case of the latter, arguments need to include that it may help prevent costly negative ethical impacts on the one hand and to show how it supports productivity in the development process on the other.

It may sometimes be necessary to appoint a special ethics officer with substantial knowledge, training, and experience in the field of ethics, and ideally as this intersects with technology, especially computation, AI, and robotics. If many organisations were looking to hire such individuals, demand for them might outstrip supply, which might leave some organisations empty-handed and with less capacity to conduct Ethics by Design.

3.2 Integration with preferred design approaches, structures and tools

Ethics by Design must, by definition, be embedded into the development process, and this means developers are very likely to need to change how they work. This is likely to require alterations to team structures, such as adding additional roles, creating new communication channels, and building additional review and decision processes. Additionally, it will require new 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 replaced.[29]

[29] SIENNA D5.4, op cit.

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

In some cases, it may not be technically possible to meet every Ethics by Design requirement due to a lack of suitable development tools. However, one should be extremely rigorous in investigations for suitable tools and their paucity should not be used to excuse to avoid ethical design.

3.2 Organisational management

At a higher level, the successful implementation of Ethics by Design requires that an organisation has proper management, staff who are able and committed to providing and maintaining the organisational structures and processes necessary for Ethics by Design. For example, it should be considered how the implementation of Ethics by Design affects the various dimensions of an organisation's management strategy, including overall objectives, quality management, portfolio management, risk management, data management, and stakeholder relationship management. Not all of such organisational measures have been specified by the SIENNA Ethics by Design approach.

3.3 Resistance among developers

In implementing Ethics by Design, there may be significant resistance among developers to the changes required, as some aspects of development will be more onerous than before. It is said that the best developers are many times more effective at what they do than the average.[30] Organisations are therefore unlikely to want to lose those developers to competing organisations should they be unhappy with the level of changes required.

Best practice to retain these high-value individuals is to actively involve them in the implementation of Ethics by Design and to empower them to lead on required changes, rather than to impose them from above. Resistance to implementation may be more likely, or even inevitable if the interpersonal dynamics of the development teams are not taken into consideration. More generally, an organisation's IT management should encourage a common culture of responsibility.

3.4 Stakeholder communication

In Ethics by Design, users and other stakeholders are actively involved in the development of systems. It may not always be easy to find and communicate with representatives of all possible groups of affected individuals. One needs to arrange meetings with stakeholders, ensure there is adequate representation of all their relevant interests, ensure stakeholders have an adequate understanding of what is at stake, and ensure they are comfortable enough to speak frankly. This can require a lot of time and effort, as well as sufficient trust in the management and organisational structure, as well as in the people involved.

[30] SIENNA D5.4, op cit.

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

4. Research Ethics and Ethics by Design for AI in academic education

Having reviewed the current approaches for Research Ethics and Ethics by Design for AI and the success factors for implementing them, we now focus on how their implementation in academic education should take place. The aim of this chapter is twofold: first, it sets out to discuss the objectives for academic education on Research Ethics and Ethics by Design for AI in academic education; and second, it aims to examine ways in which such objectives can be met. This chapter therefore references and builds on the work carried out in the SIENNA project and as outlined in chapter 2 above.

The materials offered in this report form a foundation for a module within a university course, or as an educational framework that can be more fully developed, tested, and improved by others. A sample syllabus is offered in Annex 1. Then in chapter 5 we consider how this content can be reconfigured as training outside the academy, including for Horizon Europe applicants.

4.1 Types of students and course objectives

Ethics by Design for AI should be an important component in the curricula of students who pursue degrees in AI. However, as many AI developers have completed tangentially-related programs such as general computer science, data analytics, or even just general programming, an Ethics by Design for AI course could be relevant for students of these programs as well. The same may apply to students of IT law or innovation management.

Additionally, AI is a technology with practical applications in many sectors of society, especially healthcare, finance, mobility, government, law, retail and media. As the development of AI systems may include individuals whose primary expertise is on (applying innovations in) these application areas, it may sometimes be useful to offer an AI Ethics by Design course to these individuals as well.

For all these students, we propose the following course objectives:

- To develop an understanding of the ethical values that may be negatively impacted by AI research and development, as well as the ethical requirements that these values impose on AI research and development.
- To develop an understanding of the SIENNA Ethics by Design methodology by which negative impacts may be averted during the development process of an AI system.

The first objective can be met through a brief introduction into ethics of AI (e.g., key problems, causes and possible solutions), and lectures on key values in the context of using and developing AI systems, and the requirements these values impose on research and development that involve AI. Further familiarisation can be achieved through the use of case studies, where students apply the theoretical knowledge they have gained to detect possible ethical issues in the design and development of specific technologies.

The second objective can be met through lectures on the SIENNA Ethics by Design methodology. Students can use case studies to practise applying the Ethics by Design methodology and adapting it for their own ends and to fit their particular contexts.

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

4.2 Summary of course content

This subsection provides a summary of content on Research Ethics and Ethics by Design for AI suitable for adaptation and application in a number of academic contexts. For instance, it can be integrated as a section of a theoretical or practical ethics component for existing university courses and programmes, or it can be developed into a stand-alone module or unit within those courses or programmes. It can be used as an educational framework or template for further development according to the needs and contexts of those courses and programmes, and it is suitable for application in various disciplines, whether in the humanities, social, or applied sciences. There is sufficient flexibility such that it can be adopted for different levels. For instance, by decreasing the number of readings or stretching the time available for certain topics, the content can be suitable for early or entry level students (e.g. at first or second year undergraduate degree level), or by adding reading or extending the scope of the application and assessment it may be made suitable for higher or honours level courses and modules (e.g. third or final year undergraduate degree level). With the inclusion of increased or more complex assessment, as well as more in depth analysis, it can also serve as a foundation for master's level courses, programmes as well as for units or modules therein.

4.2.1 Introduction

In order to lay the groundwork for thinking about research ethics and Ethics by Design in the context of AI, students need to be introduced to key background information. This includes:

Key problems. These can include the kinds of errors and harms caused by poorly designed or developed AI, including the kinds of very public mistakes that become apparent after it is deployed. This can include well known examples like Tay Chatbot on Twitter[31], or problematic ways in which Google Search distinguishes and classifies different hair types under the categories for "professional hair" and "unprofessional hair".[32]

Key causes. These can include insufficient thought or attention to consequences during design or development, a lack of diverse representation in the workforce, and an insufficient understanding of the inner workings of the design of the AI system, amongst others.

Key solutions. These can include post-development technical solutions, although these tend to mask rather than fix the deeper ethical and social issues. They can also include that ethical requirements are imposed on AI design and development, so that AI systems are properly developed from the ground up. The argument should be made that these solutions often help prevent economically costly negative ethical impacts on the one hand and support productivity in the development process on the other. Examples of global ethical AI initiatives can be offered, and these could include any or all of the following: IEEE, ISO, IETF, WEF, UNESCO, Governments (EG: Singapore, NYC, California, White House, Australia, Denmark), Industry (FATML, XAI, CertNexus, Google, Microsoft, IBM). The approach that we recommend is to think about ethical issues *during* the design phase.

Introduction to Ethics by Design. An introduction to Ethics by Design will need to include:

- *Explanation of terms.* This includes explanation of key terms and concepts, including Ethics by Design, ethical impact, algorithmic bias, etc.

[31] Wakefield, J. (2016). Microsoft chatbot is taught to swear on Twitter. *BBC*. Published 24 March 2016. https://www.bbc.com/news/technology-35890188
[32] Alexander, L. (2016). Do Google's 'unprofessional hair' results show it is racist? *The Guardian*. Published 8 April 2016. https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-

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

- *Context and history*. This includes explanation of how Ethics by design is part of the "values by design" approach and related to other concepts like Privacy by Design, Security by Design.
- *Process for application*. This includes explanation of the fact that Ethics by Design principles can be incorporated into any design methodology.

4.2.2 Values and ethical requisites

Students need to become familiar with key values in the context of using and developing AI systems, and that these values impose requirements on research and development that involves AI.

Outline of key values. It needs to be clear what these key values are as well as their relation to key principles. The values have a solid foundation in that they: (1) are as universal as possible; (2) stand in relation to EU principles and the UN Charter on Human Rights; (3) and are detailed and concrete. It should be explained that unethical AI would be in conflict with such values, and may violate the related principles. Examples can be offered, for instance, regarding human agency, i.e. the *capacity* that humans have to reason, make choices, and act on those choices, and its relation to human dignity, freedom, and autonomy i.e. self-governance, including over one's choices and decisions, and without determination from outside actors. For instance, the possibility to live as one chooses (autonomy) requires not only that a person has the capacity to do so (agency), but also that they are not restricted or coerced (freedom) and that their choices, where they do not harm or unfairly interfere with the choices of others, are respected (dignity). AI that is covert, for example, might interfere with a person's autonomy, because it limits the knowledge from which their full range of choices can be made. This impacts on the extent of their freedom, i.e. to accept or reject the AI, thereby violating their dignity. Especially if they are unaware of the deception and any related coercion and harms, thus limiting their capacity for agency in particular or general terms (depending on the scale of deception). Finally, it needs to be clear that values can lead to "ethical requisites", ethical requirements that are to be imposed on the design and development process.

Categories of values. Explanation and examples should be offered for any or all of the following: (1) Respect for Human Agency; (2) Privacy and Data Governance; (3) Fairness; (4) Individual, Social & Environmental Well-being; (5) Transparency; (6) Accountability and Oversight.

Some detail will need to be provided on these values. Here we offer some abbreviated insight into this content, though it would need to be adapted and developed as appropriate.

Human Agency

Respect for human agency encompasses the values of autonomy, dignity, and freedom, so these should be explained. Autonomy, dignity, and freedom are some of the fundamental rights upon which the EU is founded, and they are enshrined in the UN Declaration of Human Rights.

Respecting autonomy means allowing people scope to decide for themselves what is right and wrong, and the way they should live their lives. Human autonomy can take many forms, since autonomy means that each person can decide for themselves what their personal form of autonomy is. AI systems can restrict human autonomy by restricting human choices or decisions and 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. For instance, dealing with a robot at a hotel reception might limit the type of assistance one can obtain and influence the decision one can make about the stay.

Respecting dignity means that every human being possesses an intrinsic worth, which should be protected. Humans derive dignity from their capacity to determine for themselves what is right and wrong (their autonomy). This means they have the right not to be treated as "a means to an end" (an instrument in the service of other people), but as a unique entity that has an inherent worth.

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

Respecting freedom means leaving people free to exercise their autonomy and live with dignity. Most importantly, freedom requires that individuals have the ability to make their own decisions on matters that are important to them (as long as they do not harm others). Respect for dignity and autonomy entails that no person can be told that an aspect of their freedom is unimportant if the person involved thinks it is. Next to the freedom to act, this includes freedom from constraints that conflict with one's autonomy, such as coercion, deception and manipulation.

Key principles (drawn from SIENNA work,[33] which in turn draws significantly from the HLEG-AI report[34]):

- AI systems should not try to: control people; remove basic freedoms; subordinate, coerce, deceive, manipulate, or dehumanise people; stimulate dependency or addiction. AI applications should be designed to give humans sufficient control of the system.
- 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. For example, the application of AI in sex robots may objectify the individuals the sex robots are supposed to represent, with associated harms.
- It should be clear to people that they are interacting with an AI. They should be informed about the system's abilities and limits, 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.

Privacy and Data Governance

Privacy is an important individual and social right, and AI applications need to recognise, respect, and assure data subjects that privacy is being honoured.

Key principles:

- AI must respect the right to privacy: The collection, use, and control of personal information is a feature of privacy, and where AI collects, produces, communicates, and stores personal information, privacy must be taken into account. Further, privacy is also a way of marking a space where governments and other significant institutional actors are prevented from entering, or are significantly constrained in their actions. Given that AI will play a central role in how personal information is created, shared, stored (by whom and when), the principle of privacy is an essential feature of Ethics by Design for AI.
- AI's use of data must be *actively* governed: Where privacy is concerned, there are two parallel features that are a part of any effective Ethics by Design. First, there must be a series of practices in which the value of privacy is ensured – these must be active practices, in which humans are involved, and given the dynamic nature of AI and personal information, these processes must be constantly revisited and updated. A second essential feature is that people must have assurance that the right to privacy is being honoured. Here, it is essential that people from whom personal information is being collected, etc., and to whom personal information is being given, are assured that those practices receive active and effective oversight.
- In order to meet the EU's guidance on personal information, AI must follow these general principles, as laid out in the EU *Guidance Note on Ethics and Data Protection*:

[33] Resseguier, A., Brey, P., Dainow, B., Drozdzewska, A., Santiago, N. and Wright, D. SIENNA D5.4: Multi-stakeholder strategy and practical tools for ethical AI and robotics (forthcoming).
[34] 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

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

    - Personal data must be pseudonymised or anonymised;
    - Data collection, use, communication and storage must be minimised;
    - Relevant and up-to-date cryptography must be applied;
    - AI must use data-protection focused service providers and storage platforms; and
    - There must be arrangements to enable data subjects to exercise their fundamental rights (e.g.as regards direct access to their personal data and consent to its use or transfer).
- Data usage should be auditable by humans. In order to ensure and assure people that privacy conditions are being met, any data usage by AI must be auditable by humans. For instance, using Model Cards, Datasheets for Datasets, XAI, etc. If this is not possible, it may require that the AI application cannot be used for the given purpose until data usage is auditable.
- *If relevant*, it needs to be clear that an application must explain:
    - *How an individual can withdraw consent.* Withdrawal of consent is a core aspect of research that ensures research participants, the sources, and targets of personal information human agency are respected, and that privacy is maintained. Clear and easily accessible methods to withdraw consent are required. In cases where, at a certain point, it may no longer be possible for individuals to withdraw their personal information, this possibility should be communicated to research participants.
    - *How it will ensure lawfulness, fairness, transparency of data processing.* Given that many jurisdictions will have different specific laws and legal requirements, governance policies of AI will need to have effective means that ensure and assure people how specific laws and legal requirements are being met.
    - *Technical and organisational measures to safeguard the rights of data subjects.* A base assumption of privacy is that data subjects – research participants, people who are the sources of personal information, and/or who are the targets of personal information have rights over how that personal information is used. AI applications that honour privacy need to have both technical and organisational measures that ensure and assure that the rights of those data subjects are being protected.
    - *Strong security measures to prevent data leakages.* Cyber security vulnerabilities pose significant challenges to AI and threaten individual and political notions of privacy – if research participants' personal information is not secure, then privacy is at risk. Cyber security practices need to be constantly updated, and require human training and awareness of human vulnerabilities in addition to technical cyber security measures.

Fairness

Fairness is a constitutive value of liberal democracies and refers to two complementary principles. First, that people are treated in ways that are equivalent, wherever necessary. Second, that the processes by which people are treated, are publicly explicable and amenable to appeal. Fairness is thus concerned with how social institutions regulate the relationships between people.

Insofar as AI will affect the ways that people are treated, there is a strong responsibility to ensure that treatment is fair:

- *People should be given equal rights.* The basic principle here is that all people should have their human rights respected, and all should receive equal treatment unless unequal treatment can be justified from an ethical perspective. AI must be developed such that this principle of equal rights of all people is central to any design, use, or application. AI, and related policies, must also be designed such that the impacts of AI on human rights are monitored, and if there is unjustified unequal treatment, then that AI must be able to be changed or updated.
- *People should be given equal opportunities.* The basic principle here is that all people should have access to, and/or the capacity to particular social goods and services necessary to a minimally

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

decent social life. In the future, AI is likely to play an increasingly important role in determining, facilitating, and limiting access to basic social goods and services, and as with the principle of equal rights, AI, and related policies, must also be designed such that the impacts of AI on equality of opportunity are monitored, and if there are unjustified limitations on opportunity or unjustified differential treatment, then that AI must be able to be changed or updated.

- *People should not be advantaged or disadvantaged undeservedly*. The key point here is that people will receive differential treatments in how they access and use social goods and services, and that any differential treatment must be justified. AI that prevents someone from accessing a particular service due to their race, gender or other (especially protected) characteristics would be unjustified. AI that recognises that people with limited physical mobility may require preferential access to particular buildings would be justified. As before, if and when AI determines differential treatment, there must be some methods to identify and audit why that differential treatment occurred, and if such differential treatment can be justified.
- *Avoidance of algorithmic bias in input data, modelling, algorithm design*. One of the unique features of AI is the way that algorithmic bias can arise without a specific intention to create such biases. Algorithmic bias is essentially differential treatment without justification. The specific challenge of AI and its application is that this algorithmic bias can occur in ways that are hard to detect and harder to explain. This is because, in part at least, the skills needed in input data, modelling, and algorithm design, are heavily specialised and require significant technical expertise.
    - Algorithmic bias is a *specific* concern which needs *specific* mitigation techniques: As different algorithms used in different contexts will have different impacts on people, the mitigation strategies will need to be developed in ways that are specific to the particular biases and contexts.
    - Applications should specify: how to ensure data about people is representative and reflects their diversity; how errors will be avoided in input data; how potential biases will be monitored and identified; how biases will be mitigated; how the algorithmic design will be checked to ensure it does not target certain groups of people unfairly.
- *Universal accessibility*. This draws again from the basic principle of opportunity of access. AI must be designed such that its uses and particular applications are accessible universally. AI systems should be designed so that they are usable by different types of end-users with different abilities. This becomes particularly important for social goods and services that are necessary for basic social activity. AI systems must therefore consider, anticipate, and respond to different end-users, who may have different needs, different abilities, and different capacities.
- *Fair impacts*. This a further feature of fairness, where the impacts of AI must be distributed in ways that are both equal and justifiable. AI must be designed, developed, and overseen such that particular AI applications do not distribute benefits and burdens unequally. There may be situations where differential distribution of benefits and burdens occur, and if so, such differential distribution requires (public) justifications.
    - For instance, evidence that possible negative social impacts on certain groups must be considered in advance, monitored, and responded to.
    - Steps to ensure the system does not unjustifiably discriminate – or cause others to unjustifiably discriminate – can be outlined here.

Individual, Social & Environmental Well-being

AI systems should not harm individual, social, or environmental well-being. Well-being, whether understood in reference to individuals, society, or environmental, is a fundamentally important ethical value, and AI systems should, as a general rule, be designed such that they do not cause harm to, or degrade, the well-being of individuals, societies, or the environment.

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

- *AI systems should consider the welfare of all stakeholders*. A thorough and comprehensive ethical risk analysis of AI needs to take into account all stakeholders who may be affected directly or indirectly by AI. This needs to occur before AI is rolled out (i.e. in the R&D stages), when it is used, and after the event. Stakeholders need to be identified, and wherever possible, the stakeholders or relevant representatives must be consulted with throughout the lifecycle of the AI product or service.
- *Documented efforts to consider environmental impact.* Any consideration of the environmental impacts must be documented and publicly accessible, wherever possible.
- *If needed, steps to mitigate negative impacts.* If and when negative impacts on well-being are anticipated or observed, there must be systems and policies in place to engage in mitigation efforts to reduce the negative impacts.
- *Consider and mitigate harm to online communications.* Given the fundamental importance of information technologies to communication and the individual and social well-being that comes from that, new AI products and systems must be designed and overseen such that these products and systems do not further degrade online communications. This is particularly in reference to phenomena such as fake news, deepfakes, filter bubbles, echo-chambers, political manipulation of individual citizens, and the degradation of democratic institutions.
- *Not reduce safety in the workplace.* AI must also be designed and implemented in such a way as to ensure that each individual's physical safety and psychological well-being are attended to and protected in the workplace. AI development and implementation must take into account existing standards for safety, for instance: IEEE P1228 (Standard for Software Safety).

Accountability & Oversight

Essential to ethical AI is the principle that someone can be held to account for problems that arise when AI is involved in something of ethical importance. This is the basic premise of traceability. Parallel with accountability is the need for oversight. Oversight refers to the formalisation of a set of processes where AI, its R&D, implementation, use, and governance are observed and recorded in relevant ways. Oversight and accountability are the default settings: Unless compelling reasons are provided, methods and means of accountability and oversight must be included in any AI R&D process.

- *Accountability*. This principle ensures that people who build, operate, or are institutionally involved in AI, its uses, and outcomes can be held responsible for the actions and/or effect of AI.
- *Oversight.* In order for oversight to be functionally useful, AI applications require that humans can understand, supervise, and control design and operation of AI. The formalisation of oversight should include procedures for risk assessment, and mitigation efforts must be documented. People, particularly users or members of communities affected or likely to be affected by particular AI applications must be able to report concerns. Such reporting mechanisms must be usable and accessible. The reporting mechanisms must also detail how such external concerns or complaints will be evaluated and mitigating actions taken.
- *Transparency.* This principle underpins effective accountability and oversight: Developers must be able to explain how and why a system acts the way it does.
- *Anticipatory accountability.* As a prerequisite, applications for the development, use, and public release of AI must explain how undesirable effects will be detected, stopped, and prevented from reoccurring. This is a cornerstone of an accountability process.
- *Ethical risk assessments.* Increased ethical risks will increase the need for formal ethical risk assessments: Not all AI applications have the potential to generate serious ethical concerns, and some are more ethically risky than others. Drawing from the requirement of anticipatory accountability, if a preliminary ethical analysis suggests particular risks, and those risks are significant, then a formal ethical risk analysis may be required.

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

- *Independent Audits*. All AI systems should be auditable by independent third parties. This includes the development process by which it was created. Audits must also be able to offer justificatory explanations for all relevant decisions such that third parties can determine not just what was done, but *why* it was done.

<u>Transparency</u>

At its core, AI is being used to aid people in different ways. However justifiable that goal is, people must also be able to understand and assess how AI is operating. Not only do AI systems need to *ensure* that key ethical values are observed and respected, but people need to be *assured* that these processes are functioning effectively, and if or when this may not be occurring, people need to be able to appeal against AI decisions.[35] Transparency is fundamental to achieving these ends.

Key principles:

- *Humans must be able to understand 'how'.* This includes how the AI functions, how particular decisions are reached, and which aspects of those decisions are amenable to review.
- *Enables human agency, data governance, accountability, human governance*. Transparency is a way to ensure that human agency is central to any AI systems and decisions. Transparency enables governance processes to function, ensures accountability for decisions, and ultimately keeps the human as central to governance of AI more generally.
- *This requirement for transparency applies to* all *elements of the AI.* Data, functionality, processes by which it is designed, deployed and operated, and how particular decisions have been made. If – as with particular forms of AI – explicability is not possible, then a set of policies that explain and justify why such opaque AI has been used in the given situation, and clear lines of *human* accountability must be identified.
- *This report holds that the best practice AI is eXplainable AI "XAI"*. Wherever possible, AI decisions should be explainable. For particular contexts in which such decisions have significant ethical importance, such as a decision to use lethal force against a human, this AI *must* be explicable. Opaque or 'black-box' AI should not be used in situations of significant ethical weight: taking people's lives, decisions about who gets medical treatment that will save their lives, or the provision of/access to other primary social goods.
- *Enable traceability of the AI system during its entire lifecycle.* Traceability refers to an audit trail by which any decisions to install or use AI can be traced back to a particular person or people, and that they may be morally responsible for the outcomes that that AI produces.
- *It must be clear to end-users that they are interacting with an AI.* In order for human agency to be respected, it must be transparent to users if and when they are interacting with an AI.

To operationalise transparency, and assure people understand how it is being used, AI must engage in open communication of:

- AI's purpose
- AI's capabilities & limitations
- The benefits & risks of the AI
- The decisions made by the AI
- Governance processes
- Records must be kept of decisions about ethics made during construction

[35] The connection between transparency, and the need to both ensure and assure people that key values are observed and respected is described in Robbins, S., and Henschke, A. (2017). The Value of Transparency: Bulk Data and Authoritarianism. In *Surveillance And Society*, 15: 3/4.

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

4.2.3 The generic AI development model and ethical guidelines

Having explained the key values and associated ethical requisites for AI research and development, it should be explained to students how these ethical requisites may be mapped onto research and/or development processes as concrete ethical guidelines. Whereas ethical requisites are concerned with the final characteristics of an AI system, ethical guidelines focus on the steps by which the AI system is created, with a view to ensuring that each step is ethical.

A variety of development methodologies are used in AI projects, examples being Agile, CRISP-DM and V-model. They are distinguished by the manner in which they organise the development process. However, all have at least some level of similarity with the generic design model for AI as presented in the Ethics by Design approach of the SIENNA project (and described earlier in chapter 2 of this report):[36]

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.* This is the process of collecting, verifying, cleaning, formatting and integrating the data.
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 the system, including an evaluation against the original objectives and requirements.

This model can be used to illustrate how ethical requisites can be mapped onto the different phases of a particular development process to create ethical guidelines for each of these phases. Following the guidelines (usually formulated as tasks) at each of the phases ensures that the final system instantiates the ethical requisites and therefore does not violate any ethical values. Developers can create their own mapping of the ethical requisites on a slightly different development process by adapting and expanding the following example.

Design Phase

Privacy & Data Governance:

- Check whether the objectives are compatible with the privacy and data governance requirements.
- Non-adherence to any of these would result in serious non-compliance.
- Assess whether the plans for what data will be used are fair and appropriate.
- If the proposed data source is unfair or inappropriate, either change the data source or modify the objective so that that data source is not needed.

Specification of requirements

[36] SIENNA D5.4, op cit.

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

Build Ethics by Design into the project and create an Ethics by Design implementation plan which specifies:

- How Ethics by Design will be embedded in the development process.
- People responsible for actions and monitoring.
- Design an ethical compliance architecture.
- Tools.
- Organisational structures and procedures.

High-level Design

Fairness:

- Undertake an accessibility assessment of the interface.
- Ensure that the system meets relevant accessibility standards.
- Transparency
- Design mechanisms by which the AI system will record its own decisions in a way humans can review them.

Individual, and Social and Environmental Well-being

Demonstrate (and evaluate) how the system:

- Will be constructed in an environmentally friendly way.
- Whether the system could cause physical harm to people, animals the environment.
- Includes design features to minimise harms.

Data collection and preparation

Fairness:

- Make sure data from one demographics group is not used to represent another unless it is justifiably representative
- Transparency
- Ensure that it can be explained how personal data is used, shared, and stored.
- Accountability & Oversight
- Make sure it is clearly established what kind of sample is needed, what kind of sample has been taken, and that it can be explained what it will be used for.

Detailed design and development

Fairness:

- Check for algorithmic bias, particularly computational bias, once data commences to be processed
- Privacy & Data Governance
- Make sure that roles and responsibilities are clear for governance and management of data assets and that all relevant staff understand them
- Transparency
- 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

Testing and evaluation

Transparency:

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

- Test whether users understand that they are interacting with an AI.
- Accountability & Oversight
- Develop and deliver training to users to help develop accountability practices (including teaching about the legal framework applicable to the system).
- Whenever possible, ensure practical processes exist for third parties (e.g. suppliers, consumers, distributors/vendors) or employees to report potential vulnerabilities, risks, or biases in the system. Ensure mechanisms exist to examine and act upon such reports.

4.2.4 Research ethics

In the previous section, it was explained how the key values and associated ethical requisites presented in subsection 4.2.2 may be used during the design and development of AI systems. These values and requisites, however, also apply to undertaking research in the area of AI.

In addition to these values, it should be explained to students that there are additional requisites that need to be taken into account for those involved in AI research. These are requisites that apply to R&D more generally, and that apply to digital technologies more generally. They include:

- 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.

4.2.5 Student practice and utilising case studies

In this section, we offer an outline of case study use in the module by focusing on detailed design and development of AI and robotics. This method allows students to apply the theoretical knowledge they have gained to detect possible ethical issues in the design and development of specific technologies, and to consider concrete steps for eliminating or mitigating them by employing the Ethics by Design approach.

<u>Conducting audits</u>

Students can be asked to conduct audits of established algorithms containing significant ethical shortcomings, e.g. Equivant's (formerly Northpointe) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is employed by a number of U.S. courts to predict a defendant's likelihood of recidivism.

This particular case study would bring to the fore issues of fairness, privacy and data governance, transparency, and accountability. Specifically, students could be asked to check the algorithm for biases, including computational biases, training data biases, algorithmic focus biases, algorithmic processing

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

biases, transfer context biases, and interpretation biases,[37] and identify the nature of such biases. Example auditing programs could be described, e.g. Fairtest,[38] FairML,[39] and three naïve Bayes approaches.[40]

As for issues of privacy and data governance, students could be asked to investigate what kind of data is used by the algorithm and which measures—if any—are in place to protect the identity of their owners, as well as to prevent data leaks.

Additionally, students could be asked to check whether the developers are transparent about the design and implementation of the algorithm, as well as to investigate the extent to which developers, users and additional parties involved are able to understand the functioning of the algorithm.

Furthermore, students could be asked to analyse the algorithm's impact in terms of well-being, e.g. by employing an environmental perspective and investigating how the system affects the non-human world, where the data is stored, how the servers are maintained and whether they use large amounts of (scarce) natural resources.

By encouraging students to examine algorithms in order to identify potential ethical issues, they would gain first-hand experience regarding the ethical implications of the design and development of technology that does not conform to Ethics by Design principles and guidelines. Additionally, encouraging students to critically evaluate algorithms and examine them for biases and ethical issues would enhance students' critical sense and ethical awareness, as well as a familiarity with employment of Ethics by Design, increasing the likelihood that they will employ these methods in their own design and development planning.

4.3 Teaching methods

For a fruitful Ethics by Design course, instructors should be mindful of the background and expertise of the participants and tailor the course around them,[41] e.g. by applying different methodologies for the teaching of participants with backgrounds in computer science and engineering, as opposed to social sciences and humanities, including ethics and philosophy.

For instance, it may be worth bearing in mind that computer scientists tend to be application-first thinkers and rely on inductive methods. This may lead to a series of possible issues for teachers, in that such students may be resistant to inputs and feedback coming from instructors who both lack a technical background and employ significantly different methodologies from the ones usually employed in computer science. They may underrate the relevance of Ethics by Design for their own work as a result of these tendencies and preferences. Additionally, depending on their background and the type of work they usually undertake, students with a background in computer science may struggle to accept the grey areas of certain issues uncovered by Ethics by Design, especially if they are used to black-and-white scenarios.

---

[37] Danks, D., and London, A. J. (2017). Algorithmic bias in autonomous systems, *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI),* pp. 4691-4697.
[38] Tramèr, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J. P., Humbert, M., Juels, A., and Lin, H. (2017). Fairtest: Discovering unwarranted associations in data-driven applications, *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 401-416.
[39] Adebayo, J. A. (2016). FairML: ToolBox for diagnosing bias in predictive modeling, Doctoral dissertation, Massachusetts Institute of Technology, retrievable from https://dspace.mit.edu/handle/1721.1/108212.
[40] Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification, *Data Mining and Knowledge Discovery,* 21: 2, pp. 277-292.
[41] Meyer, E. (2014). *The Culture Map: Breaking through the Invisible Boundaries of Global Business*, New York, PublicAffairs.

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

In order to overcome such issues, teachers could approach Ethics by Design from a pragmatic perspective, e.g. by asking students to design an artefact and—either along the way or after they completed the design, depending on whether one prefers to design the artefact together as a group or let the participants design it and subsequently analyse it together—highlight how the design is value-laden, how it may contain ethical implications, e.g. unintended biases, and how one might prevent such biases. In fact, students may be more likely to listen to the principles and apply them in real life if they have not been fed to them, but they have realised autonomously that they should be an unavoidable part of the design and development processes which should not be overlooked. Additionally, for a fruitful course, teachers could

1) either prove to their students that they have a solid understanding of the technicalities behind their work, or ask them to show how they work,
2) highlight the ethical issues that may arise from design methodologies that do not satisfy Ethics by Design principles and guidelines, and
3) stress the importance of applying Ethics by Design principles and guidelines in the design process even if most ethical issues will not be fully solvable and, as a result, the design of the technology will likely always be somewhat imperfect from an ethical standpoint.

Lastly, teachers should also be mindful of the fact that social scientists and philosophers do not always have first-hand experience of the design process from a technical perspective such that they are more likely to apply Ethics by Design principles in multidisciplinary settings, working alongside computer scientists. In this sense, it is vital that such students learn how to successfully approach and discuss Ethics by Design issues with people coming from different backgrounds. In particular, as philosophers may become tech ethicists, it is particularly important to make sure that they can gain the necessary skills to learn how to convey the significance of Ethics by Design to non-ethicists, be they future students or co-workers.

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

5. Research Ethics and Ethics by Design for AI in company training

In order for Ethics by Design to be effectively applied to AI, key stakeholders outside the academy, like users, designers, policy makers, and so on, must be educated in the general principles of Ethics by Design, and how it relates to AI. For such education and training to be worthwhile for companies and professionals – it needs to offer competitive advantages. 'If being an ethically sound AI product increases market opportunities, while not being ethically sound harms sales, then businesses are more likely to ethical requirements by choice' [42]

Central to this is a certification program. Such programs are common in the private sector – for instance:

> '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' [43]

We suggest that the main aims of Ethics by Design for AI are integrated into existing education and training services, such as within continuing and professional development programmes. Just as vendors will ideally:

> 'adopt ethical AI certification voluntarily, [the] 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' [44]

Such certification can cover the AI systems themselves, the people using them, and the training programs. Furthermore, it is a requirement of most existing certification schemes that the company is certified, especially so as to deliver training programmes effectively. This section of the report looks at how to educate and train Ethics by Design for AI to people working in the private sector, and the training programs that could lead to certification.

5.1 Summary of course content

the private sector typically requires that people who deliver training, including on ethics, have sufficient understanding of the products and applications with which they are working or providing to others. This includes sufficient knowledge and understanding such that they can recognise the effects that these products and applications will have on others. When considering Ethics by Design for AI, this means that those working with relevant AI products and applications may need some education and training beyond how a particular AI product or application works, but also to include who is likely to use it, what normal uses might involve, and to be able to anticipate innovative uses. Moreover, for the private sector, they need to be able to recognise the ethical aspects of AI. A deep understanding of the principles of Ethics by Design is useful but not necessary – instead, education and training for ethical AI can give a more specified

[42] SIENNA D5.4, op cit, p. 87.

[43] Ibid.

[44] Ibid.

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

and granular awareness of how a particular AI product or application relates to specific ethical issues or principles. The education and training needs to be tailored and specific:

> '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.'[45]

We therefore offer here three central themes to AI education and training: AI and the Product/Application, AI and the People, and AI, Values and Principles.

5.1.1 AI and the Product or Application

When considering practical applications of Ethics by Design for AI, education and training needs to be developed and delivered in a way that is relevant and specific to the particular professionals, and that is going to vary depending on the specific AI *Product or Application*. For instance, the ethics of AI products and applications involved in facial recognition technology (FRT) for border control will be different from the ethics of AI products and applications involved in delivering care to elderly patients, and different from the ethics of AI products and applications involved in autonomous vehicles used in underwater mining. The point is that it is not going to be useful to deliver education and training on the ethics of FRT for border control to professionals engaged in elder care, or to those working on underwater mining. Thus, education and training for professionals must be as bespoke as possible, tailored to the specific *product* that the professionals are involved in. Second, as different AI products and applications will be used in different ways, they will impact people differently. The education and training must then effectively integrate what the product does with the specific ethical issues (as outlined in 5.1.3 below) and do so in a way that is relevant and applicable to the particular professionals.

5.1.2 AI and the People

Second, education and training must include consideration of the *People* who are likely to be impacted by the product. Professionals involved at all levels of AI product and application development, marketing, sales, service, and oversight need to recognise that products will impact people in ways that might be positive, but might also cause suffering or unhappiness, and could impact their rights in important ways. Education and training must therefore include coverage of the key stakeholders who are likely to be affected or impacted by the AI product and applications. This needs to include both how they may be affected, and options to remove or minimise unwanted or harmful impacts. Formal processes like user impact assessments, stakeholder analysis, and so on, can be useful tools here. A key point of education and training is thus to show participants how AI products and applications will impact people's lives.

5.1.3 AI, Values and Principles

The final segment of education and training is focussed on *Values and Principles*. Having looked at what AI products and applications will do, who they will likely impact, we must also illuminate *why it matters*. This is where ethics of AI becomes essential – it is not enough to say to professionals that a particular product might be gathering personal information about people, but to show them how that information

---

[45] SIENNA D5.4, op cit, p. 94.

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

impacts key values, and perhaps violates a principle, and why it matters. Many AI products and applications will interact with a range of ethical principles, and education and training must effectively explain this with sufficient depth and detail. This includes identifying those principles, explaining to participants why they matter in a way that is relevant and compelling to them, and then offering some guidance as to how they could navigate situations where different principles are in tension. The purpose here is not to give recipients of education and training full courses in Ethics by Design for AI, but to make them aware that particular products have ethical importance, and to give them some capacity to navigate the relevant terrain.

5.2 Teaching methods

This approach to education and training draws from the recognition that professionals likely have significant knowledge and lived experience of the products and applications, and typically have encountered challenges to do with at least some of the ethical principles. Education and training must therefore proceed in a way that draws from, and is responsive to, the professional's knowledge and experience. Effective education and training must therefore seek to understand and recognise the likely background knowledge and experience of the participants. In parallel, however, such education and training should be deeply tied to the ideas of Ethics by Design. As such, those involved in the delivery of education and training ideally should have some experience of the technology themselves, as well as a comprehensive and nuanced understanding of Ethics by Design.

It may be hard to find individuals who meet all the desired criteria, so co-delivery methods are encouraged. Where the education and training is delivered by a team, it is ideal if some team members have relevant and lived experience that will be applicable to, and recognised by, the participants, while other members bring the deep knowledge of ethical theory and Ethics by Design. For team teaching to be effective, each member of the team must be familiar with the skills, knowledge, and experience of other team members, or at least willing to learn and collaborate. That is, all members of the team do not need to have a deep knowledge of ethical theory, but those from a practitioner background need to be aware of, and appreciative of, the Ethics by Design approach. Likewise, the teams do not need everyone to be experienced in particular AI products or processes, but those skilled in Ethics by Design need to be aware of, and appreciative of, the knowledge and expertise of AI practitioners.

The certification of such education and training is essential to incentivise uptake, and this process requires some level of oversight and quality control:

> '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.'[46]

Finally, any such education and teaching needs to be grounded both in sufficiently universal principles, while simultaneously being bespoke enough such that it meaningful applies within the specific context in which it is taught. As we saw in chapter two, Ethics by Design has a significantly universal foundation, in that the core ethical principles undergirding it are typically thought to be universal. For instance, a human

[46] SIENNA D5.4, op cit, p. 97.

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

right like privacy is considered important for every person, regardless of their location. At the same time, however, this education and training needs to be developed, and delivered, in a way that is tailored to the specific recipients. Again, a program on FRT will need to be different from one on undersea autonomous vehicles. Moreover, recipients will need to have different modes and forms of delivery. One set of recipients might desire or benefit from a two hour introduction to the ethical challenges of FRT, while another group might need a two week intensive course on how to design particular principles into AI that supports elder care. Accordingly the education and training for these different courses will need to differ in some respects. Similarly, some recipients may need accreditation as a result of undergoing a particular course. Short courses that supplement the content with micro-credentialling will require some form of assessment, and may need to be delivered by a registered education provider. To ensure the quality of the delivered products and services, particularly where accreditation is involved, educational designers and/or those with significant experience in course design and delivery must be involved. The key point here is that education and training in Ethics by Design for professionals is likely to be highly variable, but this variability must not lead to a decline in quality of the education and training that is delivered.

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

6. Conclusion

This report has sought to outline, develop, and promote an Ethics by Design approach that builds on the work undertaken in both the SIENNA and SHERPA projects. It offers approaches to teaching Ethics by Design both in and outside the academy and provides a range of tools and methods to facilitate that practice. In so doing, the report acknowledges that there may be some obstacles and challenges in providing this training and presents some methods to help mitigate those difficulties.

As we note in the Introduction, piloting of these materials has already begun and will continue beyond the lifetime of the SHERPA project. We therefore expect the materials provided in this report to be amended and developed as the field broadens, as interest grows, and the approach becomes well known. This is especially likely given the EC's promotion of an Ethics by Design approach for applicants of Horizon Europe funding programmes, especially for AI.

If the materials in this report help to provide a foundation for training plans within the academy, for political and funding institutes like the EC, and within industry, then key expectations for this endeavour will have been met. The same applies if an Ethics by Design training programme can be offered commercially through the non-profit spin-out company ORBIT.

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

Annex 1. Research Ethics and Ethics by Design for AI and Robotics: Course Description/Indicative Syllabus

GENERAL INFORMATION

Course: Ethics by Design for AI and Robotics

Example credits: European Credit Transfer and Accumulation System (ECTS-credits): 5 to 10

Prior knowledge: *depends on level / course / programme*

Instructors: *to be added*

AIMS

The course aims to provide students with the theoretical, practical and methodological tools needed to approach the ethics of AI and robotics in a comprehensive and systematic manner. It aims to train students to recognise ethical issues in AI and robotics in all stages of the technological process, from its inception and design to its implementation, and encourages them to engage with such issues by applying Ethics by Design principles and methods.

CONTENT

The course is structured around three main modules:

- Module 1: Recognising Ethics by Design (weeks 1-4);
- Module 2: Analysing Ethics by Design (weeks 5-7);
- Module 3: Applying Ethics by Design (weeks 8-10).

Week	Topic	Content	Assignment
Week 1	Introduction to the ethics of AI and robotics	The main goal of this class is to reflect on the notion of AI and robot ethics. Sample material:	

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

		Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In W. M. Ramsey and K. Frankish (Eds.), *The Cambridge Handbook of Artificial Intelligence*, 316-334. Cambridge: Cambridge University Press. Boer, D. (2015). The robot's dilemma. Working out how to build ethical robots is one of the thorniest challenges in artificial intelligence. *Nature*, *523*, 24-26.	
Week 2	Inception and design of AI and robotics: Ethical issues	The main goal of this class is to reflect on the ethical issues that arise during the inception and design stage of the technology. Sample material: Guersenzvaig, A. (2021). *The Goods of Design: Professional Ethics for Designers*. Lanham, Boulder, New York, London: Rowman & Littlefield. Van den Hoven, J., Vermaas, P. E., & Van de Poel, I. (2015). *Handbook of ethics, values and technological design.* Dordrecht, Heidelberg, New York, London: Springer.	Assignment: Choose an artefact and write a 2-pager on the ethical issue that may arise during the inception and design of the technology
Week 3	Development of AI and robotics: Ethical issues	The main goal of this class is to reflect on the ethical issues that arise during the development stage of the technology. Sample material:	Assignment: Choose an artefact and write a 2-pager on the ethical issue that may arise during the development of the technology

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

		Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press. Luca, M., Kleinberg, J., & Mullainathan, S. (2016). Algorithms need managers, too. *Harvard Business Review*, *94*(1), 20.	
Week 4	Implementation of AI and robotics: Ethical issues	The main goal of this class is to reflect on the ethical issues that arise during the implementation stage of the technology. Sample material: Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*. Stahl, B. C., & Wright, D. (2018). Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, *16*(3), 26-33.	Assignment: Choose an artefact and write a 2-pager on the ethical issue that may arise during the implementation of the technology
Week 5	Ethics by Design principles	The main goal of this class is to present SIENNA's 5-layer model of Ethics by Design: • Ethics by Design values; • Ethical requisites; • Ethics by Design guidelines; • AI methodologies; • Tools & methods.	

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

Week 6	Ethics by Design models	The main goal of this class is to present SIENNA's model for the ethical development of AI and robotics: • Specification of objectives; • Specification of requirements; • High-level design; • Data collection and preparation; • Detailed design and development; • Testing and evaluation.	
Week 7	Ethics by Design guidelines	The main goal of this class is to present SIENNA's guidelines for the ethical deployment and use of AI and robotics, pertaining to the stages of (1) planning and management, (2) acquisition, (3) deployment and implementation, (4) monitoring.	
Week 8	Applying Ethics by Design	The main goal of this class is to start applying Ethics by Design principles, methods and guidelines to a technology in use, by starting to investigate the ethics of a technological artefact in AI or robotics and writing a case study.	Assignment: Choose an artefact and write an overview of its ethical issues
Week 9	Case study	The main goal of this class is to further develop the W7 assignment into a full case study on the ethical issues related to the design, development and implementation of the technological artefact of choice, by employing Ethics by Design principle, methods and guidelines.	Case study: Using W2-4 assignments as a starting point, write a systematic and comprehensive analysis of the ethical issues related to the design, development and implementation of an artefact of your choice, and explain how Ethics by Design can help prevent or mitigate such issues.
Week 10	Presentation & discussion of the case studies	The main goal of this class is to present and discuss student case studies.	

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

INSTRUCTIONAL MODES

Interactive lectures or lectures, seminars, small group work, tutorials.

Presence duty: yes

TESTS

Written assignments or essays, case studies, presentations, group projects

PREREQUISITES

None

MATERIALS

Links to required literature will be made available online in advance.

INDICATIVE BIBLIOGRAPHY AND RECOMMENDED FURTHER READING

Coeckelbergh, M. (2020). *AI ethics*. Cambridge, London: MIT Press.

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics*, *12*(4), 701-715.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., ... & de Wildt, T. (2018). Ethics by design: Necessity or curse?. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 60-66.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Cambridge: Mit Press.

Friedman, B., & Kahn Jr, P. H. (2007). Human values, ethics, and design. In J. A. Jacko (Ed.), *The human-computer interaction handbook*, 1267-1292. Boca Raton: CRC press.

Iphofen, R., & Kritikos, M. (2021). Regulating artificial intelligence and robotics: Ethics by design in a digital society. *Contemporary Social Science*, *16*(2), 170-184.

Lin, P., Abney, K., & Bekey, G. A. (2012). *Robot ethics: the ethical and social implications of robotics*. Cambridge, London: MIT Press.

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243-256.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

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

Shilton, K. (2013). Values levers: Building ethics into design. *Science, Technology, & Human Values*, *38*(3), 374-397.

Van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and engineering ethics*, *19*(2), 407-433.

Van Wynsberghe, A., & Robbins, S. (2014). Ethicist as designer: A pragmatic approach to ethics in the lab. *Science and engineering ethics*, *20*(4), 947-961.

Verbeek, P. P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, *31*(3), 361-380.

Verbeek, P. P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In *Philosophy and design*, 91-103. Dordrecht: Springer.

Verbeek, P. P. (2011). *Moralizing technology*. Chicago: University of Chicago Press.

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

Annex 2. Research Ethics and Ethics by Design for AI and Robotics: Training Description

ABOUT THE COURSE

Artificial intelligence (AI) already has a huge impact on society and the economy. This impact is expected to grow rapidly. AI technologies have many advantages and can open new avenues for thinking about and engaging with the world. The potential of AI for good practical and moral outcomes is counterbalanced by growing concerns about ethical and human rights implications of some applications and approaches.

In response to this, the EU and other policy bodies are developing various policies and regulations. Within the Horizon Europe research framework programme, for example, AI is now a topic to be assessed during ethics reviews. Researchers, developers and users of AI will be asked to clarify how they will address these issues.[47] Ethics by Design is one such approach promoted as a way to ensure that ethical considerations are considered comprehensively and systematically during the design and development processes. It is now included in Horizon Europe ethics review as the cornerstone of its new AI ethics requirement, and it is also gaining influence in research ethics for AI more broadly.

But how do you *do* Ethics by Design? This training course provides an overview of the concept, principles, and implementation of Ethics by Design. This includes an introduction of some procedures for applying Ethics by Design approaches to research and development processes in Artificial Intelligence and robotics. It is designed by members of the SHERPA project who played an important role in shaping the approach adopted by the EC. If you want to understand what Ethics by Design is and how it can be done, then this course is for you.

OBJECTIVES

The course aims to provide participants with the tools needed to recognise, analyse and apply Ethics by Design approaches to the development of AI and robotics. This will include an outline of the key ethical values upon which Ethics by Design is based, the characteristics any AI system should have in order to exhibit these values, the tasks required in order to produce an ethical AI, and methods by which to demonstrate compliance.

This training will enable participants to identify concrete ethical issues and implementation processes that meet the requirements for ethical AI and robotics from the initial inception of an idea through to the final design and production of a technology.

[47] Cf. § 3.8 Development, deployment and use of Artificial Intelligence (AI) and other new and emerging technologies in the EC's recently published guidance for Horizon Europe applicants entitled 'Identifying serious and complex ethics issues in EU-funded research', which builds on guidance developed in SIENNA: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/guidelines-on-serious-and-complex-cases_he_en.pdf

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

METHODOLOGY

This course will rely on active learning techniques to create an engaging learning environment capable of enhancing participants' performance and facilitating higher-level thinking skills.[48] In particular, participants will be presented with a combination of interactive lectures and seminars, short presentations and discussion points.

In additionally, participants will be asked to apply Ethics by Design to case studies related to either established AI systems containing significant ethical issues, e.g. Equivant's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, or AI systems they may have encountered in the workplace. Such exercises will help participants to apply the theoretical and methodological principles of Ethics by Design through practice, thereby learning how to detect possible ethical issues in the design and development of the technology. Issues could include those related to fairness, privacy, data governance, transparency, and accountability. Methods to avoid as well as eliminate or mitigate these issues by employing an Ethics by Design approach will also be covered.

PREREQUISITES

No technical background is required to attend this course.

WHO IS THIS COURSE FOR?

This training course is primarily aimed towards researchers, developers, and technicians planning to develop or deploy AI and robotic technologies in Europe, as well as other stakeholders involved in proposal development and project delivery, such as university research administrators or consultants.

INSTRUCTORS

To be added.

SAMPLE AGENDA:

09:30-09:45	Welcome and introduction
09:45-11:15	Introduction to Ethics by Design: Recognising ethical issues during the inception, design, development, and implementation of AI and robotics
11:15-11:30	Break
11:30-12:00	Q&A and discussion
12:00-13:00	Lunch break

[48] Miller, C. J., McNear, J., and Metz, M. J. (2013). A comparison of traditional and engaging lecture methods in a large, professional-level course. *Advances in physiology education*, 37: 4, pp. 347-355.

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

13:00-14:30	Analysing Ethics by Design: EbD principles, models, and guidelines
14:30-14:45	Q&A and discussion
14:45-15:00	Break
15:00-16:30	Applying Ethics by Design to AI and robotics: Case studies
16:30-17:00	Q&A and discussion
17:15-17:30	Closing

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

Annex 3. Existing courses on Ethics by Design

Courses on Ethics by Design

- Artificial Intelligence and Ethics in Design (IEEE) [intended for industry professionals focused on integrating artificial intelligence and autonomous systems within their companies or to their customers and end users]
- Digital Ethics by Design - Executive Masterclass (TU Delft) [This masterclass is primarily aimed towards experienced mid-level to senior-level managers in companies working on innovations in the digital domain. It is also suitable for people working at a similar level for governmental and non-profit organisations. Academic researchers and (PhD) students will not be admitted to this masterclass.]
- Ethics by Design Workshop Series: Overview & Human Rights (General Assembly) [anyone, no prerequisites]
- Ethics by Design Workshop Series: Ethics of Algorithms & Artificial intelligence (General Assembly) [anyone, no prerequisites]
- Ethics by Design Workshop Series: Democracy, Privacy & Technology (General Assembly) [anyone, no prerequisites]

Courses on AI ethics which include or reference Ethics by Design topics

- AI, Ethics, and Society (Georgia Tech College of Computing), 'Fairness in AI/ML - Objective: After completing this module, students will be able to understand and apply basic AI/ML techniques to data scenarios, with a focus on identifying fairness and bias issues found in the design of decision-making systems' 'Bias Mitigation and Future Opportunities - Objective: After completing this module, students will be able to utilize tools and methods to quantify bias and examine ways to use algorithmic fairness to mitigate this bias, taking into consideration ethical and legal issues associated with it' 'You will be able to practice your learned knowledge by writing coherent and well- structured critiques of situations and papers, leading and participating in class discussions, and designing your own algorithmic solutions' [computer science]
- AI Safety, Ethics, and Policy (Columbia University), 'Fairness, bias, and inequality' [computer science]
- Artificial Intelligence Ethics in Action (LearnQuest, Coursera) 'Week 3 project - Game Theory Algorithm Design: Predicting ethical issues before they arise in model theory' [data science]
- Design for Artificial Intelligence (IED), 'Based on the aim of facilitating its use by people and companies, the course is aimed at optimising the conceptualisation, design and development of new products and services and ensuring an ethical approach in all phases of the process' [computer science]
- Ethical and Social Issues in AI (Cornell, 2017, apparently discontinued), Among the topics: 'Inherent trade-offs in algorithmic fairness' 'Computational ethics for AI' [computer science]

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

- Ethics, Privacy, AI in Society (Imperial College London), 'Incorporate ethical principles of the key ethical frameworks into the design of artificial agents, according to standard methodologies' [engineering, computer science]

- Ethics for AI and Robotics (University of Michigan), 'As we design intelligent artifacts that make their own decisions about how to act, and as they act within the human world, we ask how we can ensure that they will act ethically' 'Do we mean that humans must be ethical as we design and deploy intelligent systems? Do we mean that the systems we design and deploy must be capable of deciding what is ethical for them to do? Most likely, the answers to both questions will turn out to be "Yes!" The follow-on question is "How do we do that?". The semester will be organized around seven major topic areas: (1) Safety, (2) Background, (3) Trust, cooperation, and society (4) Bias and fairness (5) Surveillance and privacy (6) Trust for corporate entities: corporations, governments (7) Existential risk' [computer science]

- Ethics for Engineers: Artificial Intelligence (MIT), 'In this weekly seminar we will inquire into the problems of ethics by starting with engineering cases that raise ethical quandaries and bringing to light the deep issues that underlie them. […] This course takes the approach that engineering is not merely about design and implementation. Engineering aims at goods for both the individual and for society, and the thoughtful pursuit of engineering necessitates an understanding of those goods and therein an understanding of both the individual and of society.' [engineering]

- Ethics in Artificial Intelligence (University of Bologna), 'Machine Learning, Big Data and the issues of Bias and Discrimination', 'AI Explainability and Transparency' [computer science, philosophy of law]

- Ethics of AI (London School of Economics in collaboration with GetSmarter), 'On completion of this course, you'll walk away with: [...] An understanding of the roles and obligations of governments and multinational corporations in responsibly designing and deploying AI' [social sciences]

- Ethics of Artificial Intelligence (Linköping University), 'The course focuses on three main areas of moral relevance for autonomous systems and AI: responsibility for decisions made by artificial agents, bias/discrimination as a result of AI use, and the importance of participation in the development of AI systems.' [applied ethics]

- Ethics of Artificial Intelligence (Politecnico di Milano), 'Week 1 - Describe the reasons for an ethical analysis applied to AI. Recognize how the notion of responsibility is challenged when designing and using AI tools' [engineering]

- Ethics of Artificial Intelligence (University of Edinburgh), 'This course will cover philosophical issues raised by current and future AI systems. Questions we consider include: […] How do we prevent learning algorithms from acquiring morally objectionable biases?', 'The value alignment problem', 'Racist AI' [philosophy]

- FYS: Ethics and Technology (Swarthmore College), 'Machine Learning and Algorithm Bias' [digital humanities]

- Introduction to the Ethics of Artificial Intelligence (University of Bonn), 'Bias and Nudging' 'What we can do ('design justice')' [philosophy]

- Technology and Design Ethics (LinkedIn Learning), 'Many of the technology courses in our library focus on how to make things. But sometimes, designers and developers need to pause, look around, and evaluate how the things they make affect the people who'll end up using them. [T]his

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

course provides concrete approaches for evaluating and acting on the ethical questions technologists encounter constantly. It provides a framework for ethics in tech, discusses matters related to privacy and security, explains how to discover and define virtues' [designers and developers, professionals in tech]

- The Ethics and Governance of Artificial Intelligence (Harvard in collaboration with MIT, Spring 2018, apparently discontinued), among the topics: 'autonomy, system design, agency, and liability' 'algorithmic bias' 'ownership, control, and access' 'governance, explainability, and accountability' [computer science]

- The Ethics of Artificial Intelligence (Vanderbilt University), 'the course will divide the week between (1) learning the structure and workings of AI through programming workshops; lectures; or research articles in the field of computer science and engineering, and (2) participating in plenary or small group discussions and writing assignments based on literary, critical, philosophical, new media, and/or business case-study readings', 'the instructors will convey important AI methods, and broader concepts, through high-level algorithm descriptions, examples, having students reflect on their own cognitive processes, and rich visualizations of an AI's computational processing' (section 1), 'Section (2) interrogates the meaning, history, and applicability of "intelligence" from various angles, fostering human-centered thinking as the foundation of ethical and civic engagement' [arts, engineering, law, medicine, management, nursing]

Topics in Philosophy of Science (Queen's University) 'Biased algorithms' [philosophy]

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