Ethical Design for Trustworthy Solutions: An Introduction to the Course

Ioannis Patias and Vasil Georgiev

Faculty of Mathematics and Informatics
University of Sofia St.Kliment Ohridski“
5 James Bourchier blvd., 1164, Sofia, Bulgaria

patias@fmi.uni-sofia.bg

Abstract. Autonomous and intelligent systems cannot be trusted if they produce outcomes for which it is unclear or even does not exist assigned responsibility. No system can or should be neither blind trusted nor blind distrusted. The students of the Master's program of Embedded and Autonomous Systems learn how to create, deploy, and operate autonomous and intelligent. Thus, it is essential for our students to also get familiar with methodologies and practice using guidelines related to how to assign systems' outcomes responsibility. This will allow the end users and the other people affected by the systems' operation results to trust such systems, and further use and recommend them. The aim of the paper is to describe the new discipline introduced in our Master's program. In the discipline, the IEEE initiative on ethically aligned design and the guidelines for trustworthy artificial intelligence are presented, discussed and covered in a practical manner. The used project-oriented approach aims to provide the students apart with strong theoretical background, also with real practical instruments and experience on their use. The result is ready to use methodologies and tools applicable in any project or product related to intelligent and autonomous systems, using artificial intelligence.

Keywords: Autonomous and Intelligent Systems, Trustworthy AI, Course Introduction.

1 Introduction

Engineers and people with so-called STEM (science, technology, engineering, and mathematics) background and education usually do not have philosophical or ethics education. They are not familiar with the concepts and terminology used by philosophers and ethicists.

Philosophical and ethics concepts and terminology include many levels of abstraction in their meanings. Starting with basic and common understandings, they continue to generalize at different levels using other also predefined foundational terms. Any exchange of such multi-disciplinary information between peo-

ple with so different backgrounds is ineffective namely because of those layers of abstraction understanding.

However, the lack of theoretical background on philosophy and ethics does not limit people in using some philosophical terms, and acting in ethical manners. Oversimplification of philosophy or underestimation of ethics will not contribute to their wider consumption and application. An introduction to the history of moral philosophy and ethics on the other hand will contribute a lot to a cross-disciplinary information and ideas exchange.

Awareness should be raised on the necessity, education and application of classical philosophy, and ethics to engineers, and people with STEM background. Autonomous and intelligent systems (A/IS) designers and developers can take advantage of the engagement with applied ethics. This will help the multi-disciplinary dialog across specialists with different background from philosophy, trough technology, until reaching applied legal framework (see Fig. 1).

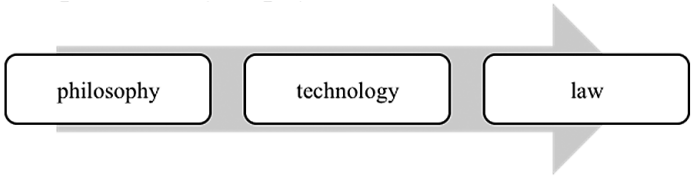

Fig. 1. Multi-disciplinary dialog across specialists with different background.

For the purpose, two successful examples will be covered as theoretical background in the course. One is the IEEE initiative for ethically aligned design and the other is the EU initiative for the provision of the guidelines for trustworthy AI, both described in brief in the following sections.

2 Ethically aligned design

Ethically Aligned Design, First Edition (EAD1e) [1] represents a vision for prioritizing human well being with A/IS. Although it does not necessarily reflect the official policy or position of Institute of Electrical and Electronics Engineers (IEEE), it is published under the IEEE Global Initiative on Ethics of A/IS [2]. EAD1e is a structured set of high-level ethical principles, and includes description of key issues, but also practical recommendations. The main purpose of EAD1e is to inspire its audience to take action. The target audience includes engineers, designers, and manufacturers of A/IS, but also academics, and policy makers.

The Ethically Aligned Design (EAD) Conceptual Framework described in EAD1e is based on the mapping of the three pillars of EAD, to the general principles of EAD. By applying these principles on any specific products, services, systems or combination of systems based on A/IS the engineers, designers, and

manufacturers of A/IS can apply a simple form of impact assessment and due diligence process. This process helps them to match the described principles into their practice.

2.1. The pillars of EAD

The three pillars of the EAD Conceptual Framework cover anthropological, political and technical aspects (see Fig.2):

1. Universal Human Values
   A/IS should be designed to respect human rights, and human values, and increase well-being, safeguard our environment and natural resources.
2. Political Self-Determination and Data Agency
   A/IS should protect political freedom and democracy, improve government effectiveness, accountability, and trust, and protect our privacy.
3. Technical Dependability
   A/IS should deliver services that can be trusted, monitored, validated and verified.

They aim to structure the three aspects and relate them to the respective fundamental values.

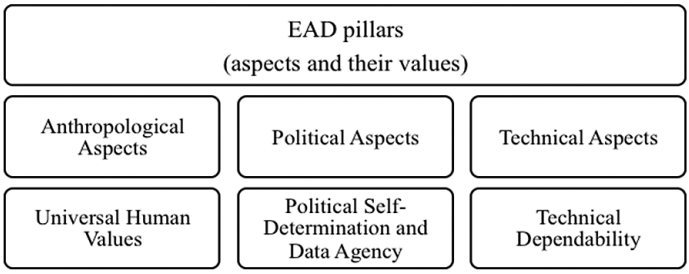

Fig. 2. The pillars of EAD, presented as the covered aspects and the respective fundamental values.

2.2. The general principles of EAD

The general principles of EAD should apply to all types of A/IS, guide their behavior and inform standards and policymaking. They cover all the aspects, by starting with ethical design, going through development, deployment, and adoption, up-to finally even decommissioning of A/IS. They also cover all the roles starting with the designers and manufacturers, going through operators, and other users, up-to any other stakeholders.

The eight general principles of EAD (see Fig. 3) are:

1. Human Rights
2. Well-being
3. Data Agency
4. Effectiveness
5. Transparency
6. Accountability
7. Awareness of Misuse
8. Competence

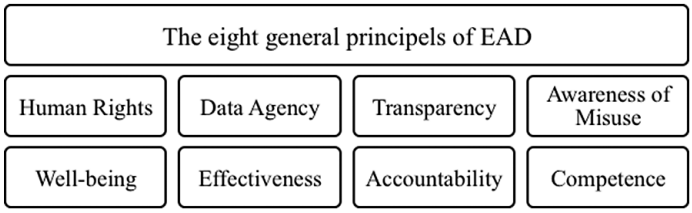

Fig. 3. The eight general principles of EAD.

2.3. Conceptual Framework mapping of pillars into principles, and practice

EAD1e is a product of the IEEE Global Initiative on Ethics of A/IS and aims to service the global policy-making by providing tangible and visible results. Supporting the idea that the society should move "From Principles to Practice" [3] regarding the governance of emerging A/IS developed an action-oriented conceptual framework.

The development of A/IS must be done with respect to our ethical principles. A/IS applications must be validated in practice for honoring our will for political self-determination and data agency. The A/IS we design, develop, and deploy must respect our fundamental human values, and be trustworthy, provable, and accountable.

There are ten Chapters in EAD1e, namely General Principles, Classical Ethics in A/IS, Well-being, Affective Computing, Personal Data and Individual Agency, Methods to Guide Ethical Research and Design, A/IS for Sustainable Development, Embedding Values into Autonomous and Intelligent Systems, Policy, and Law. EAD1e Chapters in order to provide a tangible action-oriented tool are organized by "Issues" and "Recommendations". On the one hand the issues aim to identify A/IS design ethical matters, and on the other the recommendations to provide guidelines on how they should be treated (see Fig. 4).

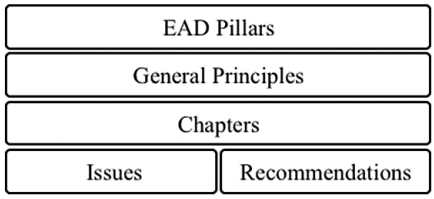

Fig. 4. The action-oriented conceptual framework of EAD.

The overall aim of IEEE efforts is in the direction of introducing the IEEE 7000 Series Standards [4]. Starting from 7000 – Model Process for Addressing Ethical Concerns During System Design, 7001 – Transparency of Autonomous Systems, 7002 – Data Privacy Process, and going to 7013 – Inclusion and Application Standards for Automated Facial Analysis Technology.

3 Trustworthy AI

At the end of 2018, the European Commission set out its vision for AI, which supports "ethical, secure and cutting-edge AI made in Europe" [5, 6]. Three pillars support this vision: (i) investments in AI, (ii) socio-economic changes, and (iii) ethical and legal framework. The Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG) to support its vision implementation. AI HLEG mission was to draft of two deliverables: (1) AI Ethics Guidelines and (2) Policy and Investment Recommendations.

Here we will discuss the structure and applicability of AI Ethics Guidelines [7] as a tool for impact assessment and due diligence process definition. The basic idea is that trustworthiness as prerequisite for people and societies in development, deployment and usage AI systems.

According to AI HLEG, trustworthy AI has three components. Those components are essential for the entire life cycle of a system, namely the system needs to be lawful, ethical, and robust (see Fig. 5).

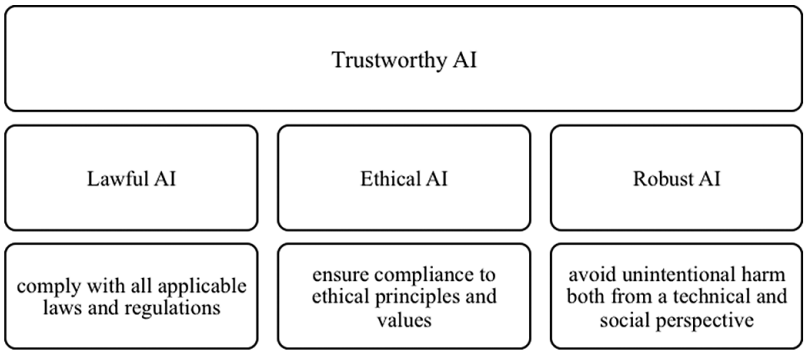

Fig. 5. Trustworthy AI components.

It is important to notice that the Guidelines focus on the second (ethical), and the third (robust) components of trustworthy AI. The reason for not dealing explicitly deal with the first (lawful) component is that what is provided under the existing legislation should be covered, but there are often situations where, their treatment may go beyond existing legal frameworks.

Guidance is provided in three layers of abstraction, from the most abstract to the most concrete. The first, abstract layer is based on an approach founded on fundamental rights. It identifies the ethical principles and their correlated values that must be respected in the development, deployment and use of AI systems.

The second layer sets the frame. It provides guidance on how trustworthy AI can be realized, by listing seven requirements that AI systems should meet (see Fig. 6). Both technical and non-technical methods can be used for their implementation.

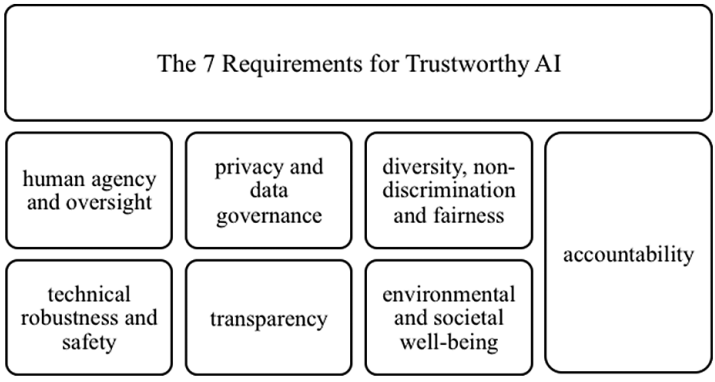

Fig. 6. The seven requirements for trustworthy AI.

The third and most concrete layer defines a list. It is a concrete and non-exhaustive trustworthy AI assessment list aimed at operationalizing the key re-

quirements. This assessment list needs to be tailored to the specific use case of the AI system.

This assessment list is meant to guide AI practitioners to achieve trustworthy AI, when tailored to the specific use case in a proportionate way. The list does not provide concrete answers to address the raised questions; it rather encourages finding ways of how trustworthy AI can be operationalized, and defining the potential steps in this regard.

When using the assessment list in practice, attention should be paid not only to the areas of concern but also to the questions that cannot be (easily) answered. One potential problem might be the lack of diversity of skills and competences in the team developing and testing the AI system, and therefore it might be necessary to involve other stakeholders inside or outside the organization. It is strongly recommended to log all results both in technical terms and in management terms, ensuring that the problem solving can be understood at all levels in the governance structure.

4 ED4TS – the course

The basic course materials are on the one hand EAD1e and on the other the European Commission AI HLEG guidelines. They both focus on the provision of guidelines and policy and investment recommendations, to serve technologists, educators and policymakers.

The aim of this course is to develop the quality of students in applying in real systems, and applications design, the scientific analysis, resources, high-level principles, and actionable recommendations, which will ensure their ethics readiness. The program focuses on pragmatic tools and their application to solve real problems.

Through lectures, case studies, exercises, test examples and tasks students will acquire both basic knowledge and understanding of the key factors for successful applications of guidelines, policies, and recommendations.

Within the course project, students will have to demonstrate practical skills through the realization of a working example of the application of guidelines, policies, and recommendations.

As a result, students will be able to handle cases related to the application of such guidelines, policies, and recommendations in the development of complex embedded and autonomous systems

5 Conclusions

The use and the impact of A/IS increase, and many institutions are trying to establish societal and policy guidelines for their ethical principles, to ensure that they will operate in a beneficial to the people and the environment way. Techno-

scientific communities need to go beyond simple functional and technical solutions, and build trust between people and technology. We need to develop a positive, non-dogmatic way when include human values in AI applications, and solutions. We need to include ethical practices assuring human well-being at individual and collective level in the A/IS design. The proposed course provides a solution to those challenges.

References

1. EAD1e (2018), Ethically Aligned Design, First Edition, "From Principles to Practice" https://standards.ieee.org/industry-connections/ec/ead1e-infographic.html, accessed February 2021
2. IEEE SA (2017), IEEE Standards Association, https://ethicsinaction.ieee.org/, accessed February 2021
3. IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2016), Ethically Aligned Design – Version on a request for input, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf, accessed February 2021
4. IEEE P7000™ Standards Working Group (2017), https://ethicsinaction.ieee.org/p7000/, accessed February 2021
5. COM(2018)237, Communication from The Commission to The European Parliament, The European Council, The Council, The European Economic and Social Committee and The Committee of The Regions, Artificial Intelligence for Europe, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2018:0237:FIN:EN:PDF, accessed February 2021
6. COM(2018)795, Communication from The Commission to The European Parliament, The European Council, The Council, The European Economic and Social Committee and The Committee of The Regions, Coordinated Plan on Artificial Intelligence, https://eur-lex.europa.eu/resource.html?uri=cellar:22ee84bb-fa04-11e8-a96d-01aa75ed71a1.0002.02/DOC_1&format=PDF, accessed February 2021
7. AI HLEG (2019), High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, accessed February 2021