FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Responsible AI development: Challenges and the European approach to AI governance

BACHELOR'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

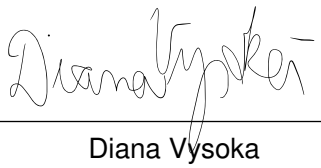
Media Informatics and Visual Computing

by

Diana Vysoka
Registration Number 01633081

to the Faculty of Informatics
at the TU Wien
Advisor: Ao.Univ.Prof. Mag.iur. Dr.iur. Markus Haslinger

Vienna, 4th February, 2021



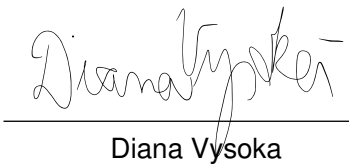
Diana Vysoka | Markus Haslinger

Erklärung zur Verfassung der Arbeit

Diana Vysoka
Herzgasse 13, 1100 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 4. Februar 2021



Diana Vysoka

Abstract

Development of the artificial intelligence has been on the rise for the last decade, despite the two AI Winters in the twentieth century. AI systems accompany humans on every step, from assisting them in daily tasks to just entertaining them. However, AI is also being deployed in more critical fields, such as autonomous driving, predictive policing, well-fare, finance, and many others. Operators of these systems incorporate the decisions of intelligent systems in their own decision making, which means that such system's decision has a direct impact on the society. To maximize the effect and to ensure that the system does not harm the society and does not cause harm, AI governance on a large scale is necessary. The institutions of the European Union face a challenge to create legal frameworks that would directly address these issues and enforce responsible AI development. To create an environment where a user can fully rely on artificial intelligence, designers, developers, and deployers of the AI must follow the principles of transparency, accountability, lawfulness, ethics and trustworthiness, and to make these systems understandable to the users. Such governance does not only depend on the organizational measures, but also the advancements of the technology. As an example, to explain black box systems and to provide insights into their decision making, other algorithms can be applied to accomplish this task. This work analyzes the principles of responsible AI development and its challenges. Furthermore, the initiatives of the EU's institutions are analyzed from the perspective of these principles.

Keywords: Responsible AI, AI governance, AI and society, Explainable AI.

Contents

Acronyms

AGI Artificial General Intelligence. 6, 7, 40, 61

AI Artificial Intelligence. 5, 6

AI HLEG High-Level Expert Group on Artificial Intelligence. 68–70, 73, 78–80, 84

ALTAI Assessment List for Trustworthy AI. 69, 77

ANI Artificial Narrow Intelligence. 6, 11, 40, 61

ASI Artificial Super Intelligence. 6, 61

CEPEJ European Commission for the Efficiency of Justice. 70

CFR Charter of Fundamental Rights of the European Union. 18

CoE Council of Europe. 16, 19

COMPAS Correctional Offender Management Profiling for Alternative Sanctions. 33

ECHR European Convention on Human Rights. 18

IAI Interpretable AI. 45, 46, 85

NeurIPS Conference on Neural Information Processing Systems. 37

RAI Responsible AI. 37, 38, 40, 67

SDGs Sustainable Development Goals. 25, 26

UDHR Universal Declaration of Human Rights. 15, 18

UN United Nations. 14, 15, 26

XAI Explainable AI. 42, 45, 46, 67, 85

CHAPTER 1

Introduction

"Danger is not AI taking over the world, but misuse and failures."

Prof. Dr. Virginia Dignum

Chair of Social and Ethical Artificial Intelligence Department of Computer Science, UMEA University

1.1 Problem

Since the 1930s, when Alan Turing defined an abstract computing machine, also known as a universal Turing machine, the AI development has experienced its ups and downs.[1] The research in the AI field experienced its First Winter in 1973, when many research activities and investments in this field were stopped mostly because the AI did not deliver the promised impact.[2] Some research continued, but the Second AI Winter happened later in 1988, again for the similar reasons of over-promising by the developers and researchers, high expectations from users, and huge media propaganda.[3]

[1] B. Copeland, "Artificial intelligence," *Encyclopedia Britannica,* August 11, 2020, https://www.britannica.com/technology/artificial-intelligence [Accessed on February 3, 2021].

[2] Ben Dickson, "What is the AI winter?" TechTalks, https://bdtechtalks.com/2018/11/12/artificial-intelligence-winter-history/ [Accessed on February 3, 2021].

[3] Ibid.

Nevertheless, the research did not stop completely, and the AI development advanced. Nine years later, after the Second AI Winter, a supercomputer defeated the world champion in chess, Garry Kasparov.[4]

Despite this winding curve of the AI development, the AI has been on the rise for a decade.[5] AI systems accompany the society on every step of the way. As it is applied in different scenarios, the artificial intelligence impacts the daily lives of humans. From helping doctors to diagnose diseases with high accuracy, detect frauds in welfare, to just serve the end-users in the form of voice assistant, AI is widely applied to support human decision making.

In the analogy where a human decision-maker takes into account his or her assistant's suggestion, the decision-maker must find both the suggestion and the person who suggests it trustworthy. First, they must comply with the fundamental principles, be it ethical or legal principles. Second, it is useful if the decision-maker understands the reasoning behind each suggestion. If the decision-maker does not have access to the explanation but only to the given suggestion, such suggestion becomes a worthless piece of information because it does not have any value without the context. Additionally, if a decision proves to be wrong, there must be an entity responsible for the consequences.

Now, if the term *assistant* is substituted by *AI system* in the aforementioned example, it is obvious that there is an emerging need to understand the reasoning behind system's prediction and the factors taken into account. It is necessary that these systems comply with laws and fundamental values and that there is always one entity, be it a natural or juridical person, that is held liable for the potential harm caused by an AI system. Furthermore, all the efforts to develop an AI systems are redundant if the system is not trustworthy and the user cannot rely on it. To overcome these challenges and to eliminate the negative consequences of misuse and failure of AI systems, principles of Responsible AI shall be employed in every step of the product development life-cycle.

The AI-based products shall be developed with the responsibility in mind. However, expecting this responsible approach from every designer, developer, and deployer of the artificial intelligence would not be a sufficient safeguard to ensure that every high-risk AI system employs the Responsible AI principles.

Therefore, a large scale AI governance is necessary. One way to govern AI on a large scale is to put a legal framework in place that would be effective in every Member State of the European Union. A legislation proposal on AI is due in the first quarter of 2021, which is two months after completing this work.[6]

[4] Dustin Waters, "Garry Kasparov vs. Deep Blue: The historic chess match between man and machine," WashingtonPost, https://www.washingtonpost.com/history/2020/12/05/kasparov-deep-blue-queens-gambit/ [Accessed February 3, 2021].

[5] "Are we facing an 'AI Winter' or is our relationship with AI evolving?" OpenAccessGovernment, May 11, 2020, https://www.openaccessgovernment.org/relationship-with-ai/86742/ [Accessed on February 3, 2021].

[6] "Artificial Intelligence," European Commission, last modified January 8, 2021, https://ec.europa.eu/digital-single-market/en/artificial-intelligence [Accessed on February 3, 2021].

1.2 Objective & Motivation

The aim of this bachelor's thesis is to research Responsible AI, its principles, and challenges connected to the design, development, and deployment of the systems powered by AI. One of the crucial principles of Responsible AI is understandability, which heavily depends on the scientific research in the machine learning field. This work aims to look closely at each of the Responsible AI's principles, its challenges and present the latest development in the field. In the second phase, this work reviews the approach of the European Union's institutions towards Responsible AI governance.

Thus, the research question of this thesis is defined as follows:

- What are the challenges of the design, development, and deployment of the Responsible AI?
- How is the European Union adjusting to the advancements of artificial intelligence and to the emerging need to govern this powerful technology on a larger scale?

1.3 Approach

To answer the above-stated research questions, the author analyzes the currently effective legislation acts of the European Union. These acts, to a certain extent, govern the development of Responsible AI, although artificial intelligence was not as widely applied as nowadays, at the time when these acts were formulated.

To objectively assess the impact of artificial intelligence on the society and to showcase the necessity of the Responsible AI, the Chapter 4 showcases both positive and negative instances of the deployment of AI systems.

After the legal frameworks and the urgency to govern AI are presented, the Chapter 5 defines principles of Responsible AI, which shall be an integral part of the emerging technology's product development life-cycle. The chapter summarizes each of the principles; however, it focuses on understandability, accountability, and transparency.

Finally, the Chapter 6 summarizes the milestones of the AI governance in the European Union by reviewing the existing statements and reports of the EU's institutions. The goal of this chapter is to assess the readiness of the EU to engage in public discussion with relevant stakeholders and to implement the feedback from the public discussions. Furthermore, the chapter observes what does it take to assure that the first legislative proposal on AI in the history of the European Union is going to be of high quality and based on decisions informed by opinions, experiences, and suggestions of the leading experts in the field.

CHAPTER 2

Introduction to Artificial Intelligence

With the increasing advancements of Artificial Intelligence (AI) and its contribution to the society and economy, it is clear that research and development of this powerful technology must be fostered and supported by both domestic governments and international bodies, such as the institutions of the European Union. For organizations dealing with vast amounts of data, the artificial intelligence presents a tool of strategic importance. Mainly, it is advantageous when analyzing the past events (in terms of user behavior, medical records, stock market, and other use-cases) to predict the future. Due to the capability of intelligent systems to learn from giga- to terabytes of data, such systems are better at understanding patterns and relations between data points than human minds.

This chapter aims to provide a high-level understanding of artificial intelligence. In the field of AI, different approaches to learning have been developed that are of advantage in different context. In recent years, the development of deep neural networks has been on the rise, additionally to the standard statistical methods used in machine learning.

The objective of this work is to address the *applied* artificial intelligence itself rather than the formal algorithmic research in the AI field. Therefore, this chapter discusses the seven patterns of the AI, to draw an understanding of its application and common use-cases.

2.1 Definition

To provide equal understanding to both non-technical and technical readers, it is necessary first to define the term *Artificial Intelligence*. In this paper, the definition of the AI follows the definition proposed within the European Commission's Communication on AI:

"Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones, or Internet of Things applications)." [7]

From the perspective of the intelligent system's capabilities, AI can be categorized into the following categories:

Artificial Narrow Intelligence (ANI) is a term used to describe intelligent systems designed to handle a specific task.[8] They address a concrete problem, such as language translation, playing computer games or image and pattern recognition. An example of such an application is the self-driving car, voice assistants, or real-time translator.

Artificial General Intelligence (AGI) is a term that describes a system with comprehensive knowledge and cognitive computing capabilities that allow the system to operate beyond the scope of a specific task (in comparison to ANI) while its performance is indistinguishable from that of a human.[9] It is difficult to predict when the break-even between human and artificial intelligence will occur. Still, the experts predict (on average) that by 2040 AGI will be comparable to the human intelligence.[10]

Artificial Super Intelligence (ASI) is more powerful than AGI, as such intelligent system does not only perform as high as human, but even exceeds human capabilities.[11] The point in time when the technological advance is so rapid that to

[7] European Commission, *Communication from the Commission, Artificial Intelligence for Europe (COM/2018/237 final)* (Brussels, Belgium: European Commission, 2018).

[8] "Narrow AI," DeepAI, https://deepai.org/machine-learning-glossary-and-terms/narrow-ai [Accessed on December 7, 2020].

[9] Ben Goertzel, *Scholarpedia* 10(11):31847, s.v."Artificial General Intelligence" (Online: Scholarpedia, 2015), http://www.scholarpedia.org/article/Artificial_General_Intelligence [Accessed on January 11, 2021].

[10] Roberto Saracco, "Computers Keep Getting Better ... Than Us," IEEE Future Directions, https://cmte.ieee.org/futuredirections/2018/01/21/computers-keep-getting-better-than-us/ [Accessed December 7, 2020].

[11] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford, England: Oxford University Press, 2014), p. 63.

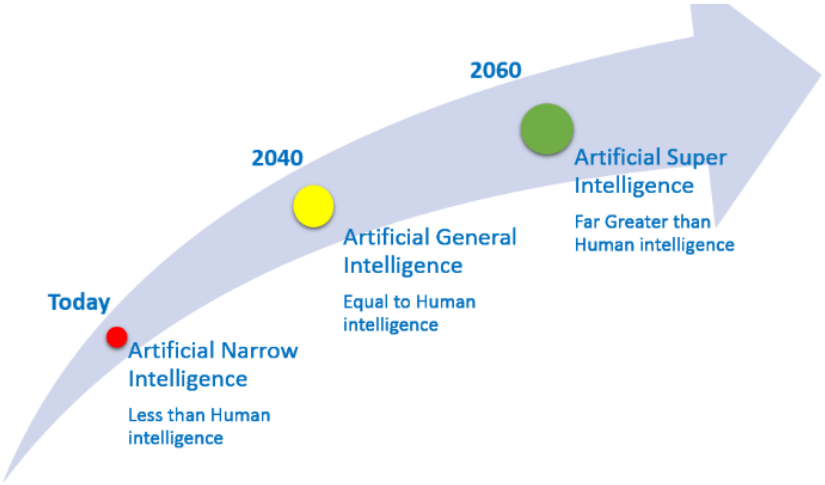


Figure 2.1: Prediction of AI development[15]

human it appears instantaneous is referred to as *singularity*.[12] AI is one of the main drivers of the singularity. Chalmers says that the idea of the singularity is that a machine will become better at designing machines than humans, which will produce a sequence of machines that will be more intelligent than their predecessors, resulting in an uncontrollable development.[13]

The artificial narrow intelligence is nowadays widely used, while the artificial general intelligence is underdeveloped to date. As shown in the Figure 2.1, AGI is expected to be achieved by 2040. Twenty years later, researchers predict to achieve artificial super intelligence.[14]

2.2 Artificial intelligence, machine learning and deep learning

As it often leads to confusion among people with no technical background, the relationship between artificial intelligence, machine learning, and deep learning is discussed in this section. These three terms are often used interchangeably. The following chapters also use artificial intelligence, machine learning models, and algorithms interchangeably, as it is sufficient for the scope of this work, and the distinguishion on that granular level is not necessary. However, there are slight differences between these concepts, as shown in the Figure 2.2.

[12] David J. Chalmers, "The Singularity: A philosophical Analysis" *Journal of Consciousness Studies* 17, (2010):1, https://www.researchgate.net/publication/233701623_The_Singularity_A_Philosophical_Analysis [Accessed on February 3, 2021].
[13] Ibid.
[14] Pratul Kumar Singh in Saracco, "Computers Keep Getting Better ... Than Us."
[15] Ibid.

Artificial intelligence is a study of intelligent agents that are any devices that perceive their environment and take actions that maximize their chance of successfully achieving their goals.[16] The AI scientists study ways to build intelligent systems that can creatively solve problems that normally require human cognition. Artificial intelligence is an interdisciplinary study that often draws upon computer science, mathematics, psychology, neuroscience, linguistics, and many others.[17] Some of the typical AI problems are planning, decision making/reasoning, natural language processing, movement, and perception.

Machine learning is a subfield of artificial intelligence, and machine learning methods are ways to create artificially intelligent systems. Machine learning is based on mathematical and statistical models capable of extracting patterns from data(finite set of *features* of the data), learning from them, and then providing a solution to the aforementioned problems the model was trained to solve.[18] A basic example of a machine learning task is to predict the person's origin based on her or his name, skin color, or mother tongue. All of these attributes are called features in machine learning. Some machine learning algorithms are Linear Regression, Logistic Regression, Decision Trees, Naive Bayes, k-Nearest Neighbors, or Random Forest. A single mathematical function is also called a neuron.[19]

Note: *This work often uses the term machine learning model or simply a model. These terms refer to concrete instances of machine learning algorithms that are already trained to carry out a task and solve a machine learning problem.*

Deep learning is a field of machine learning where algorithms' goal is to solve the aforementioned problems; however, in more complex scenarios, where a finite set of features cannot be found, such as in computer vision.[20] In a regular computer vision task, a user would be interested in classifying an object in the picture. He could describe to object, its colors, its shapes, basically every attribute of the object. However, these attributes cannot be easily interpreted by machines. Furthermore, the object's position could be different in every picture, which would be very difficult to describe to a single machine learning model. Therefore, the deep learning introduces an approach where neurons are interconnected to form so-called *neural networks*. This name comes from the analogy with the human brain, as these algorithms imitate the human neural networks' function. Examples of deep learning methods are Convolutional Neural Networks, Recurrent Neural Networks, Generative Adversarial Networks, or Deep Reinforcement Learning. The shallower layer of the network, the more basic features such as contours, corners, and colors

[16] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach,* 3rd ed. (Harlow, Essex, England: Pearson Education Limited, 2016), p. 34.
[17] Ibid. p. 5-14.
[18] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning* (Cambridge, Massachusetts, USA: MIT Press, 2016), p. 2-3.
[19] Russell and Norvig, *Artificial Intelligence: A Modern Approach,* p. 728.
[20] Goodfellow, Bengio and Courville, *Deep Learning,* p. 3.

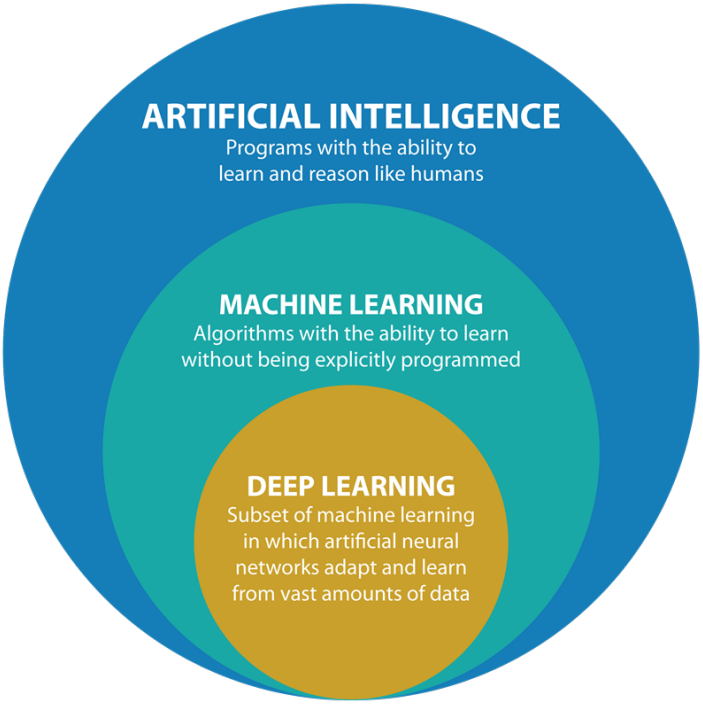


Figure 2.2: Relation between artificial intelligence, machine learning and deep learning[23]

are recognized.[21] The deeper layer of the network, more complex features of objects are recognized, such as faces or fingerprints.[22]

2.3 Seven patterns of Artificial Intelligence

Humans use AI-powered devices or software daily, often without even realizing it. From personalized news feed on social media, product suggestions in online shops, health tracking apps, to voice assistants, AI became an inseparable part of our lives. To provide an understanding of all possible use-cases of AI, seven patterns of AI can be defined, as shown in Figure 2.3.

2.3.1 Hyper-personalization

Many software systems use machine learning to develop a personalized user profile. This profile is a virtual model of each user, created based on their preferences, virtual behavior

[21] Goodfellow, Bengio and Courville, *Deep Learning,* p. 6.
[22] Ibid.
[23] "What Is Artificial Intelligence, Machine Learning And Deep Learning?" Argility, https://www.argility.com/data-analytics-ai-ml/ [Accessed on January 11, 2021].
[24] "The Seven Patterns Of AI," Cognilytica, April 4, 2019. https://www.cognilytica.com/2019/04/04/the-seven-patterns-of-ai [Accessed on December 7, 2020].

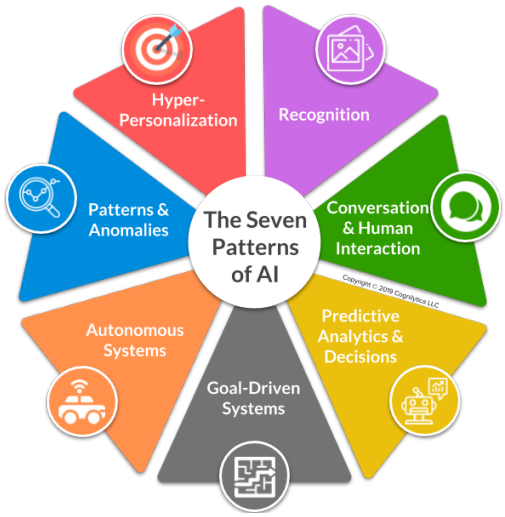


Figure 2.3: Patterns of AI[24]

patterns, and other parameters that are use-case dependent.[25] As the system learns and acquires more information about the user, the model adapts to provide better results. The profile is then used for various purposes, such as personalized recommendations, displaying relevant content, etc. Hyper-personalization is widely used in social media platforms such as Facebook or Instagram, streaming services such as Netflix, video sharing platforms such as Youtube, or online travel agencies as Booking. All of these platforms use machine learning to adapt their products and content to every individual user.

2.3.2 Patterns and anomalies

Pattern recognition is an ability of a computer system to recognize patterns in provided data. Machine learning achieves good results in recognizing regularities and identifying outliers. Applications of this pattern achieve excellent results in risk and fraud detection,[26] as machine learning is based on statistical computations. The goal of such a system is to learn patterns in a specific dataset, understand the dependencies between the data points and then decide whether a new data point fits the known pattern. Such systems are useful in the social insurance field to detect whether there are any patterns in the history of sick leaves of patients; or in the banking field, where unauthorized use of a credit or debit card can be detected.

[25] Gilad Maayan, "Hyper Personalization: Customizing Service With AI," Computer, https://www.computer.org/publications/tech-news/trends/hyper-personalization-customizing-service-with-ai [Accessed on February 3, 2021].

[26] Pradheepan Raghavan and Neamat El Gayar, "Fraud Detection using Machine Learning and Deep Learning," *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (Dubai, UAE: IEE, 2019), p. 1, https://www.doi.org/10.1109/ICCIKE47802.2019.9004231.

2.3.3 Autonomous systems

Cambridge dictionary defines an autonomous system as "a system that is able to operate without being controlled directly by humans." [27] The goal of the autonomous system developers is to create a machine capable of solving a complex task with little to no human interaction. A precondition of this behavior is a capability to perceive the environment, predict its actions, and act in accordance with them to maximize the outcome. Many autonomous systems are already being used on a daily basis, such as smart dust cleaners, lawn trimmers, medical robots that conduct surgeries, or self-driving cars. All of these smart systems could be categorized as ANI. More advanced ones are still in development, such as the humanoid robot Atlas designed by Boston Dynamics.[28]

2.3.4 Goal-driven systems

Building upon the capability to learn patterns from seen data, machine learning algorithms enable machines to learn rules and strategies. These strategies are then used to solve puzzles, win games or simply optimize problems such as finding the shortest path, optimizing resources. The most famous example of a goal-driven system is Deep Blue, developed by IBM. In 1997, Deep Blue defeated the former world champion in chess, Gary Kasparov.[29]

2.3.5 Predictive analytics

When a smart system learns a pattern from training data and understands the past, it gains the ability to predict the future output of new datapoints it has never seen before. This AI pattern is helpful for humans to make decisions based on a long history that only a computer can evaluate. These systems are often used for predicting customer behavior or stock market forecast.

2.3.6 Conversation and human interaction

Conversational AI enables humans to interact with computers naturally – with voice and in written language. Systems powered by this type of AI are also usually capable of understanding context and leading a dialog, rather than pure question-answer communication. As the user expresses their intent, the machine first translates it into a formal language, then determinates the intent, finds the proper answer, and translates it back to the natural language. Examples of such intelligent apps are voice assistants such as Amazon Alexa, Google Duplex, or various chatbots.

[27] *Cambridge Dictionary,* s.v. "autonomous," https://dictionary.cambridge.org/dictionary/english/autonomous [Accessed on December 7, 2020].

[28] https://www.bostondynamics.com/atlas [Accessed on December 7, 2020].

[29] "Deep Blue," IBM, https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/ [Accessed on December 7, 2020].

2.3.7 Recognition

Mostly realized by deep learning, its goal is to detect, recognize, classify or identify objects. The input data for deep learning algorithms are complex patterns such as image, video, text, or other unstructured data. Nowadays, almost every smartphone uses deep learning algorithms for biometry such as face, voice or fingerprint recognition. Recognition is also used to automate invoice processing, where neural networks extract information from unstructured business documents. An example of such software is Rossum.

CHAPTER 3

Overview of the European legislation acts

As discussed in the previous chapter, for an AI model to make intelligent decisions, it first has to be trained on big training data sets, which, based on the use-case and application, often include personal information of the users. This means that the output of such a system solely depends on the quality of the training data used for machine learning. As intelligent systems learn from the collected data that mostly represent human behavior, such systems learn to mimic the human decision making.

As humans make mistakes and are often biased, a smart system learns to reproduce the bias, potentially introduces new ones, and therefore makes biased decisions after its deployment.[30] The data itself can also be biased - Google's algorithm observed that men are more likely to interact with job ads for high-paid jobs, and therefore, women were less likely to be shown such job ad.[31] If no quality measures are in place, such biased system could violate the fundamental and human rights of citizens, if applied in a domain with substantial impact on citizens such as predictive policing. Predictive policing goes beyond the ability of the human mind to analyze past offenses and predict possible future patterns of crime, such as which individuals are likely to become involved in a crime or sentence severity based on profiling and probability of becoming a repeat offender.[32] If

[30] Giovanni Sartor and Francesca Lagioia, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence,* (Brussels: European Parliament, 2020), p. 1, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf [Accessed on December 8, 2020].

[31] Amit Datta, Michael Carl Tschantz and Anupam Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," *Proceedings on Privacy Enhancing Technologies* 2015, no. 1 (2015):1, https://doi.org/10.1515/popets-2015-0007.

[32] Committee of experts on internet intermediaries (MSI-NET), *Algorithms And Human Rights - Study On The Human Rights Dimensions Of Automated Data Processing Techniques And Possible Regulatory Implications (DGI (2017)12),* (Strasbourg, France: Council of Europe, 2018), p. 10-11.

such model learns from biased data, it also predicts biased decisions. Such approaches require extensive oversight and appropriate safeguards, as they may be highly prejudicial in terms of ethnic and racial backgrounds.[33]

To avoid such pitfalls in general, regardless if the traitor is a human person or a technology, the EU has a strict legal framework when it comes to human rights, data privacy, and liability. This chapter summarizes the European legislation that is relevant in the context of AI but does not explicitly address it. Firstly, human rights and corresponding legislative acts are described, including relevant rights in the context of AI (human rights that could be violated by AI). Secondly, the data and privacy protection legislative acts are discussed, as the right to privacy is a substantial part of human rights, and AI heavily depends on the data. Finally, as the AI-based systems are eventually products and someone has to be held liable for their impact, this chapter briefly discusses the product liability.

3.1 Human & fundamental rights

3.1.1 United Nations - Universal Declaration of Human Rights (1948)

After the World War I., the League of Nations was established, as the first intergovernmental organisation, to prevent future wars and conflicts between countries.[34] With the World War II, the League of Nations showed to be ineffective by failing its primary purpose - maintaining the world peace. Vast majority of the countries world-wide were directly or indirectly affected by the World War II, that, in the European space, ended on May 8, 1945.[35] After the war, the Member States of the League of Nations rejected the idea of restoring the League, establishing the United Nations (UN) instead. Fifty governments and hundreds of nongovernmental organizations met in San Francisco on June 26, 1945, where they signed the Charter of the United Nations, a new constitutional framework of the UN that came into force on October 24, 1945.[36] The Statute of the International Court of Justice is integrated in the Charter.

The purpose of the United Nations is defined in Article 1 of the Charter of the United Nations as:

"To maintain international peace and security, and to that end: to take effective collective measures for the prevention and removal of threats to the peace, and for the suppression

[33] Committee of experts on internet intermediaries, *Algorithms And Human Rights - Study On The Human Rights Dimensions Of Automated Data Processing Techniques And Possible Regulatory Implications.*

[34] Charles Townshend, "History - World Wars: The League Of Nations And The United Nations," BBC UK, last modified February 17, 2011, http://www.bbc.co.uk/history/worldwars/wwone/league_nations_01.shtml [Accessed on December 13, 2020].

[35] "75Th Anniversary Of The End Of World War II," The National WWII Museum New Orleans, https://www.nationalww2museum.org/war/topics/75th-anniversary-end-world-war-ii [Accessed on December 13, 2020].

[36] Claude Welch, "Universal Declaration Of Human Rights: Why does it matter?" UBNow, December 17, 2015, http://www.buffalo.edu/ubnow/stories/2015/12/qa_welch_udhr.html [Accessed on December 13, 2020].

of acts of aggression or other breaches of the peace, and to bring about by peaceful means, and in conformity with the principles of justice and international law, adjustment or settlement of international disputes or situations which might lead to a breach of the peace;"[37]

In other words, the objectives are maintaining international peace and security, protecting human rights, delivering humanitarian aid, promoting sustainable development, and upholding international law.[38] These objectives later motivated the formation of the Sustainable Development Goals, discussed in the section 4.1.

At its founding, the UN had 51 member states.[39] To date, the UN recognises 195 sovereign states in the world, of which 193 (including all the members of the European Union) are the members of the UN. The remaining 2 sovereign states that are not members of the UN are Palestine and Vatican City.[40]

Three years later after the founding of the UN, on December 10, 1948 in Paris, France, the Universal Declaration of Human Rights (UDHR) was proclaimed by the United Nations General Assembly.[41] The Declaration consists of 30 rights and freedoms, that can be split into two groups: civil and political rights, such as the right to life, the right for freedom, the right for fair trial or the right to privacy; and economic, social and cultural rights, such as the right to social security, health and education.

After the declaration of UDHR, the General Assembly requested the Commission on Human rights to draft two covenants: the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR).[42] Together with UDHR, these three documents form the International Bill of Rights. The covenants are legally binding on States which have signed and ratified them, in contrary to UDHR, which is is not legally binding and enforceable in a court. The implementation of the ICESCR by the member states is monitored by the Committee on Economic, Social and Cultural Rights.[43] The implementation of ICCPR by the member parties is monitored by the Human Rights Committee.[44]

[37] United Nations, *Charter of the United Nations,* October 24, 1945, 1 UNTS XVI, Available at: https://www.refworld.org/docid/3ae6b3930.html [Accessed on December 13, 2020].

[38] "What We Do," United Nations, https://www.un.org/en/sections/what-we-do/index.html [Accessed on December 13, 2020].

[39] "UN Membership: Founding Members," Dag Hammarskjöld, https://research.un.org/en/unmembers/founders [Accessed on December 13, 2020].

[40] "Member States," United Nations, https://www.un.org/en/member-states/index.html [Accessed on December 13, 2020].

[41] "Universal Declaration Of Human Rights," United Nations, https://www.un.org/en/universal-declaration-human-rights/index.html [Accessed on December 13, 2020].

[42] "Human Rights Explained: Fact Sheet 5:The International Bill Of Rights," Australian Human Rights Commission, https://humanrights.gov.au/our-work/education/human-rights-explained-fact-sheet-5the-international-bill-rights [Accessed on December 13, 2020].

[43] "Committee On Economic, Social And Cultural Rights," United Nations Human Rights Office of the High Commissioner, https://www.ohchr.org/en/hrbodies/cescr/pages/cescrindex.aspx [Accessed on December 13, 2020].

[44] "Human Rights Committee," United Nations Human Rights Office of the High Commissioner,

3.1.2 Council of Europe - European Convention on Human Rights (1950)

Founded after the World War II, the Council of Europe is one of the oldest and the biggest European organisation, which unifies 47 member states (including all 27 members of EU) and promotes the main principles of the Human Rights.[45] The Council was founded by ten member states (Belgium, Denmark, France, Ireland, Italy, Luxembourg, Netherlands, Norway, Sweden, and United Kingdom) on May 5, 1949.[46]

After the end of the World War II, the governments of the European countries were determined to ensure that tragedy of this kind would not repeat in the future. Winston Churchill, who was a prime minister of the United Kingdom during the war, delivered a speech at the University of Zurich on September 19, 1946, where he pointed out there was a need for

"...remedy which, if it were generally and spontaneously adopted by the great majority of people in many lands, would as by a miracle transform the whole scene and would in a few years make all Europe, or the greater part of it, as free and happy as Switzerland is today. What is this sovereign remedy? It is to recreate the European fabric, [...], and to provide it with a structure under which it can dwell in peace, safety and freedom. We must build a kind of United States of Europe."[47]

Taking in count that five years prior to Churchill's speech at the University of Zurich the United Nations was established, there was a discussion how would it conflict with the suggested "United States of Europe", which later became the Council of Europe (CoE). According to Churchill, there was "no reason why a regional organisation of Europe should in any way conflict with the world organisation of the United Nations". On the contrary, he believed that the larger synthesis could only survive if it is founded upon broad natural groupings.[48]

To take the first steps for the collective enforcement of the rights stated in the Universal Declaration in November 1950, the governments of European countries have agreed to form the European Convention on Human Rights.[49] The Convention is legally binding to the member states and interpreted by the European Court of Human Rights, with the seat in Strasbourg, France.[50]

---

https://www.ohchr.org/en/hrbodies/ccpr/pages/ccprindex.aspx [Accessed on December 13, 2020].

[45] "About The Council Of Europe - Overview," Council of Europe Office in Yerevan, https://www.coe.int/en/web/yerevan/the-coe/about-coe/overview [Accessed on December 14, 2020].

[46] Ibid.

[47] "Winston Churchill, speech delivered at the University of Zurich, 19 September 1946," Council of Europe, https://rm.coe.int/16806981f3 [Accessed on December 14, 2020].

[48] Ibid.

[49] Council of Europe, "European Convention for the Protection of Human Rights and Fundamental Freedoms," (1950), *Council of Europe Treaty Series 005* p. 5. https://www.echr.coe.int/documents/convention_eng.pdf [accessed on December 13, 2020].

[50] Ibid. Art. 19.

3.1.3 The Charter of Fundamental Rights of the European Union (2000)

European Union was officially founded in 1993, with the Maastricht Treaty coming into force.[51] Similarly to the previously described organisations, the first idea of creating the European Union was created after the World War II to ensure economic and social prosperity of the member countries and was preceded by the European Coal and Steel Community and the European Economic Community.[52] The European Union refers to the values of the Council of Europe as a core of its social and economic politics.

Over the time, with the changes in society in the social, technological and scientific development and the expansion of EU policies which directly affect fundamental rights, there was a need to formulate legislation act effective and legally binding in every member state of the European Union.[53] This legislation act is called The Charter of Fundamental Rights of the European Union and came into effect with the Lisbon Treaty on December 1, 2009, nine years after its declaration.[54]

The Charter of Fundamental Rights of the European Union consists of 54 articles that build upon human rights and interpret these rights in a way that addresses the challenges of the modern world. As an example, human right defined in the Art. 3[55] of UDHR addresses the security of a person, whereas the Art. 34 of the Charter[56] is interpreted as social security and social assistance, and seems like it extends its Art. 6,[57] "Right to liberty and security." Similarly, both documents address the respect for privacy and family, however, the Charter adds protection of personal data in its Art. 8, as it is a problem of the twenty-first century.[58]

The Charter consists of 6 chapters: Dignity, Freedoms, Equality, Solidarity, Citizen's rights, Justice. For the relevant Articles of this Charter, refer to the Subsection 3.1.4.

[51] "The History Of The European Union," European Union, https://europa.eu/european-union/about-eu/history_en [Accessed on December 14, 2020].

[52] Ibid.

[53] "European Charter Of Fundamental Rights: Five Things You Need To Know," European Parliament, December 1, 2019, https://www.europarl.europa.eu/news/en/headlines/society/20191115STO66607/european-charter-of-fundamental-rights-five-things-you-need-to-know [Accessed on December 14, 2020].

[54] Ibid.

[55] UN General Assembly, *Universal Declaration of Human Rights*, December 10, 1948, 217 A (III), Art. 3, https://www.refworld.org/docid/3ae6b3712c.html [Accessed on December 14, 2020].

[56] European Union, *Charter of Fundamental Rights of the European Union,* October 26, 2012, 2012/C 326/02, Art. 3, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012P%2FTXT [Accessed on December 14, 2020].

[57] Ibid. Art. 6.

[58] Ibid. Art. 8.

3.1.4 Selected rights

All of the previously described documents are centered around human rights and therefore have many articles in common. For the brevity, this subsection selects the articles of the CFR that are relevant in the context of AI. These articles are also either directly mentioned in ECHR and UDHR or implied. Therefore, this chapter does not address articles of each act separately.

Title	Article	Name
Dignity	Art. 1	Human dignity
	Art. 3	Right to the integrity of the person
Freedoms	Art. 6	Right to liberty and security
	Art. 7	Respect for private and family life
	Art. 8	Protection of personal data
	Art. 10	Freedom of thought, conscience and religion
	Art. 11	Freedom of expression and information
	Art. 14	Right to education
	Art. 15	Freedom to choose an occupation and right to engage in work
Equality	Art. 20	Equality before the law
	Art. 21	Non-discrimination
	Art. 23	Equality between women and men
	Art. 26	Integration of persons with disabilities
Justice	Art. 47	Right to an effective remedy and to a fair trial
	Art. 48	Presumption of innocence and right of defence

Table 3.1: Relevant rights in the context of AI

The AI systems are discriminatory, as described later in the case of COMPAS, the American tool used to predict the likelihood that a criminal defendant would reoffend (Subsection 4.2.3). Political participation of a citizen can be manipulated by intelligent bots that spread disinformation on social media and therefore manipulate political processes.[59] An example of the right to privacy is a prediction of sexual orientation based on data from online dating websites.[60] The right to freedom of expression can be violated by sentiment analysis of social media posts, and following removal of the post to make an impression on the user that there is only positive content on the platform and therefore keep the user on the platform for longer time.[61] The streaming company Netflix was sued to include the subtitles in the movies that the company streams, to

[59] Mark Matonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, (Online: Data Society, 2018), p. 12, https://datasociety.net/library/governing-artificial-intelligence/ [Accessed on February 3, 2021].

[60] Matonero, *Governing Artificial Intelligence*, p. 13.

[61] Ibid. p. 14.

treat people with disabilities with human dignity and to integrate them.[62] Later in this work, in the Subsection 4.2.2, discrimination is addressed in the hiring processes, which indirectly impacts the right to engage in work. To make the complaint procedures more efficient, companies use automated data processing to handle these complaints and to remedy the customers if they are not satisfied with the service, which could potentially violate the right to effective remedy, as an AI system cannot carefully analyze the case and take all the relevant factors into consideration.[63]

3.2 Data & privacy protection

As the protection of privacy and personal data is an integral part of the previously discussed documents, this section further describes the concrete conventions agreed upon by the CoE to guarantee the right for data and privacy protection. Historically most significant legislation act is the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, also known as Convention 108. The development of the telephony systems later instigated the creation of the Directive 97/66/EC. The topic got even more relevant in the recent years, when the automated data processing became an essential part of the information systems and big data processing accompanied us in every step. The aim of this chapter is to present the history of the data & privacy protection safeguards in the European countries.

3.2.1 Convention 108 (1985)

With the development of the information technology and internet, data became an important component of the IT systems. Computers made it even easier to collect, process and store data. In response to the recent development back then in 1981, the member states of the Council of Europe agreed on a convention that would bind them to incorporate the data protection agenda into their domestic laws. This document was ratified four years later, in 1985, under the title Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.[64] This Convention builds upon the Human right defined in the Article 12 of the UDHR and Article 8 of the Charter of Fundamental Rights of the EU.

The Convention 108 is the first international document that enforces the protection of the individual against abuses of the personal data and unlawful processing of such data. In addition to providing standards for the collection and processing of personal data, it outlaws the processing of personal data such as "racial origin, political opinions or

[62] Jonathan Hassell, "Netflix captions lawsuit settlement – how the perception of why you've improved your accessibility is vital for ROI," Hassell Inclusion, October 24, 2020, https://www.hassellinclusion.com/blog/netflix-captioning-settlement/ [Accessed on January 31, 2021].

[63] Committee of experts on internet intermediaries (MSI-NET), *Algorithms And Human Rights,* p. 25.

[64] Council of Europe, *Convention for the Protection of Individuals with Regard to the Automatic Processing of Individual Data,* January 28, 1981, ETS 108, https://rm.coe.int/1680078b37 [Accessed on December 15, 2020].

religious or other beliefs, as well as personal data concerning health or sexual life" in the absence of proper legal safeguards in the domestic law.[65] By defining the fundamental terms in the data processing, such as personal data, automatic processing or controller, it became a basis for the upcoming regulations discussed later in this chapter.

3.2.2 Directive 97/66/EC (1997)

With the development of the telecommunications sector, equivalent level of protection of the right to privacy, with respect to the processing of personal data and the free movement of such data, the European Parliament and the Council of the European Union drafted the Directive 97/66/EC that came into force in 1997.[66]

In this directive, the aim is to enforce security, confidentiality of the communications, as well as define safeguards for traffic and billing data for the subscribers. Subscribers are defined as natural or legal persons who are party to a contract with the provider of publicly available telecommunication services.

3.2.3 Convention on cybercrime (2004)

Conscious of the changes brought about by the digitalisation and continuous globalisation of computer networks, and the potential risk that the computer networks could be used for committing crimes, the member states of the Council of Europe have agreed to sign the Convention on cybercrime, also known as the Budapest Convention. The Convention is the first international treaty on crimes committed via the Internet and other computer networks and came into force in 2004.[67]

Its main objective is to pursue a common criminal policy among the Member States aimed at the protection of society against cybercrime, such as illegal access, interception, data interference, misuse of devices or other computer-related forgery, as described in the Chapter 2 of the Convention.[68]

3.2.4 General Data Protection Regulation (EU) 2016/679 (2016)

The GDPR succeeded the Data Protection Directive 95/46/EC from 1995. There are a couple of differences between the two legislative acts. First, a directive sets out goals that each EU Member State must achieve, but gives the Member States the freedom

[65] Council of Europe, *Convention for the Protection of Individuals with Regard to the Automatic Processing of Individual Data,* Art. 6.

[66] European Union, *Directive 97/66/EC of 15 December 1997 of the European Parliament and of the Council Concerning the Processing of Personal Data and the Protection of Privacy in the Telecommunications Sector,* December 15, 1997, FXAL98024ENC/0001/01/00, https://www.refworld.org/docid/3ddcc6364.html [Accessed on December 15, 2020].

[67] Council of Europe, *Convention on Cybercrime,* November 23, 2001, https://www.refworld.org/docid/47fdfb202.html [Accessed on December 15, 2020].

[68] Ibid. Chapter II, Section 1.

to formulate their own laws on how to reach these goals.[69] A regulation, however, is a legislative act that is automatically effective in every Member State.[70] Therefore, the GDPR has ensured that in every country, the same rules on the data privacy and protection apply, with no difference. This is a very useful step, as it standardizes the rules of the data acquisition, processing and storage across the European Union. The GDPR was adopted in 2016, but became enforceable two years later, on May 25, 2018.[71]

The main idea of the Regulation is to prohibit any kind of personal data processing.[72] The GDPR, however, sets out ten exceptions, such as explicit consent from the data subject or when the personal data are manifestly made public by the data subject.[73] If one of the ten exceptions applies, the data processing must comply with requirements set out in the Articles of the GDPR.

The Regulation defines following terms:

Data subject as "a natural person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name" [74]

Personal data as “any information relating to an identified or identifiable natural person (‘data subject’); ... such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;"[75]

Consent as "any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;"[76]

Pseudonymisation & Anonymisation "processing of personal data in sucha manner that the personal data can no longer be attributed to a specific data subjectwithout the use of additional information, provided that such additional information iskept separately and is subject to technical and organisational measures to ensure thatthe personal data are not attributed to an identified or identifiable natural person"[77] Pseudonymized data therefore fall within the scope of GDPR, as they are reversible and the data subject is re-identifiable, as declared in GDPR Recital 26.[78] The same

[69] European Union, "Regulations, Directives and other acts," European Union, https://europa.eu/european-union/law/legal-acts_en [Accessed on January 30, 2021].

[70] European Union, "Regulations, Directives and other acts."

[71] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data,and repealing Directive 95/46/EC (General Data Protection Regulation)" (2016) *Official Journal of theEuropean Union* L 119, Art. 84 (2)

[72] Ibid. Art. 9 (1).

[73] Ibid. Art. 9 (2)

[74] Ibid. Art. 4(1)

[75] Ibid.

[76] Ibid. Art. 4 (11).

[77] Ibid. Art. 4 (5).

[78] Ibid. Recital 26.

Recital specifies that personal data, that underwent anonymization, and therefore the data subject can neither directly nor indirectly be identified with such data, do not have to comply with the Regulation.

Controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;"[79]

Processor processes the data on behalf of the data controller.[80]

Processing as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;"[81]

Profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;"[82]

When it comes to personal data processing, the GDPR Art. 5 requests that the data must be processed lawfully, transparently and the data must be collected for specific purpose. The volumes of data must be kept minimal, only to store data that is necessary for the processing.[83] The controller must keep the data the data accurate, and delete or rectify the data on the request of the data subject.[84] The controller must put organisational and technical safeguards in place to protect the data.[85]

The GDPR requires that the data controller takes appropriate actions to inform the data subject about his rights, such as right of access to the data, right to rectification of the data, right to erasure, right to restriction of processing, right to data portability, right to object to automated decision making including profiling and it is the controller's obligation to inform the data subject of a personal data breach.

The above described transparency and the automated decision-making including profiling often collide with the main principles of the AI. Transparency cannot always be guaranteed when it comes to the applications based on artificial intelligence, especially if "explanations" of automated decision-making are requested, as later described in the Chapter 5. This

[79] GDPR Art. 4 (7).
[80] Ibid. Art. 4 (8).
[81] Ibid. Art. 4 (2).
[82] Ibid. Art. 4(4).
[83] Ibid. Art. 5.
[84] Ibid.
[85] Ibid.

is due to the "black box" nature of AI. Due to its discriminative nature, as a result of learning from existing (biased) data, profiling can have negative impact on rights of the data subjects, as described in the Chapter 4.

3.3 Product liability

To create a safe environment for the consumers across the Europe, the Council of the European Communities has adopted the Product Liability Directive (85/374/EEC) in 1985.[86] The main statement of the Directive is that the producer is fully liable for damage caused by a defect in his product, although the producer did not intend to harm.[87] However, the injured person must prove the damage, the defect and the causality between the two.[88] In regard to the artificial intelligence, the definition of product is not very sufficient, as product is defined as "all movables." [89]

Apart from the product liability, there are two more widely used terms - strict and vicarious liability.

Strict liability is given if a manufacturer produces products that are inherently dangerous, such as chemicals or explosives.[90] Strict liability also applies if a person owns wild animals.[91] In such cases, there is not need to prove negligence if a consumer (or other person) is harmed.[92]

If an entity acts on behalf of another and causes harm in course of this acion, the other entity is held vicariously liable.[93] The best example for this is an employee, that acts within the scope of his employment and unintentionally causes harm. The employer is held vicariously liable for employee's tort.

[86] Council of the European Communities, "Council Directive 85/374/EEC on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products" (1985) *Official Journal* L 210.

[87] Ibid. Art. 1.

[88] Ibid. Art. 4.

[89] Ibid. Art. 2.

[90] Mark Weycer, "Strict Liability vs Product Liability," Weycer Law Firm, August 10, 2019, https://weycerlawfirm.com/blog/product-liability-vs-strict-liability/ [Accessed on January 30, 2021].

[91] Ibid.

[92] Ibid.

[93] Gerald N. Hill and Kathleen Thompson Hill, s.v. "vicarious liability," *The People's Law Dictionary,* (New York, NY, USA: MJF Books, 2002), https://archive.org/details/B-001-001-744/page/n427/mode/2up [Accessed on January 30, 2021].

CHAPTER 4

Impact analysis of the AI development on the society and the human rights

This chapter looks at the impact of artificial intelligence on the society. As everything has its pros and cons, artificial intelligence is no exception. Firstly, the chapter analyzes how AI can be an efficient tool to solve the world's problems and reduce the gap between developing and developed countries. This positive impact is considered in the context of the Sustainable Development Goals (SDGs). Secondly, to provide understanding and to express the urgency for Responsible AI development, this chapter describes a couple of cases where algorithms and emerging technologies harmed the society.

4.1 AI as a tool to achieve Sustainable Development Goals

SDGs were conceived at the United Nations Conference on Sustainable Development in Rio de Janeiro in 2012.[94] the conference aimed to produce a set of universal goals that would be adopted by all members of the UN to meet the urgent environmental, political and economic challenges we face.[95] Three years later, the Sustainable Development Goals were adopted by all United Nations member states as a call to action to fight poverty,

[94] United Nations Development Programme, "Background on the goals," UNDP, https://www.undp.org/content/undp/en/home/sustainable-development-goals/background.html [Accessed on November 17, 2020].

[95] Ibid.

Figure 4.1: List of Sustainable Development Goals [97]

protect the planet and ensure that all people enjoy peace and prosperity by 2030.[96] The 17 goals are displayed in the Figure 4.1.

One of the ways to achieve these goals and solve the current global issues is to employ the emerging technologies and the latest research in the field of AI. To date, a plethora of social impact and high-tech projects and start-ups have been founded that apply machine learning models, that either directly or indirectly solve the social, political, and economic challenges. As described later in this section, it can be concluded that the role of artificial intelligence in achieving the SDGs is significant, especially if designed responsibly. Out of 169 targets across all goals, AI could help to achieve 134 targets but possibly inhibit 59 targets.[98]

This section aims to highlight the positive impact of AI on SDGs, although the author acknowledges that there is also a negative impact on them. The negative impact has been sufficiently (in terms of the scope of this work) displayed in the following section. Therefore, this chapter focuses on the positive sides by showcasing successful projects that impact the SDGs in positive terms.

SDGs are interdependent and therefore it is not possible to provide examples of projects that would address only one of the goals. Therefore, later in this section, the classification of Environmental, Economic and Social Impact proposed by Vinuesa et al. is employed,[99] instead of goal distinction of the UN.

[96] Department of Economic and Social Affairs Sustainable Development, United Nations, "The 17 Goals," SDGs UN, https://sdgs.un.org/goals [Accessed on November 17, 2020].

[98] Ricardo Vinuesa et al., "The role of artificial intelligence in achieving the Sustainable Development Goals," *Nature Communications* 11, no. 233 (2020):1, https://doi.org/10.1038/s41467-019-14108-y.

[99] Ibid. p. 2.

4.1.1 Social Impact

To increase the quality of life of people with disabilities, Hand Talk, a Brazilian company, created an app that uses AI to translate Portuguese into sign language.[100] The app aims to help Portuguese speaking people with deafness and hearing loss to understand their counterparts in a dialogue. Another example is Microsoft's Seeing AI[101] and Google's Lookout[102] apps that enable visually impaired people to identify elements (objects, people, text, etc.) present in their environment, thanks to voice assistant powered by automatic image recognition. To help foreign families in the U.S. integrate, Talking Points has built a translation system, that fosters communication between non-English speaking parents with their children's teachers at school.[103]

An important indicator of life quality is the support and care for the elderly. Accenture London Liquid Studios, together with Age UK, ran a pilot of HomeCare, a companion for the elderly to assist with everyday tasks, and living independently.[104] In this pilot, AI was applied to create a human-centered platform to provide support and assistance in areas such as health appointments, medicine reminders, grocery shopping, exercise, and staying connected with the close people.

Affordable and clean energy, also part of the Social Impact category, could be achieved by the development of smart grids, coordination of decentralized power plants, and peer-to-peer algorithmic energy trading. Platforms such as eFriends Energy, an Austria-based energy supplier, that utilizes the methods of peer-to-peer energy trading to manage the excess energy.[105]

Artificial intelligence has a significant impact on the education, too. In developing countries, there are over 773 million people illiterate, of whom the majority is women.[106] To address this problem in Africa, two online ed-tech platforms have been put into operation. Daptio analyzes student's weaknesses and strengths and adjusts the online curriculum to his or her preference.[107] A mobile learning app, Eneza Education, provides lessons and assessments to the students through web communication or SMS messages and the students can also ask teachers their questions in a live chat.[108]

Health & well-being, the third development goal, is also constantly being worked on. Starting with mental health and suicide prevention, The Trevor Project uses sentiment

[100] https://www.handtalk.me/en [Accessed on November 17, 2020].
[101] https://www.microsoft.com/en-us/ai/seeing-ai [Accessed on November 17, 2020].
[102] https://support.google.com/accessibility/android/answer/9031274?hl=en&ref_topic=7513948 [Accessed on January 24, 2020].
[103] https://talkingpts.org/ [Accessed on November 17, 2020].
[104] Accenture Applied Intelligence, *Realising the economic and societal potential of responsible AI in Europe* (Accenture, 2018), p. 8.
[105] https://www.efriends.at/ [Accessed on January 23, 2021].
[106] "Literacy", UNESCO, http://uis.unesco.org/en/topic/literacy [Accessed on February 6, 2021].
[107] Linsey Alexander, "Companies providing AI tutoring in Africa," July 23, 2020, https://borgenproject.org/tag/daptio/ [Accessed on February 3, 2021].
[108] Ibid.

analysis and natural language processing to determine the risk of suicide in LGBTQ youth.[109] To help diagnose a disease of civilization, cancer, a Harvard-based team of researchers created an AI-based technique to help oncologists identify breast cancer cells with greater precision than doctors that did not use the technique. Doctors that used the AI-technique were able to accurately identify 99.5% of cancerous biopsies.[110] With nearly 1.7 million new cases of breast cancer diagnosed globally each year, this research result could help yearly from 68,000 to 130,000 more women to receive accurate diagnoses.[111] Next, Powerful Medical, a Slovakia-based company, develops a solution to support doctors, primary non-cardiologists, to strengthen early diagnosis of cardiovascular diseases and competent decision making in primary care.[112] Apps like this can help to even the gap and increase the healthcare quality in developing countries, as knowledgeable doctors and medical staff are rarely available.[113]

4.1.2 Economic Impact

Writing job descriptions is not a straight-forward process as it is supposed to address a wide audience, which in the field of technology, is still dominated by men. It has been proved that the diversity in a company has a positive impact on the company's revenues (up to 19% revenue increase),[114] therefore it is in a company's interest to address all potential applicants with relevant education, skills, or experience, regardless of their gender or other characteristics. Atlassian, a software company developing products for software development teams, used Textio, a smart text editor capable of making a job description more inclusive, increasing the percentage of women from 10% to 57% in two years.[115] Employing more women in technical positions could also partially lead to reducing the pay-gap, as tech jobs are generally better paid.

4.1.3 Environmental Impact

To increase food security, AI is applied in the field of agriculture (precision agriculture) to improve harvest quality and accuracy. The goal of precision agriculture is to help

[109] https://www.thetrevorproject.org/ [Accessed on November 18, 2020].

[110] Ellyn Shook and Mark Knickrehm, *Reworking the Revolution* (Online: Accenture Strategy, 2017), p. 7, https://www.accenture.com/_acnmedia/PDF-69/Accenture-Reworking-the-Revolution-Jan-2018-POV.pdf [Accessed on January 23, 2021].

[111] Ibid.

[112] https://www.powerfulmedical.com/ [Accessed on January 23, 2021].

[113] Day Translations Team, "How AI is Helping Undeveloped and Developing Countries," Day Translations, November 31, 2018, https://www.daytranslations.com/blog/helping-undeveloped-countries/ [Accessed on February 7, 2021].

[114] Rocio Lorenzo et al., "How diverse leadership teams boost innovation" (Online: The Boston Consulting Group, 2018), p. 2, https://image-src.bcg.com/Images/BCG-How-Diverse-Leadership-Teams-Boost-Innovation-Jan-2018_tcm9-207935.pdf [Accessed on January 23, 2021].

[115] Tim Halloran, "How Atlassian went from 10% female technical graduates to 57% in two years," Textio, December 12, 2017, https://textio.com/blog/how-atlassian-went-from-10-female-technical-graduates-to-57-in-two-years/13035166507 [Accessed on January 23, 2021].

detect diseases of plants, animals, to detect pests, and to measure nutrition indicators of the plants and soil. It is also possible to predict weather conditions to plan the season. Wadhwani AI uses image recognition to track pests to advise farmers about the amount of pesticides they use, intending to reduce the amounts.[116]

Another example from the innovative farming field is Plenty, which employs the latest tech like IoT sensors and machine learning to grow crops vertically indoors using only light, water, and nutrients. Its system reputedly uses only 1% of the water that is wasted in conventional farming.[117]

4.2 Negative impact of the AI systems on the society

4.2.1 Right to education: Government's final grade assessment algorithm and its negative impact on

In 2019, with the outrage of the COVID-19 pandemic, the schools around the world have had difficulties facilitating the final exams of their graduates. In many countries, the exams did not take place, eg. no written exams in Slovakia,[118] in others, the oral exams were canceled, eg. in Austria.[119] Instead of the usual assessment approaches, the final grades were calculated based on various statistical methods. In some cases, the statistical methods showed to be accurate and accepted by the assessed students and their parents. On the other hand, some of these approaches caused a public outcry and legal actions.

In the United Kingdom, the Office of Qualifications and Examinations Regulation (Ofqual), a non-ministerial government department that regulates qualifications, exams and tests, has developed a grade estimation tool that would substitute the final exams of thousands of students that graduated in 2020.[120] Instead of examining the students knowledge, their final grades were estimated based on their previous performance and the estimated grade by their teachers. Despite fair intentions and reported discussion with relevant stakeholders, the algorithm predicted lower grades than what the teachers would expect their students to get in 39.1% cases.[121] Such approach can be considered

[116] https://www.wadhwaniai.org/ [Accessed January 23, 2021].

[117] https://www.plenty.ag/about-us/ [Accessed on January 23, 2021].

[118] Ministry of Education, Science, Research and Sport of the Slovak Republic, "Opatrenia ministerstva školstva - písomné maturity sú zrušené," Minedu.sk, March 24, 2020, https://www.minedu.sk/opatrenia-ministerstva-skolstva-pisomne-maturity-su-zrusene/ [Accessed on December 20, 2020].

[119] Lisa Nimmervoll, "Im Corona-Jahr wird Maturanten die mündliche Prüfung erlassen," Der Standard, April 7, 2020, https://www.derstandard.de/story/2000116619608/im-corona-jahr-wird-maturanten-die-muendliche-pruefung-erlassen [Accessed on December 20, 2020].

[120] Richard Adams, Sally Weale and Caelainn Barr, "A-level results: almost 40% of teacher assessments in England downgraded," The Guardian, August 13, 2020, https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded [Accessed on December 20, 2020].

[121] Ibid.

profiling in terms of GDPR, as introduced in Chapter 3. Accurate prediction for A-levels was substantial, as the students in the UK apply to universities before taking the exam and obtain a conditional offer. The offer can be revoked if the final grade is lower than predicted by the student's teacher.

The following input parameters were taken into account to calculate student's grade per subject:[122]

- Statistical distribution of grades per subject and school from the three previous years.
- The rank of each student within their school, based on Centre Assessment Grade (CAG). It is the teacher's objective judgment of the grade that the student will most probably achieve in their final exams, based on the student's capabilities throughout their studies.
- The previous exam results for each student in question and also the students who graduated before 2020.

It has been reported that such an approach to grade prediction resulted in awarding a similar grade to a student that graduated years before the 2020 and had a similar ranking within their school.[123] The grade is based upon the assumption that if a student scored low in the past, he or she would score low on their finals. Due to the nature of the profiling, the score is additionally discriminative in a way that it compares the student's past to the grades of students who graduated the years before 2020, and takes the final grades of "similar" students into account.

To design such a system to provide value to the society, instead of causing harm, better decisions must be taken that result from open discussion with relevant stakeholders. It must be clear who is responsible for the result, shall it be success or failure.

Roger Taylor, the Chairman of Ofqual claimed that their "goal has always been to protect the trust that the public rightly has in educational qualifications."[124] As a consequence, Taylor took the responsibility and decided to step down from his position.[125]

However, such failure of a system has a significant impact on reputation and public trust in the algorithmic systems. The failure to design the algorithm well would have certainly resulted in discrimination and indirect violation of the right to education -

[122] "A-levels: How controversial algorithm behind moderation row works," Sky News, August 16, 2020, https://news.sky.com/story/a-levels-how-controversial-algorithm-behind-moderation-row-works-12048780 [Accessed on December 20, 2020].

[123] Ibid.

[124] BCS, The Chartered Institute for IT, *The exam question: How do we make algorithms for the right thing?"* (Swidon, England: BCS, 2020), p. 8.

[125] Hannah Richardson, "Ofqual chief Sally Collier steps down after exams chaos", BBC, August 25, 2020, https://www.bbc.com/news/education-53909487 [Accessed on February 3, 2021].

rejected offers from the universities for which the students have sufficient skills but were falsely underestimated. Evaluating students based on their profiles, rather than their real knowledge, discards their opportunity to change their future under the consideration that if a student has enough time and resources, they can prepare for the finals and outperform their grade history and decline the similarity of their results in regards to the alumni from the previous years.

Such unsuccessful grade predictions were not only made in the UK's school system but also worldwide in the International Baccalaureate Diploma Programme.[126]

4.2.2 Equality and right to job: Hiring algorithms preferring male candidates over female ones

It is a challenge to match open positions with candidates that are a perfect fit for a given role. It is also rarely the case that a person looks for a new job, it is approximately every 3 to 5 years (the median number of years, in general, is 4.6 years) and the job-hopping ratio is very dependent on the location, age and occupation.[127] To reduce the time-to-hire and the cost-to-hire, it is therefore necessary to innovate also in the field of hiring, such as targeted advertisement or automatic pre-screening of the candidates.

However, as soon as the automated processing comes into place, it brings about the discrimination and certain bias. There are three well-known cases of such discrimination in the hiring process - Amazon's tool for automatic pre-screening, Google's targeted advertising for high paid jobs tendentially shown to men, rather than women; and an old case from the 1970s, a discriminative admission process of the St. George's Hospital Medical School in London.

In 2014, Amazon Inc. built an in-house tool for the job applicant's resume review, based on the natural language processing to pre-screen the candidates that would fulfill the initial criteria.[128] This tool was rating the employees on a scale of five stars. Eventually, it came to the attention of the employees that the ratings are not gender-neutral and after extensive analysis, it was found that the training data (the data from the previous 10 years of the successfully hired people) contained hidden bias.[129] What is now obvious, was not obvious back then, when the tool learned that the majority of the applicant are men and therefore more men were eventually a good fit and hired - and this is the decision making that the HR tools was trained to mimic.

[126] Theodoros Evgeniou, David R. Hardoon, and Anton Ovchinnikov, "What Happens When AI is Used to Set Grades?" Harvard Business Review, August 13, 2020, https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades [Accessed on December 20, 2020].

[127] Alison Doyle, "How Long Should an Employee Stay at a Job?" The Balancecareers, November 8, 2019, https://www.thebalancecareers.com/how-long-should-an-employee-stay-at-a-job-2059796 [Accessed on January 23, 2021].

[128] Akhil Alfons Kodiyan, *An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool* (2019):1, https://www.academia.edu/42965919/An_overview_of_ethical_issues_in_using_AI_systems_in_hiring_with_a_case_study_of_Amazons_AI_based_hiring_tool [Accessed on January 23, 2021].

[129] Ibid.

A similar situation had been happening for a decade in the 1970s in London, in course of the St. George's Hospital Medical School's admission process, until it was revealed in the late 1980s.[130] It was found that the system generated lower score for women and individuals from the racial minorities, although, on the application, there were no direct information that would identify the race of the applicant. This data was extracted from the name of the candidates and their places of birth. Back then, it was usual that a woman would take some time off due to family commitments, as well as a foreigner would have difficulties with colloquial and technical terms. At that time, only 17.5% of all the applicants would be offered a place in the cohort per year, so these two factors would be deciding factor to filter out students that would tendentially not succeed in their careers due to the aforementioned reasons.[131] Later, the system would learn this pattern and reproduce it for the future decisions. After this issue was revealed, the School contacted potentially discriminated applicants and re-did the process again, resulting in some applicants getting the offer.[132]

A more recent case, however not as significant, is the targeted job advertising powered and sold by Google. The researchers from Carnegie Mellon University conducted an experiment with series of fake accounts, that were identical (also in terms of the search of history) - with a single exception of the gender.[133] From all the high-paying jobs, the ads were shown to the male group of the fake users 1852 times, whereas to the female group only 318 times.[134] Looking objectively at this problem, this discrimination does not necessarily have to be caused by biased model or the training data (collected from the group of users where women were not interested in high-paying jobs) that Google used, but also the advertiser that paid for the targeted advertisement services could have set up the target group that indicated that male candidate would be preferred over the female ones.[135]

4.2.3 Right to fair trial and due process: US' predictive policing system biased towards citizens of African American origin

When designed properly and trained on unbiased data, the use of an evidence-based risk assessment tool in predictive policing can have a positive impact on the fairness and consistency of court decisions. Based on the machine learning fundamental feature - learning from the previously seen data and predicting outputs based on similar input characteristics - such a system could result in sentencing fairness. Similar charges under

[130] Stella Lowry and Gordon Macpherson, "A blot on the profession," *British medical journal* 296, no. 6623 (1988):657.

[131] Ibid.

[132] Ibid.

[133] Julia Carpenter, "Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you." The Washington Post, July 6, 2015,https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/ [Accessed on January 23, 2021].

[134] Ibid.

[135] Ibid.

similar circumstances would result in similar convictions and similar sentences. Despite this consideration, the same feature of the machine learning could result in the violation of the European Charter of Human Rights, concretely The Right for Fair Trial.

The evidence-based risk assessment tools have already been tested and put into operation. In the UK, the Harm Assessment Risk Tool for predict recidivism and foster consistency in the court's decision making has been tested.[136] Similar tool has been introduced in the US, causing a public discussion whether the rights of a defendant has been violated, after his request to gain access to the methodology used in the design and implementation of the algorithm has been rejected.[137] In this case from 2016, in the state of Wisconsin, US, an evidence-based risk assessment tool Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was used to predict a recidivism risk of the defendant, Eric L. Loomis, as a part of a pre-sentencing investigation report.[138] The predictive tool caused controversy and resulted in the defendant filing a motion for post-conviction relief after the trial court referred to the assessment result of the intelligent system in their decision.[139]

Mr. Loomis was charged for the following crimes, of which in (1), (2), (4) and (5) he was a Party to a Crime:[140]

- (1) First-degree recklessly endangering safety
- (2) Attempting to flee or elude a traffic officer
- (3) Operating a motor vehicle without the owner's consent
- (4) Possession of a firearm by a felon
- (5) Possession of a short-barreled shotgun or rifle

He admitted that he drove a car without the owner's consent and later attempted to flee a traffic officer, but denied the other three charges.[141] The COMPAS assessment showed that Mr. Loomis was of high risk in all three recidivism categories - pretrial recidivism, general recidivism and violent recidivism.[142] In response, the defendant filed a motion for post-conviction relief because the court relied on a tool that allegedly "infringed on both his right to an individualized sentence and his right to be sentenced on accurate

[136] Marion Oswald et al., "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality," *Information Communications Technology Law* 27, no. 2 (2018), http://www.doi.org/10.1080/13600834.2018.1458455.

[137] State v. Loomis, 881 N.W.2d 749 (Wis. 2016) at 46, https://www.leagle.com/decision/inwico20160713i48 [Accessed on February 6, 2021].

[138] Ibid. at 755.

[139] "State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing," *Harvard Law Review* 130, no. 5 (2017).

[140] Loomis, 881 N.W.2d at 755.

[141] Loomis, 881 N.W.2d at 755.

[142] Ibid. at 756.

information".[143] Mr. Loomis also argued that the tool has violated his due process rights, by violating his right to be sentenced based upon accurate information (which he could not assess because of the proprietary nature of COMPAS), violating his right to individualized sentence, and that the tool is biased towards his gender.[144]

The court later rejected the arguments, stating that the gender served non-discriminatory purposes[145] and that the individualization of his sentence was guaranteed through the individual and objective approach of the judge to whom the tool was supposed to assist.[146] It was also accentuated that the judge has a full power to overrule any intelligent tool and make a decision that completely differentiates from the predicted decision.[147] This, however, is a very vague claim. The public has a belief that technology is unbiased, as it does not possess consciousness and feelings. The truth is that technology reflects the bias and beliefs of its designers and developers, whether or not they incorporate their subjective views purposely.[148] Even if striving for inclusion and objectivity in the course of development, the data sets used to train the models have their flaws and could potentially learn to discriminate certain groups of people because the algorithms observe unapparent discriminative tendencies in the training data.[149]

Independent investigative journalists at ProPublica, serving the public interest, have analyzed the COMPAS recidivism algorithm and published a report on their website.[150] For brevity and to depict the problems of COMPAS algorithm, the following paragraph summarizes ProPublica's findings. Their findings show certain biases in the predicted risk score of the people of color. To analyze the bias, ProPublica developed a machine learning model based on logistic regression, that took into account various demographic and socio-economic characteristics such as race, age, criminal history, future recidivism, charge degree, and gender. By adjusting the parameters to evaluate how the output changes based on the above-mentioned characteristics, ProPublica found that:

- Histogram distribution of the COMPAS decile score of the black defendants was distributed quite evenly among 10 categories - 1 (low risk) and 10 (high risk), whereas the distribution of the COMPAS decile score of the white defendants was tendentially converging towards zero as the risk category increased. In other words,

[143] State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing," p. 2.
[144] Loomis, 881 N.W.2d at 758.
[145] Ibid. at 755–58.
[146] Ibid. at 759-61.
[147] Ibid.
[148] Lee Rainee and Janna Anderson, "Theme 4: Biases exist in algorithmically-organized systems," *Code-depentent: Pros and Cons of the Algorithm age* (Washington, DC, USA: Pew Research Center, Internet & Technology, 2017), p. 57.
[149] Ibid.
[150] Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [Accessed on January 6, 2021].

the highest count of the white defendants was in the lowest risk category (1), while the lowest count of the white defendants was in the highest category (10).

- Black defendants' high-risk score was predicted false positive in 45% of the cases, while the white defendants only 23% of the cases.
- White defendants' low-risk score was predicted false positive in 48% of the cases, whereas the black defendants were falsely predicted to be low risk in only 28% of the cases.
- Age was a significant factor, ProPublica analyzed that defendants younger than 25 years old were 2.5 times as likely to get a higher score than older defendants.

It remains for further discussion whether or not, and to what extent, a judge could be subconsciously influenced by automatic decision-making tool in the belief that a machine must be right and therefore falsely decide in accordance to the tool's false output instead in a favour of the defendant. It is also a matter of question how the use of a tool and its reliability is communicated to a judge that has no technical background and possibly does not have an understanding of how machine learning works.

CHAPTER 5

Responsible AI: Principles and challenges of its development

Objective of this chapter is to research the principles of the *Responsible AI (RAI)* and state of the art approaches to achieve it. Based on the previous Chapter 4, the current section concludes that there is a need for Responsible AI that adheres to the legal standards and does not hinder citizens in exercising their rights.

Responsible AI is not a new term introduced by this thesis, it has been introduced in the context of AI a couple of years ago. The leading technology companies[151] around the world and the research community itself makes efforts to develop scientific methodologies to solve the problem on technological level. This finding is also supported by the recent approach of Conference on Neural Information Processing Systems (NeurIPS) who demanded that the researchers who submit their work also prepare a statement of their research's impact on ethics.[152] Additionally, Association for Computing Machinery organizes ACM Conference on Fairness, Accountability and Transparency every year since 2018.[153]

It is worth mentioning that different authors manifest their definitions of the RAI. However, many of the reviewed definitions have several principles in common, such as transparency, trustworthiness, interpretability, and accountability.

[151] Example: Google (https://ai.google/responsibilities/responsible-ai-practices/), Microsoft (https://www.microsoft.com/en-us/ai/responsible-ai-resources) [Access on January 11, 2021].

[152] Neural Information Processing Systems Conference (NeurIPS), "Getting Started with NeurIPS 2020," Medium, 2020, https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28. [Accessed on January 12, 2021].

[153] ACM FAccT Conference, "ACM Conference on Fairness, Accountability, and TTransparency (ACM FAccT)," FAccT Conference, https://facctconference.org/index.html [Accessed on January 21, 2021].

In the first section, this chapter defines Responsible AI as a superclass of eight properties that the research community considers important to achieve responsible design, development, and deployment of the AI. In the next section, Ethical and Lawful AI is compared in terms of effectiveness. One of the biggest challenges of the RAI is to develop machine learning models that are understandable to a human observer. Therefore, we first demonstrate the urgency of understandable models and define fundamental terms in this research field. In Section 5.4, the research problems of understandable models are introduced and the scientific state of the art approaches to solve the problem are summarized. Finally, the chapter discusses the organizational approaches to achieve Transparent, Trustworthy, and Accountable AI.

5.1 Principles of Responsible AI

Although the term Responsible AI suggests that it is the machine who is responsible, it is, however, the responsibility of the machine's creator for the development of such a system that adheres to the fundamental values and principles and ensures human well-being in a sustainable world.[154] To fulfill these goals, the AI should take societal, moral, and ethical values into account, explain its reasoning, provide transparency and respect values held by stakeholders with various cultural backgrounds.[155] To ensure such behavior of the systems, the creators should be held accountable and have frameworks available that guide them towards achieving Responsible AI.

After reviewing the literature, this chapter defines Responsible AI as depicted in the Figure 5.1. This work defines the principles as follows:

Lawful Quality of a system not to carry out tasks that are abhorrent with legal frameworks and guidelines that touch on (human) rights specified in the Chapter 3.

Ethical Quality of an entity to behave in correspondence to moral principles defined by the society.

Explainable Explainability is the extent to which a system, its reasoning, and decision making explains (that is represented in a way that it is understandable by human) on internal mechanics of a complex (deep learning) system.[156]

Interpretable Interpretability is the extent to which the model; prediction; and cause and effect are observable by a human so that they can follow what is happening,

[154] Andrea Aler Tubella et al., "Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19),* (Macao, China: IJCAI, 2019), https://doi.org/10.24963/ijcai.2019/802, p. 2.

[155] Ibid.

[156] Richard Gall, "Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI," KDnuggets, https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html [Accessed on January 21, 2021].

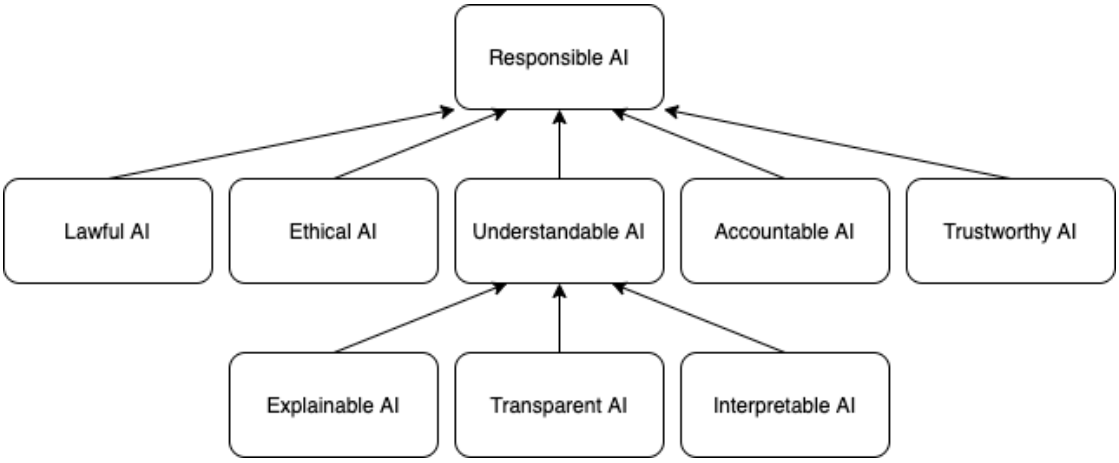


Figure 5.1: Principles of Responsible AI

and predict the output if an input changes, such as when working with decision trees.[157]

Transparent The system exposes the information about the author, internal functioning and algorithms, intended use, business model, training data, and other relevant information that helps stakeholders understand the system, its strengths and weaknesses, to the extent that provides value to the stakeholders, but at the same time does not violate the intellectual property rights of the creators.

Trustworthy Quality of a system to behave expectedly, so that the users trust it and rely on it. This could be connected with the system's accuracy and predictability of its outputs.

Accountable Quality of a creator of the system to take ownership for the machine's actions and its results, in terms of putting efforts to identify possible problems and take actions to prevent them.

Note: In the literature, interpretability and explainability are often being used interchangeably. This work, however, distinguishes between the two. To better demonstrate the difference, we provide the following example: If a person wants to make an ice cream, it is enough to follow how another person carries out this task, what is the sequence of the steps, what ingredients are used, what temperature is necessary, etc. The person is at any time capable of repeating and interpreting the process, and also to hypothesize why exactly a certain temperature or amount of the ingredients is necessary. However, if a person understands that if the temperature would be lower, the ice cream would melt, or that the ice cream would not taste well if some steps would be left out, we refer to it as explainability.

[157] Riccardo Guidotti et al., "A Survey of Methods for Explaining Black Box Models," In *ACM Computing Surveys (CSUR)* 51, no. 5 (2019):6, https://doi.org/10.1145/3236009.

Above defined qualities are mostly applied in the context of human behavior. As the development of intelligent systems is expected to achieve the AGI in the medium-term, these systems will become more proficient in mimicking the human brain to the extent that their decisions or behavior is indistinguishable from the human behavior, as described in the Section 2.1. If a machine's intelligence is comparable to the intellectual capability of humans, it ought be evaluated also in regards to the aforementioned principles (and the creators ought to be held accountable for designing them this way). Although the rise of AGI is not yet happening, the topic is also relevant in regards to ANI, as analyzed in the Chapter 4.

5.2 Ethical and Lawful as a subclass of the Responsible AI

In the literature, the term *Ethical AI* is often discussed as an approach to design, develop, and deploy AI system in a way that mitigates the AI's potential negative impact on society. This thesis specializes in Responsible AI and explicitly analyzes the AI in the legal context (*Lawful AI)*, rather than researching the ethical context (*Ethical AI*).

The reason is that ethics is a social construct of a society or a community, and therefore it is not binding and legally enforceable. The violation of ethical or moral principles does not result in any penalty and the only outcome is public shame or rejection of the traitor in the community.

On the other hand, there is a relation between laws and ethical principles. Based on the ethical principles, the laws have been formulated.[158] Countless number of legally binding documents, such as conventions, regulations, and directives have been created in the European Union (for further information, refer to the Chapter 3) to provide the safeguards for the citizens and to mediate the relationships among them. These documents have either been adopted in domestic laws or directly effective in the Member States. To ensure that the natural and legal persons follow the rules, the committees and the courts have been established that oversee and judge the implementation of the guidelines.

Figure 5.2 depicts a comparison of the law and ethics, regarding binding, punishment for violation, governance, and other characteristics.

Based on this fact, this work considers Ethical as well as Lawful AI a subclass of the Responsible AI, but focuses on the latter, as legal frameworks provide more efficient measures to govern the responsible design, development, and deployment of the AI than an ethical framework would.

[158] Lumen Learning, "Introduction to Ethics, Chapter 3: Making Ethical Decisions, Ethics and Law" Lumenlearning, https://courses.lumenlearning.com/atd-epcc-introethics-1/chapter/ethics-and-law/ [Accessed on January 17, 2021].

[159] Surbhi S., "Difference Between Law and Ethics," Keydifferences, 13 August 2018, https://keydifferences.com/difference-between-law-and-ethics.html [Accessed on January 11, 2021].

BASIS FOR COMPARISON	LAW	ETHICS
Meaning	The law refers to a systematic body of rules that governs the whole society and the actions of its individual members.	Ethics is a branch of moral philosophy that guides people about the basic human conduct.
What is it?	Set of rules and regulations	Set of guidelines
Governed By	Government	Individual, Legal and Professional norms
Expression	Expressed and published in writing.	They are abstract.
Violation	Violation of law is not permissible which may result in punishment like imprisonment or fine or both.	There is no punishment for violation of ethics.
Objective	Law is created with an intent to maintain social order and peace in the society and provide protection to all the citizens.	Ethics are made to help people to decide what is right or wrong and how to act.
Binding	Law has a legal binding.	Ethics do not have a binding nature.

Figure 5.2: Comparison of the law and ethics[159]

5.3 Understandable: The urgency to understand black-box algorithms

As the current chapter defined in the Subsection 5.1, this work considers explainability, interpretability & transparency substantial principles of the Responsible AI. The Chapter 4 reviews the decision support systems that have been implemented as a *black box*, meaning that hide their internal logic to the user and how we have seen, it causes ethical issues.[160] It is questionable how tech makers and society can trust the product powered by machine learning if they do not know the underlying rationale.[161] This work also questions the acceptance of the society to believe in a software's decisions, whereas in a similar situation, the same person would most probably question such decision of a human counterpart. Additionally, it has been predicted that "by 2018 half of business

[160]Guidotti, "A Survey of Methods for Explaining Black Box Models," p. 1.
[161]Ibid.

ethics violations will occur through the improper use of Big Data analytics".[162]

To solve the issues, the black box models must be transformed into *glass box models*, providing insights into their functioning, which can be achieved by a critical audience demanding answers to their issues (as implied from the Subsection 4.2.1) and by (Responsible) AI governance (as implied from the EU's approach discussed in the Chapter 6). Enforcement of XAI does not only depend on legal and regulatory frameworks (see Section 6.3), but the scientific research is an integral part of the equation, too. Lawmakers create regulatory frameworks that must be easy to implement by the tech makers, otherwise the frameworks become hard to follow or in the worst case, completely obsolete. To ensure the usefulness of the frameworks, advancement of the emerging technologies *must* be taken into account and the research *must* be fostered to develop strategies for explaining black boxes.

With the advantage of understanding a black box, Doshi-Velez and Kim argue that interpretability can help to assess whether other goals such as fairness (non-discrimination), privacy (sensitive data protection), reliability & robustness (good accuracy independent of input), causality (expectation that a perturbation causes a change in the output), usability (easy to operate and assists human in executing their tasks) and trust (human feels comfortable to rely on the output) are met.[163]

The need for explanation stems from the presence of incompleteness in a problem definition.[164] In complex problems, that cannot be formalized as a finite set of possible states, (such as creating a list of all the possible scenarios in which a system could provide unexpected outputs,) explanations play an important role in pointing out on these unforeseen states.[165]

5.3.1 Explanation and its properties

To provide a better understanding of the topic, it is necessary to define what *explanation* means. In the literature, every research group provides a different meaning of explanation, because they research the problem from different perspectives. Throughout the literature review, three main approaches were found.

The first one is that an explanation is a model itself. Linear models and decision trees are the best examples. However, it is vague to claim that *all* of them are interpretable, as a simple model with hundreds of features and their weights cannot be easily comprehended by a human user, although a human can directly observe them.[166] Therefore, limitations

[162] Ibid.

[163] Finale Doshi-Velez and Been Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint* (2017):2, arXiv:1702.08608v2.

[164] Ibid. p. 4.

[165] Ibid.

[166] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations,* (San Diego, California, USA: ACL, 2016), https://www.doi.org/10.18653/v1/N16-3020, p. 2-3.

of human cognition should be considered.

Another common definition is that an explanation is an interface between a human user and an intelligent system that is "an accurate proxy of the decision-maker and comprehensible to human".[167] In these terms, the word proxy is introduced to represent an entity (explanator) that is supposed to provide insights into the intelligent system's functioning. In the research community, there is a mutual consensus that the only faithful explanation with 100% accuracy is the model itself, as defined in the paragraph above.[168] As it is not always possible to expose the whole system (e.g. because of its complexity that would not be understandable to human), the concept of proxy is widely used.

Others define explanation in more human terms, as a "textual or visual artifacts that provides qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction".[169]

No matter which explanation definition is taken into account, the quality of the explanation can be judged based on the following properties:[170]

- Accuracy: How well does the explanator predict the output? Is the explanator's accuracy comparable to the black box accuracy?
- Fidelity: How well does the explanator approximate the black box model? The fidelity represents the quality of the explanation and its faithfulness to the actual decision making of the black-box model.
- Consistency: If the same explanator is used to generate explanations for two different black box models that were trained on the same task and the same data, the explanations should be similar and therefore these explanations are consistent.
- Stability: How similar are explanations generated by the explanator model for two similar outputs of the black box model?
- Comprehensibility: Does the user understand the explanation?
- Certainty: Does the explanation include the information on how certain is the black box model about its prediction?
- Degree of Importance: Some explanations provide weights of features' importance. How well are the actual weights reflected in the model's explanation?

[167] Guidotti, "A Survey of Methods for Explaining Black Box Models," p. 5.

[168] Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," p. 3.

[169] Ribeiro, Singh and Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier," p. 3.

[170] Marko Robnik-Sikonja and Marko Bohanec, "Perturbation-based explanations of prediction models," *Human and Machine Learning* (Cham, Switzerland:Springer, 2018) p. 159-175. as cited in
Christoph Molnar, "Properties of Explanations," *Interpretable MachineLearning,* (Online:Leanpub, 2020), https://christophm.github.io/interpretable-ml-book/properties.html#fn8 [Accessed on January 21, 2021]

- Novelty: If a certain data point is new to the black-box model and very different from the previously seen data points, the accuracy and certainty are likely going to be decreased. Does the explanation reflect on this fact?

- Representativeness: Is the explanation global or local?

Additionally to the aforementioned properties of the explanations, some researchers assess the quality of the explanations on an observation, how well would a human explain its reasoning in comparison to an explanator model on the same task. To evaluate the quality of explanations in this regard, Phillips et. al. defined four principles of Explainable AI:[171]

- Explanation: a decision support system provides information and rationale regarding its output

- Meaningful: the provided explanation can be understood by a user addressed

- Explanation Accuracy: the explanation corresponds to the actual reasoning

- Knowledge Limits: the system only provides a decision if the confidence score of the output is above a threshold that was specified in the course of the system designs

Based on these four principles, Philips et al. put AI in comparison to a human mind and evaluate whether or not all of these principles can be fulfilled in the decision-making process of both entities. The conclusion from the research is that even a human cannot always guarantee a high quality of the above-mentioned conditions and are often unreliable. Nonetheless, human decision making can inspire "the development of benchmark metrics for explainable AI systems".[172]

When it comes to the requirements for the model's explanation, they heavily depend on the context in which the model is deployed, on the expected exhaustiveness of an explanation as well as on the audience. Although the models vary in terms of the context where they are deployed, common patterns in the explanations can be observed. To cluster requirements of the explanation, Doshi-Velez and Kim defined the following dimensions of the explanations:[173]

- **Global and local interpretability:** If a user understands the underlying logic and can extract the reasoning of any decision, it is referred to as global interpretability. If the internal logic of a model is unknown but the reasons for a specific decision are documented in comprehensive terms, it is referred to as local interpretability.

[171] P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence,* Draft NISTIR 8312 (Gaithersburg, Maryland, USA:NIST, 2020), 2, https://doi.org/10.6028/NIST.IR.8312-draft.
[172] Ibid.
[173] Doshi-Velez and Kim, "Towards A Rigorous Science of Interpretable Machine Learning," p. 7.

- **Area, Severity of incompleteness:** Incompleteness can be present in e.g. the definition of input, domain, internal model structure. This incompleteness could have different severity - in autonomous driving, a user might want to know how the decision making generally works (high incompleteness, low severity), or might be interested in a particular set of inputs that would cause a car to crash (low incompleteness, high severity).

- **Time limitation:** It is important to determine what is the time constraint for a user reading and understanding the explanations, as different situations allow different time resources. In a very exact context such as predictive policing, the judge has enough time to study the reasons behind any automatic prediction. In the case of driving an autonomous car, the driver must understand quickly why the car reacts in a certain manner.

- **Nature or user expertise:** The last dimension depends on the stakeholders. In every context and situation, different people require different precision of the explanation. More experienced people might require in-depth explanations, while less experienced people might expect explanations that are shallow and easy to reads.

5.3.2 Explainable vs Interpretable AI

In the AI research, there are two main approaches to ensure that the reasoning of the black box systems is provided. On one side, researchers work on the interpretable machine learning models (Explainable AI) that provide explanations of the decision making of other black-box models, while being interpretable themselves and providing insights into their functioning.[174] In other words, they develop additional AI models called *explanators*. On the other side, there is criticism of that XAI efforts. As an alternative, some researchers suggest developing algorithms, that are interpretable in the first place, so-called Interpretable AI (IAI), instead of black boxes, applying them at least for the high stake decisions, in fields such as medicine, predictive policing, social security system.[175]

Anyhow, these two approaches have been subject to research since decades.[176] Arrieta et al. charted the rise of the XAI and IAI in terms of number of research papers with keywords "Interpretable AI", "Explainable AI" or "XAI". As depicted in their work (Figure 5.3), over the past eight years the number of research papers in the field of IAI had been increasing exponentially, until 2018/2019, when the research of XAI started dominating.

[174] Phillips et al., *Four Principles of Explainable Artificial Intelligence,* p. 10.

[175] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019), https://doi.org/10.1038/s42256-019-0048-x.

[176] Xu, "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," p. 2.

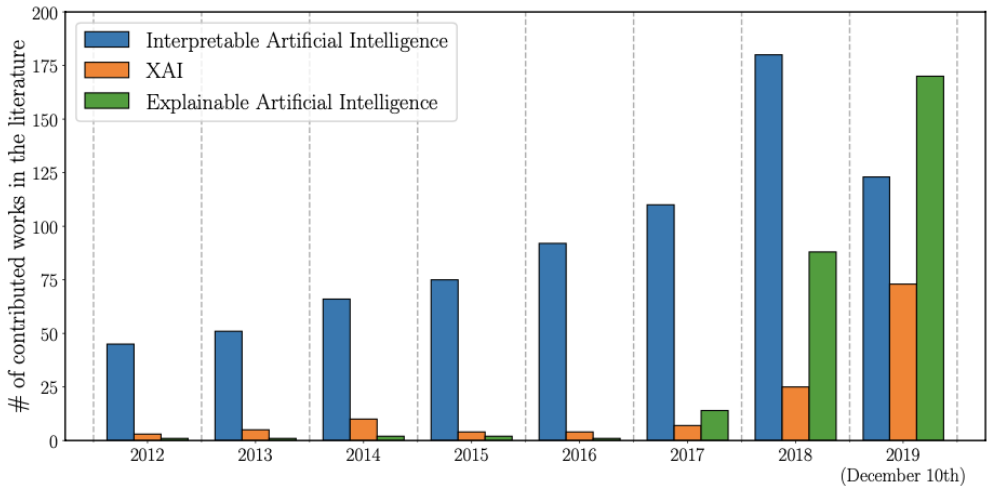


Figure 5.3: Rise of the XAI and IAI research[177]

For this work, the author assumes the term *Understandable AI* (Figure 5.1) as a sufficient keyword that represents the methods, approaches, and strategies to help human users understand the behavior of the system, its strengths and weaknesses, as well as the relation between the system's prediction (output) and its input.

5.3.3 Interpretable models vs post-hoc interpretability techniques

The core difference between post-hoc interpretability techniques and interpretable models is that an interpretable model provides insights into its functioning, and does not provide an explanation (e.g. a summary of the statistically most important input features for each decision) of why a certain output was generated, they are explanations themselves.[178] From such an interpretable model, a human observer can comprehend the decision-making process and follow its reasoning. Interpretability is therefore better applicable in the symbolic approach to AI, while the post-hoc interpretability techniques are better applicable in the sub-symbolic AI (also known as connectionist AI).

A post-hoc interpretability technique is the application of an interpretable model to explain a black box (by approximating its output), then the black box is known as an explainable model. The quality measure of such explanator is *fidelity*.[179]. Guidotti et al. define fidelity as a measure of how good the model is in the mimicking the black-box

[177] Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, (2020):3, Figure 1, https://doi.org/10.1016/j.inffus.2019.12.012.

[178] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems,* (Long Beach, California, USA: Curran Associates Inc., 2017), arXiv:1705.07874, p. 2.

[179] Guidotti, "A Survey of Methods for Explaining Black Box Models," p. 7.

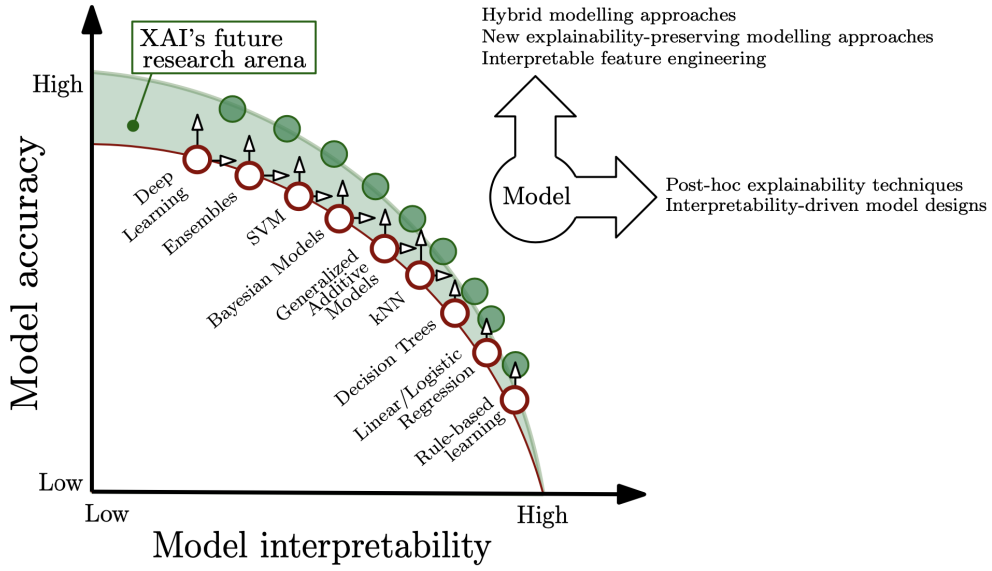


Figure 5.4: Trade-off between model accuracy and interpretability[182]

model and fidelity's quantification is expressed in terms of accuracy score.[180]

Symbolic vs. subsymbolic approach to AI The symbolic approach, is based on language-like representations (e.g. decision trees, logistic regression, Bayesian classifiers), while the subsymbolic (connectionist) approach, inspired by neuroscience (e.g. neural networks, deep learning).[181]

5.3.4 Accuracy vs. interpretability trade-off

Reportedly, the key issue in understanding the black box models is that with the higher algorithm complexity and performance (e.g. deep learning, SVM) the interpretability is getting lower, as displayed in the Figure 5.4. The research of Xu et al.[183] and Došilović

[180] Ibid.

[181] Chris Eliasmith and William Bechtel, *Encyclopedia of Cognitive Science,* s.v. "Symbolic versus Subsymbolic," (Somerset, New Jersey:John Wiley & Sons, Inc., 2006), https://doi.org/10.1002/0470018860.s00022.

[182] Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies,Opportunities and Challenges toward Responsible AI," p. 31, Figure 12.

[183] Feiyu Xu et al., "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," in *Proceedings of Natural Language Processing and Chinese Computing,* (Zhengzhou, China: Springer, Cham, 2019), http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-030-32236-6_51.

et al.[184] also assumes the correctness of this figure. There is however criticism of this claim. Rudin claims that if the data is structured well and only meaningful features are considered, mostly there is no significant difference in performance between sub-symbolic and symbolic classifiers.[185] Rudin also argues that in the field such as computer vision that uses deep learning heavily due to its performance, interpretability can be imbued into the models in a way that the accuracy is not compromitted.[186] Guidotti et al. states that to explain such predictive systems that employ images, transformations using equivalences, approximations or heuristics could be used to provide the interpretation of the model and/or the prediction.[187]

[184]Filip Karlo Došilović, Mario Brcić and Nikica Hlupić, "Explainable artificial intelligence: A survey," in *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* (Opatija, Croatia: IEEE, 2018), https://doi.org/10.23919/MIPRO.2018.8400040.

[185]Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," p. 2-3.

[186]Ibid.

[187]Guidotti, "A Survey of Methods for Explaining Black Box Models," p. 11.

5.4 Understandable AI as a subject of the scientific research

To understand the black box models, the research problems can be classified into the following categories:

- **Black Box Model Explanation Problem:** provides an interpretable predictor (explanator) that approximates the result of a black-box model with a high fidelity while providing insights into its functioning.
- **Black Box Outcome Explanation Problem:** provide an interpretable model that returns an output approximated to the output of the black box together with an explanation of the outcome, while it is not required to provide insights into the black box logic and does not necessarily have to be generalized to other cases (this is also referred to as local interpretability).
- **Black Box Inspection Problem:** provide an interpretable model that creates a representation (visual or textual) either for understanding the internal logic of the black box or reason why it returns certain predictions more likely than any other (and therefore is globally interpretable).
- **Transparent Box Design Problem:** provide a model which is locally or globally interpretable without any additional explainator.

If a model solves at least one of the aforementioned problems, it is "able to open a black box".[188]

To solve the above-mentioned problems, Phillips et al. define the following types of explainable algorithms which address each of the research problems:[189]

- **Self-Explainable Models:** these are the models that are *interpretable*, as they are mostly simple and follow logic that is implemented and therefore understandable by human.[190] Following the definitions of the aforementioned problems, these models represent the Transparent Box Design Problem. An example of such a model is Linear Regression, Generalized Additive Models, Bayesian Classifiers, and Decision Trees.
- **Global Explainable AI Algorithms:** these are the models that Rudin,[191] and Lundberg with Lee[192] refers to as models that approximate black-box models and

[188] The content of the parabox is cited from Guidotti, "A Survey of Methods for Explaining Black Box Models," p. 12. (the four research problems) and p. 16. (definition of black-box opening)
[189] Phillips, "Four Principles Of Explainable Artificial Intelligence," p. 6.
[190] Ibid.
[191] Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," p. 4.
[192] Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions," p. 2.

hereby explain it (post-hoc interpretability techniques). Following the definitions of the aforementioned problems, these models represent the Black Box Model Explanation Problem or Black Box Inspection Problem. An example of such an algorithm is SHAP.

– **Per-Decision Explainable AI Algorithms:** similar to the Global Explainable AI Algorithms, these also approximate a black box model and a decision that the black box model made, however this particular type of explainable AI is not required to generalize to other cases.[193] Following the definitions of the aforementioned problems, these models represent the Black Box Outcome Explanation Problem. An example of such an algorithm is LIME.

5.4.1 Model-specific and model-agnostic

Another distinction of the XAI approaches is classified using two dimensions:

- model-specific technique: These techniques are constrained to provide an explanation to a specific model.[194] Per se, self-explainable models are considered model-specific, as they only explain a specific class of models.[195]
- model-agnostic technique: These techniques are applied to solve the Black Box Outcome Explanation Problem, as they take an input and a prediction of a black box model into account and generate explanation.[196]

SHapley Additive exPlanations (SHAP) is one of the model-agnostic global interpretability models, that interprets predictions of black boxes. SHAP is a universal explanator, that combines other existing approaches - LIME, DeepLIFT, Layer-Wise Relevance Propagation, Classic Shapley Value Estimation.[197] By combining these approaches, the authors claim that SHAP is a universal solution that addresses the problem that it is often not clear which of the existing solutions are preferable in what context.[198]

Local Interpretable Model-Agnostic Explanations (LIME) is also a model-agnostic technique but specializes in the local interpretation of classifiers. Similar to SHAP, it provides importance values of each feature, as shown in the Figure 5.5. In this figure, a black-box model has returned a prediction of the patient's diagnosis based on the symptoms. LIME takes that prediction and symptoms as an input and returns an ordered list of the most important features (green color) and the least important features (red

[193] Phillips, "Four Principles Of Explainable Artificial Intelligence," p. 6.
[194] Christoph Molnar, "Taxonomy of Interpretability Methods," in *Interpretable Machine Learning*, (Online:Leanpub, 2020), https://christophm.github.io/interpretable-ml-book/taxonomy-of-interpretability-methods.html [Accessed on January 21, 2021].
[195] Ibid.
[196] Ibid.
[197] Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions," p. 1.
[198] Ibid.

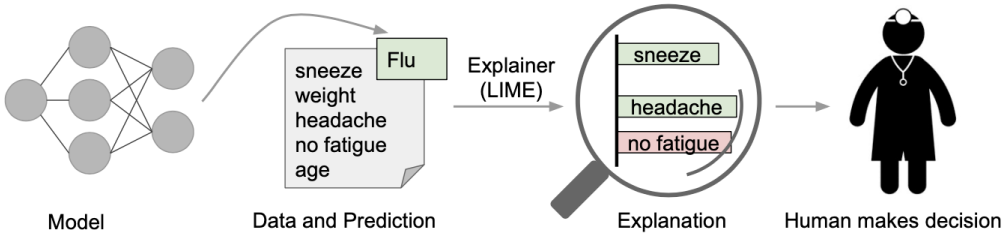


Figure 5.5: LIME: explanation of the diagnosis prediction[199]

color) for each explanation. Now, the doctor can make a well-informed decision and knows whether or not to trust the black box.

5.4.2 Explanations impact people's trust in s model and trust in an explanation

Based on the explanation's first dimension, the local and global interpretability, two levels of people's trust in the AI can be defined - the *trust in model* and the *trust in the explanation*.[200]

The quality and accuracy of the model contribute to the user's and developer's trust and decision whether or not to deploy such model. Additionally, a good explanation, which is on one hand understandable to the human and of high fidelity on the other, also influences people's trust in it and consequently the confidence they have in putting the model into operation.[201] Analogically, an explanation of the reasoning behind a prediction can influence a user's trust in the prediction and consequently whether or not the user takes a prediction into account in his decision making.[202]

5.5 Vulnerabilities of understandable AI

Although Understandable AI opens the black box and therefore enable all the stakeholders that work with the machine learning model to make well-informed decision whether or not to rely on the model, it also has certain shortcomings. With high-quality explanations, the black-box model can be reconstructed.[203] The model reconstruction stands in conflict with the confidentiality of the trade secret.

[199] Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," p. 2.
[200] Ibid. p. 6-7.
[201] Arun Rai, "Explainable AI: from black box to glass box ," *Journal of the Academy of Marketing Science* 48 (2020):3, https://doi.org/10.1007/s11747-019-00710-5.
[202] Ibid.
[203] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, Moritz Hardt, "Model Reconstruction from Model Explanations," In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* (New York, New York, USA: Association for Computing Machinery, 2019), https://doi.org/10.1145/3287560.3287562.

Another shortcoming of explainability is that an explainable model is vulnerable to adversarial attacks.[204] Adversarial attacks are attacks on the model, when the attacker either purposely or accidentally perpetuates the input data in a way that he gets targeted outputs. One example of an adversarial attack could be to manipulate the looks of an IBAN (by adding noise or inconspicuous hand-writing) before it is scan by a banking app to look like someone else's IBAN so that the algorithm falsely identifies it. Examples of such vulnerable models are LIME and SHAP.[205] The reason for that is that both of the models learn from perturbations of the input data to mimic a black box (that was trained on biased data set). Slack et al. propose a technique that hides the bias of any classifier in a way that can easily fool the two models to generate explanations that do not uncover the bias.[206]

[204] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." *CoRR* abs/1412.6572 (2015).

[205] Dylan Slack, et al., "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods," In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20),* (New York, New York, USA:Association for Computing Machinery, 2020), https://doi.org/10.1145/3375627.3375830

[206] Ibid. p. 1.

5.6 Transparent: Datasheets and Certificates

Apart from the algorithmic transparency that can be achieved by turning black box models into glass box models (so that their internal logic is visible to the user, as described in previous sections in this chapter), transparency can also be achieved by organizational measures. Researchers have proposed a Model Reporting approach to transparency. Others have proposed certifications of the AI. This chapter briefly summarizes both approaches.

5.6.1 Model Cards for Model Reporting

Mitchell et al. suggest that machine learning models should be deployed and published with accompanying "datasheets".[207] Datasheets are commonly used to provide the specifications such as hardware details, contents, materials, etc., for a particular product. Such documents are not yet in use for machine learning models.[208] In their research, Mitchell et al. suggest that these datasheets contain the information about the following (but not exclusively) points:

- Model Details
- Intended Use
- Factors
- Metrics
- Evaluation Data
- Training Data
- Quantitative Analysis
- Ethical Considerations
- Caveats and Recommendations

This work considers the model cards a good approach to secure transparency of the models and to provide the stakeholders with relevant information.

[207] Margaret Mitchell et al., "Model Cards for Model Reporting," *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19),* (New York, NY, USA:ACM, 2019), https://doi.org/10.1145/3287560.3287596.

[208] Ibid. p. 1.

5.6.2 AI Certification

As an alternative to the Model Cards, we found a white paper that describes the need for an AI certification system based on auditing of 6 areas of the AI models. Cremers et al. stated in the white paper that they planned to publish a Catalogue of Requirements for such a certification system in early 2020.[209] Throughout the literature review, we found that the plan has been postponed to 2021, therefore we cannot analyze the certification system in this work.[210] However, this could be an interesting topic for the future work. In the white paper, the authors suggest that the certification system takes the following areas into account when auditing machine learning models, which we find reasonable:[211]

- Autonomy and Control
- Fairness
- Transparency
- Reliability
- Security
- Data Protection

Additionally, the auditing system is going to take into account the ethics and regulations; current research, and the advancements in the field of AI; and go through a specific test and calibration.[212]

The certification system is being worked on by Fraunhofer Institute for Intelligent Analysis and Information Systems in cooperation with the German Federal Office for Information Security BSI.[213]

Alternatively to the European approach, the World Economic Forum, AI Global, and the Schwartz Reisman Institute for Technology and Society at the University of Toronto announced the launch of an independent working group to develop a globally recognized certification program for the responsible and trusted use of algorithmic decisioning and

[209] Cremers et al., *Trustworthy Use of Artificial Intelligence,* Fraunhofer Institute for Intelligent Analysis and Information Systems (2019):20, https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf [Accessed on January 19, 2021].

[210] Fraunhofer Institute for Intelligent Analysis and Information Systems, (November 24, 2020), *Künstliche Intelligenz sicher und vertrauenswürdig gestalten – Nächster großer Schritt Richtung KI-Zertifizierung »made in Germany«* [Press release], https://www.ki.nrw/en/certified-ai/, [Accessed on January 19, 2021].

[211] Cremers et al., *Trustworthy Use of Artificial Intelligence,* p. 15.

[212] Ibid.

[213] Fraunhofer Institute for Intelligent Analysis and Information Systems, "Künstliche Intelligenz sicher und vertrauenswürdig gestalten – Nächster großer Schritt Richtung KI-Zertifizierung»made in Germany«."

AI on 1 December 2020.[214] As of today, the working group did not yet publish a white paper on the further plans, as the group was announced a couple of weeks ago. Analysis of their work is a possible topic for the future work of this thesis.

5.7 Trustworthy: Study: Australian citizens' trust in automated decision making

In Australia, the governance of the artificial intelligence is also a topic of importance. To evaluate citizens' trust in the emerging technologies, especially the automated decision making, the Australian Human Rights Commission has conducted a study to research this issue.[215] The Australian Human Rights Commission is a national human rights institution established under the Australian Human Rights Commission Act 1986.[216] Its main duty is to advocate for the human rights of Australia's citizens, undertake inquiries, intervene in court proceedings, examine enactments, and conduct educational programs and public awareness campaigns.[217] The current section summarizes the key findings of the Commission's research.

In the study published in July 2020, the research group asked 1058 respondents (from various socio-economic backgrounds) three central questions about their awareness of automated decision making in the governmental applications and their trust in it. The respondents were categorized into the following groups: age (18-34, 35-54, 55+), education (secondary education, professional qualification, university education), employment (paid, unpaid, retired), location (capital city/ no capital city), income (low, mid, high) and others.

5.7.1 Questions and answers

The study consisted of the following questions:

1.The Australian Government, through agencies like Centrelink and the Australian Tax Office, sometimes uses artificial intelligence technology to make decisions automatically, without a human decision-maker. This is called an automated decision. Before today, were you aware that the Australian Government sometimes makes automated decisions?

[214]Roberto Zicari, "Independent certification working group launched for advancing ethical and responsible AI", Operational Database Management Systems, December 4, 2020, http://www.odbms.org/2020/12/independent-certification-working-gro/up-launched-for-advancing-ethical-and-responsible-ai/ [Accessed on January 19, 2021].

[215]The Commission provided the results of the study on request, the summary is available at: https://humanrights.gov.au/about/news/new-data-shows-australians-want-accountable-ai [Accessed on February 7, 2021].

[216]Australian Government, Attorney-General's Department, *"Australian Human Rights Commission,"* https://www.ag.gov.au/rights-and-protections/human-rights-and-anti-discrimination/australian-human-rights-commission [Accessed on January 18, 2021].

[217]Ibid.

54% of the respondents said they were aware of automated decision making in the government.

2. When the Australian Government uses a computer program to make an automated decision affecting you, like working out whether you owe money to the Australian Tax Office or Centrelink, to what extent is it important that the following 3 steps are taken?

68% of the people expressed that it is very important to them that they can object to a decision if they are not satisfied with the outcome. Almost as important as the objection, it is very important for 67% of the citizens that they obtain an explanation of why such decision has been concluded. In general, the least important was to know that the decision was made by an automated system (only 59% of the people found it very important).

3. To what extent would the following measures increase your trust in the use of artificial intelligence and automation in government decision making?

The possibility to appeal to the automated decision to a human worker would highly increase the trust in a decision for 48% of the citizens. 42% of the people claimed that it would highly increase their trust in an automated decision if all of these decisions were first approved by a human worker as well as if stronger laws and other measures were put in place to protect their human rights when automated decision making is put into operation. The least important was the option to have measures in place to prevent the government from both internally and externally share the sensitive data (only 41% of the citizens responded this measure would highly improve their trust in automated decision).

5.7.2 Findings

From this data, several implications about the citizen's requirements can be made. The citizens require:

- explanation to possibly unlawful or unfair automated decisions
- right to object
- transparency about what entity makes these decisions
- human oversight, mostly in the cases when they would like to express their disagreement with the result
- that the government takes necessary steps to protect their rights and privacy

Looking at the distribution of the respondent's demographic characteristics in the poll of the answers, there are several correlations that we point out (all of the later named demographic characteristics are not necessarily in conjunction with others):

- Men; citizens with a university degree; citizens living in the capital city; in paid employment or retired; or with a mid to high income were more likely to claim that they are aware of the automated decision making than respondents with complementary demographic characteristics (women, no university degree, low income, ...).
- People older than 55 years; people with no university degree; retired; people with low income and people not living in the capital city claimed that it is very important to them to know that the decision was made by an automated system.
- Women; people over 35+ years old; with no university degree; retired or in unpaid employment; people with low to mid-income; and people living out of the capital city claim it is very important to know a reason why a certain decision was made
- Retired people; people older than 55+; and people with low income claimed it was very important to have the opportunity to object to a potentially unlawful decision made by an automatic system
- Women; people with a university degree; people with mid-income stated that they would have much more trust in such decision if it was first checked by a human worker. Interestingly, for men, it would not make much difference.
- Women; older people; and people with at least a professional qualification claim that appealing to human decision-makers would increase their trust by much if the decision looked to them unfair.

Based on the above-interpreted data, this section concludes that trust in technology is not a constant that every person perceives similarly. Trust in technology and automated decisions correlates with the socio-economic characteristics of a person. People with unstable economic backgrounds and with possibly lower education degrees are more likely to have trust issues in technology and feel the need to obtain explanation and to object to a potentially unfair decision that more educated and economically more stable counterparts. However, this claim is solely based on the interpretation of the data provided by the Australian Human Rights Commission. The author outlines that a certain bias towards women and people with socially unstable economic background could theoretically be found in this data corpus.

Note: If a machine learning model would be trained on this data to predict whether a person has a low or high income, it would most probably predict low income for a person that answered that it is crucial for them to get the reasoning of a prediction.

5.8 Accountable: Liability and legal personhood for intelligent systems

The topic of accountability and legal personhood of the intelligent systems is a controversial topic and a subject of both philosophical research and political debate. This section

summarizes the current views on both topics of society.

5.8.1 Liability in the context of the emerging technologies

In the context of product liability for conventional products, it is quite easy to identify the person who is held liable for any harm. The products have to comply with specific requirements, depending on their nature, and in some cases, ex-ante certification before their launch on market is necessary. If a consumer operates the product in an expected manner compliant to the instructions of the producer but still gets harmed, the producer is held liable for this harm, although no intention to harm was present. This concept of liability is further described in the Chapter 3.

Due to the self-modifying and opaque nature of artificial intelligence, there are several issues with governing the liability for the products powered by AI. The problems with defining liability are:[218]

- Defining AI system and its significance: To make the legislation effective, the subject which is aimed to be regulated with such legislation must be defined very clearly. In the case of the emerging technologies, it is a huge challenge to provide a concrete definition, as any universal definition would be over- or under-inclusive.[219] If AI would be defined in general terms as an autonomous system, able to learn over time and also to self-modify, both a smart toothbrush and an autonomous vehicle would fall into the scope of such regulation, however with diametrically different significance.[220]

- Classifying between low-risk and high-risk: To address the above-mentioned challenge, such technologies could be grouped and listed in two categories, low-risk and high-risk applications. Risk is known as a product of the probability of an occurrence of an event and the damage caused by the event.[221] Bertolini argues that to date, there is not enough valid statistic data and methodologies that could help to formulate criteria for high- and low-risk categories.[222]

- Victim compensation: From the point of the economic prosperity, it is also the Consumer Sales and Guarantees Directive 1999/44/EC (CSGD), among the other directives, that provides security to the consumers to buy (technology-based) products, because they know that there are high ex-ante standards required for the product to fulfill, as well as ex-post consequences, such as 2-year warranty or

[218] Andrea Bertolini, *Artificial Intelligence and Civil Liability* (Brussels, Belgium: European Union, 2020), p. 87-93.
[219] Ibid. p. 88.
[220] Ibid. p. 88.
[221] American Chemical Society, "Risk Rating & Assessment," ASC Chemistry for Life, https://www.acs.org/content/acs/en/chemical-safety/hazard-assessment/fundamentals/risk-assessment.html [Accessed on January 23, 2021].
[222] Bertolini, *Artificial Intelligence and Civil Liability,* p. 89.

compensation, if any damage (not caused by the consumer) occurs.[223] To provide compensation, the court must identify the responsible entity for the damage, which is extremely difficult when it comes to the emerging technologies. This is the case because of the high opacity and complexity of these technologies, it is also different to identify one single point of responsibility for the litigation, and approaches to prove the evidence, due to the system's complexity (also results from learning and self-modification).[224]

- Need for narrow-tailored definition of the responsible party: To define the single entry point of litigation, there are two possible approaches. The first one, the umbrella term, (that encompasses entities, such as producer, owner, service provider, etc. and holds them all liable at the same time, and aiming to regulate technology unitarily), would require every stakeholder to insure against the same harm, that would lead to costs and efforts beyond reasonable and useful extent.[225] On the other hand, technology-specific approach might be more efficient, as in different scenarios, different entities (operators of drones, users of industrial robots, deployers of AI-based services) can best identify, control, and manage risk.[226]

- Compensable damages: Once the single entry point for litigation is identified (the entity that can best identify possible risk and ex-ante manage it), the damage caps should be defined in correspondence to the risk possessed by the given technology. For this, defining a damage caps across the categories of the systems powered by emerging technologies, would be insufficient and therefore it is necessary to consider each application separately, considering the nature of the harm, the rights it violates, and the number of victims.[227] A smart toothbrush could damage the consumer's teeth, but an autonomous vehicle, whose brakes do not function, could result in a disaster.

5.8.2 Eelectronic personhood as a legal concept

To date, the existing legal concepts encompass natural and juridical personhood. To be a legal person, it means the entity is entitled to rights and duties, posses properties, enter into contracts, sue and be sued, etc.[228] Natural personhood is granted to any human being, without exception. Juridical personhood is a legal personhood that is granted to abstract entities, such as companies or organizations, on the request of natural persons that want to form a legally recognized group.

[223] Bertolini, *Artificial Intelligence and Civil Liability,* 90.
[224] Ibid. p. 91.
[225] Ibid. p. 92.
[226] Ibid. p. 93.
[227] Ibid. p. 94.
[228] Bryant Smith, "Legal Personality," *Yale Law Yournal* 37, no. 3, (1928) https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=3259&context=ylj [Accessed on January 23, 2021].

In the recent years, the discussion has emerged whether or not to give AI-based system some kind of legal status, such as electronic personhood. In one of its statements, the European Parliament used the term *electronic personhood*.[229]

Attributing legal personhood to artificial intelligence could provide some protection to the victims, especially if the dangerous behavior of the AI agents cannot be foreseen by those who operate them, especially if the agent is autonomous and displays emergent behavior.[230] Such high-risk applications of AI could be prohibited to be put into operation unless they are certified and registered as legal persons with insurance or sufficient funds to compensate victims.[231] If the lawmakers continue to consider AI as a tool instead of a legal entity, a possible solution to liability governance would be to relax the condition of neglect and intent, and thereby move to the strict liability for high-risk autonomous systems; or to deny the validity of the tool's actions, which could stop the innovation.[232] On the contrary, if an AI agent would be registered as a legal person, it would be its principal who would be held liable, which in practical terms is no different to the basic strict liability as in case of relaxing the condition of neglect and intent in the operation of AI as a tool. [233]

From the other perspective, Bertolini interprets the *electronic personhood* in two ways: as an acknowledgment of individual rights and duties of the agent, or as an equivalent of legal personhood (possibly the juridical personhood).[234]

For the first interpretation, there is no reason to justify such step as acknowledgment of rights and duties, as any AI-based system is just a product of human intellect and does not possess such strong autonomy, that its own decision making could overrule any of the commands programmed by humans that which would allow them to pursue any goal in its way.[235] There is always at least one person who can be held accountable for the result of the system's behavior, as the person is in the best position to identify risks and put measures in place to mitigate them (either by technical means, organizational means or to decide not deploying the system at all).[236]

The second interpretation is more plausible to be interpreted in the future, as it touches on the functional perspective by putting the AI-based system into equality to the juridical persons.[237] This, however, is also only possible if precisely defined criteria are in place.

Bertolini argues that in the future, there may be many cases in which a legal personhood of AI-based systems would make sense, either by extending the scope of the juridical

[229] Bertolini, *Artificial Intelligence and Civil Liability,* p. 35.
[230] Mireille Hildebrandt, "Legal Personhood for AI?" *Law for Computer Scientists* (Online: Oxford University Press, 2019):11, https://lawforcomputerscientists.pubpub.org/pub/4swyxhx5 [Accessed on February 7, 2021].
[231] Ibid. p. 12.
[232] Ibid.
[233] Ibid.
[234] Bertolini, *Artificial Intelligence and Civil Liability,* p. 35.
[235] Ibid. p. 36.
[236] Ibid. p. 37.
[237] Ibid. p. 38.

personhood to also cover AI, or to introduce a separate legal entity - electronic personhood.[238] The liability for AI agents is a topic that is already being addressed by some Member States.[239] To prevent the European market from the market fragmentation due to the inconsistent policies and implementation of the existing liability directives, the European Commission should set out a legal framework to govern the liability of the autonomous systems.[240]

Apart from the legal personhood, the concept of citizenship is another controversial topic that is a subject of political discussion. In October 2017, for the first time in history, a humanoid robot "Sophia" has received citizenship of a country (in this case of Saudi Arabia).[241] Many refer to this act as a political choreography (to boost the social robotics market) and to Sophia as nothing more but a sophisticated chatbot.[242] Under the consideration of the Artificial General Intelligence or Artificial Super Intelligence, this robot is far from both of them, as it only delivers speeches that are inputted to the system before public performances.[243] Obviously, Sophia is based on well-developed deep learning algorithms to understand the spoken word and to answer accordingly, which only makes it an advanced form of Artificial Narrow Intelligence. As long as a machine does not equal to human intelligence in general terms independent of a task, the attribution of the citizenship is redundant.

[238] Ibid. p. 44.

[239] Tatjana Evas, *Civil liability regime for artificial intelligence* (Brussels, Belgium: European Parliament, 2020), p. 37, https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)654178 [Accessed on February 7, 2021].

[240] Ibid. p. 45.

[241] Jaana Parviainen and Mark Coeckelbergh, "The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market," *AI & Society,* (2020):1-2, https://doi.org/10.1007/s00146-020-01104-w.

[242] Ibid.

[243] Ibid. p. 2.

CHAPTER 6

Towards Responsible AI in Europe

As discussed in the previous chapters, artificial intelligence is a powerful technology of significant economic and societal value. If applied in accordance with laws and following the principles of Responsible AI, intelligent systems can be used to tackle various problems, such as environmental crisis, healthcare, social inclusion. However, if the AI systems are not designed responsibly, AI can also violate human rights and cause problems, as discussed in Chapter 4.

If no safeguards are put into place, citizens can be left powerless in their fight to protect their rights while companies encounter legal uncertainty, when an autonomous system fails to behave in an expected and appropriate manner.[244] Having different safeguards in every Member States could pose additional effort in governing the development of AI and would contribute to the Single Market fragmentation if one product has to comply with different requirements across different countries.

It is a clear goal of the European Union to become the world leader in the research and development of artificial intelligence and will push in even more in the next decades. To review how the EU plans to achieve this complex goal and to govern AI on a large scale, this chapter summarizes the steps of the EU's institutions to prepare for the upcoming changes in society and to address the challenges brought about by AI.

6.1 Milestones of AI governance in the European Union

As the development of artificial intelligence around the world had been advancing, the European Commission concluded that, although the current legal frameworks are

[244] European Commission, *White Paper on Artificial Intelligence - A European approach to excellence and trust* (Brussels, Belgium: European Commission, 2020), p. 9.

sufficient to govern state-of-the-art AI-based systems, it is necessary to review the current frameworks in dept and, if these prove to be insufficient, to the extent the existing frameworks or introduce new ones. The first announcements of the AI governance plans date to 2017, and in course of the last 3 years, the European Commission has facilitated many initiatives to review the current frameworks and identify the insufficiencies that could pose problems in the future. As a result, new guidelines, reports, and also a white paper were published to inform the stakeholders across Europe about the EU's approach to the problem. This section summarizes the most relevant milestones (in terms of relative relevance to this work) and points out the most important findings of the European Commission in each of the published documents.

6.1.1 Background: European Council requests the Commission to set out an European approach to AI

In 2017, the European Council approved the legislative priorities for the year 2018-2019 that the European Union would focus on in the upcoming period.[245] The European Commission formulated seven priorities, that did not directly address the governance of the artificial intelligence, however there was a mention of data protection, digital rights, and high ethical standards concerning the development of artificial intelligence.[246] Based on the fact that in the Joint Declaration on the EU's legislative priorities for 2017 there was no mention of artificial intelligence or robotic systems,[247] and that there is no such document available from year 2016, this chapter concludes that the year 2018 was officially the first year when the European Commission had AI governance in their agenda as a priority.

Two months before that, on 19 October 2017, the leaders of the European Council met in Brussels to discuss the approach to Digital Europe, among other topics such as defense, migration, and external relations.[248] In this meeting, the members of the Council formulated eight points to address in the course of building digital Europe: cybersecurity; a first-rate infrastructure and communications network (5G); a future-oriented regulatory framework; digitalization in the public sector and government; combating online crime;

[245] European Council, (December 12, 2018), *Council approves the EU's legislative priorities for 2018-2019* [Press release], https://www.consilium.europa.eu/en/press/press-releases/2017/12/12/council-approves-the-eu-s-legislative-priorities-for-2018-2019/ [Accessed on December 25, 2020].

[246] European Commission, *Joint Declaration on the EU's legislative priorities for 2018-19* (2017), https://ec.europa.eu/commission/sites/beta-political/files/joint-declaration-eu-legislative-priorities-2018-19_en.pdf [Accessed on December 25, 2020].

[247] European Commission, *Joint Declaration on the EU's legislative priorities for 2017* (2016), https://ec.europa.eu/commission/publications/joint-declaration-eus-legislative-priorities-2017_en [Accessed on December 25, 2020].

[248] General Secretariat of the Council, *European Council meeting (19 October 2017) – Conclusions* (2017), https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf [Accessed on December 25, 2020].

digital skills of the citizens; R&D investment efforts; and addressing technological trends including the artificial intelligence. All of these points are relevant in the context of the AI governance. High-speed infrastructure drives the data exchange among the intelligent systems. In the course of data exchange, cybersecurity must be addressed to protect the data. The virtual world is another place where we witness online crime and this crime should be combated. AI is a tool that can be trained to detect online fraud. By digitalizing the public sector and government, both AI-driven and simple systems perform automatic case handling and free resources.[249]

Another relevant meeting took place on September 29, 2017, during the Tallinn Digital Summit, where "the European Council invited the Commission to put forward a European approach to artificial intelligence by early 2018."[250] The Commission introduced the approach on April 25, 2018, as described in the Subsection 6.1.3.

6.1.2 April 10, 2018: Member States declare the Cooperation on Artificial Intelligence

On April 10, 2018, twenty-three Member States of the European Union together with the UK and Norway signed a Declaration of Cooperation on Artificial Intelligence during the Digital Day 2018 in Brussels, Belgium.[251] Throughout the year, Greece, Romania, Cyprus, and Croatia joined the initiative and also committed themselves to cooperate.[252] The current section considers the Declaration the first international document on artificial intelligence governance in the European Union, as no other preceding documents of this kind are publicly available. The goal of the Declaration was to ensure an adequate legal and ethical framework, building on EU fundamental rights and values, including privacy and data protection, as well as principles such as accountability and transparency.

To achieve these goals, the member states agreed to:

- Provide public sector data: Work together on accessibility to the public sector data and improve the re-usability of the scientific research data that emerge from publicly funded research, as the data is a substantial factor in the AI development.
- Foster research, development and innovation in the field of AI: Allocate funds to increase the quality of the research and development in the EU and its com-

[249] Matthias Daub et al., "Digital public services: How to achieve fast transformation at scale," McKinsey & Company, July 15, 2020, https://www.mckinsey.com/industries/public-and-social-sector/our-insights/digital-public-services-how-to-achieve-fast-transformation-at-scale [Accessed on January 27, 2021].

[250] General Secretariat of the Council, *European Council meeting (19 October 2017) – Conclusions*, p. 7.

[251] *Declaration of Cooperation on Artificial Intelligence*, Digital Day 2018, (Brussels, Belgium: European Commission, 2018), https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence [Accessed on December 25, 2020].

[252] Ibid.

petitiveness, modernize domestic policies to ensure that the new opportunities brought about by AI are not suppressed by laws, yet following them and to support European AI research centers & innovation hubs.

- Mitigate risks and negative impact on environment and society: Exchange the best practices in governing AI to prevent harm and violation of the EU's values. AI is already transforming the labor market, the member states should cooperate on the measures in the education and training to prepare the citizens for the operation and use of AI systems to their benefit.

- Open discussion with other member states and the Commission: Exchange the findings and best practices related to AI and legal & ethical frameworks.

6.1.3 April 25, 2018: The Commission accepts the challenge from the Council and sets out the European AI Strategy

After the Council of Europe has challenged the Commission to work out a strategy to govern AI, the Commission announced a European Strategy on AI on April 25, 2018.[253] In the Strategy, the European Commission states that it is necessary to develop a strategy that ensures the EU's competitiveness in the global AI landscape, while no one [country, citizen – Ed.] is left behind in the digital transformation; and that the new technologies will be developed in correspondence to the EU's core values, fundamental rights and ethical principles.[254] The aim of the Communication from the European Commission is to set out the three main elements of the Strategy, which are: to boost the uptake of AI in the EU, both by public and private sector; to prepare for the socio-economic changes, such as modernization of the education systems; and to ensure an appropriate ethical and legal frameworks for AI governance.[255] For the scope of this work, the latter is the most relevant point and is going to be discussed later in this section.

The EU, in general, has very strict rules and policies in regards to the consumer protection, product safety and liability, and personal data protection of the citizens, as introduced in the Chapter Legal. The EU pioneered the Privacy and Data Protection measures, as the first authority to put document of this impact into force. The GDPR was included as a part of the European AI Strategy, especially in regards to the automated decision making, processing, and profiling.[256] The Commission also accentuated the citizens' right to relevant information about the decision-making process and the logic behinds such system.[257] The chapter later discusses the alleged "right to explanation", that has often been mentioned by the general public.

[253] European Commission, *Communication from the Commission - Artificial Intelligence for Europe COM(2018) 237 final* (Brussels, Belgium: European Commission, 2018).
[254] Ibid. p. 2.
[255] Ibid. p. 3.
[256] Ibid. p. 14.
[257] Ibid.

According to the Strategy document, it is one of the key goals of the Commission to create an environment where citizens and businesses trust the technology they interact with, because they rely on the EU to provide a predictable legal environment and effective safeguards protecting citizen's fundamental rights and freedoms.[258] Based on this quote, there is a strong link between the concept of the RAI defined in the Chapter 5 and the goals of the EU in AI governance. Furthermore, the Commission points out the importance of Explainable AI (XAI)'s research.

Not only the understandability of the intelligent system, but also the manner of its "behavior" is to be approached by the Commission. The Commission announced that it would develop Ethics AI Guidelines that address issues such as the future of work, fairness, safety, security, social inclusion, and algorithmic transparency.[259] These guidelines would undergo public discussion and the feedback from the academia, private sector and civil society would be implemented. This promise was fulfilled later that year, as discussed in Subsection 6.1.4.

The Commission also promised that it would review existing legal frameworks and if necessary, extend them to better address the challenges brought about by emerging technologies or to suggest new legislation, to ensure the respect of the Union's basic values and fundamental rights.[260] This holds especially regarding product liability and consumer protection, in the Annex to this Communication, the terms such as 'producer', 'product' and 'defect' are questioned and might be redefined to "reflect the technological and other developments in the single market and global value chains."[261]

Among other points, the Commission expressed the urgency that all Member States work together on this strategy and it was also mentioned that the Commission would be working on a coordinate plan on AI with the Member States.[262] The Cooperated Plan on AI was then later published on December 7, 2018.[263]

The Commission acknowledges that to achieve the above-mentioned goals, consultation with relevant stakeholders (including experts, businesses, consumer organizations, trade unions, etc.) in the field of AI is necessary.[264] To facilitate such consultation, the Commission would set up a multi-stakeholder platform, the European AI Alliance. The Alliance would be a space for sharing best practices, encourage private investments and activities related to the development of AI.[265] the European AI Alliance has already

[258] Ibid. p. 14.
[259] Ibid. p. 15.
[260] Ibid. p. 16.
[261] European Commission, "Commission staff working document - Liability for emerging digital technologies," *Communication from the Commission - Artificial Intelligence for Europe COM(2018) 237 final,* (Brussels, Belgium: European Commission, 2018), p. 21.
[262] European Commission, *Artificial Intelligence for Europe,* p. 3.
[263] European Commission, *Coordinated Plan on Artificial Intelligence,* (Brussels, Belgium: European Commission, 2018).
[264] European Commission, *Artificial Intelligence for Europe,* p. 17.
[265] Ibid. p. 17.

organized two Assemblies, the first one June 26, 2019[266] and the second one on October 9, 2020.[267]

6.1.4 June 2018: the Commission appoints the High-Level Expert Group on AI

The European Commission announced its plan to appoint a High-Level Expert Group on Artificial Intelligence (AI HLEG) on March 9, 2018, via their press release.[268] The Expert Group is an independent advisory body to the Commission, that consists of 52 experts on AI that come from various fields, such as academia, private sector, and civil society.[269] As stated in the press release, the Expert group's main tasks are to advise the Commission on community building to form European AI Alliance of diverse stakeholders, develop guidelines for the operationalization of the ethical AI in compliance with the Charter of Fundamental Rights of the European Union described in the section 3.1.3 and to share their expertise in the course of implementation of the EU's initiative on AI.

Based on the Expert Group's deliverables, which are going to be discussed later in this chapter, AI HLEG seeks a broad public discussion with relevant stakeholders on its documents and guidelines. Public discussion enriches the meaningfulness and efficiency of the guidelines, and it is noticeable that the AI HLEG implements the feedback from the public.

AI HLEG's deliverables are:

- **Ethics Guidelines on Trustworthy AI** - made public on April 8, 2019, after the first draft from December 18, 2018, and implementation of the feedback from public discussion that ended on February 1, 2019.[270] This document sets up the requirements for Trustworthy AI and technical and non-technical methods to realize Trustworthy AI.
  - **A definition of AI: Main capabilities and scientific disciplines** - made public on April 8, 2019, together with the previous document. The Definition

[266] European Commission, "The first European AI Alliance Assembly," EC Europa, last modified August 7, 2020, https://ec.europa.eu/digital-single-market/en/news/first-european-ai-alliance-assembly [Accessed on January 28, 2021].

[267] European Commission, "Second European AI Alliance Assembly," EC Europa, last modified December 21, 2020, https://ec.europa.eu/digital-single-market/en/news/second-european-ai-alliance-assembly [Accessed on January 28, 2021].

[268] European Commission, (March 9, 2018), *Artificial intelligence: Commission kicks off work on marrying cutting-edge technology and ethical standards* [Press release], https://ec.europa.eu/commission/presscorner/detail/en/ip_18_1381 [Accessed on December 25, 2020].

[269] European Commission, "High-Level Expert Group on Artificial Intelligence," EC Europa, last modified November 18, 2020, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence [Accessed on January 27, 2021].

[270] High-Level Expert Group on AI, *Ethics guidelines for trustworthy AI* (Brussels, Belgium: European Commission, 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [Accessed on January 2, 2021].

is not considered a deliverable on its own, it serves as a basis for the common understanding of the artificial intelligence in terms relevant to other four deliverables.[271]

- **Policy and Investment Recommendations for Trustworthy AI** - made public on June 26, 2019, without preceding public discussion.[272] The AI HLEG's recommendations focus on research and academia; private sector; public sector and society in general. The Group accentuates the need to focus on data availability and infrastructure, skills and education, governance and regulations, and funding.

- **Assessment List for Trustworthy AI** - the final version made public on July 17, 2020, succeeding the draft included in the Ethics Guidelines on Trustworthy AI and implementation of the public discussion's results conducted by the European Commission between June and December 2019.[273] For more information on ALTAI, refer to the Section 6.4.

- **Sectoral Considerations on the Policy and Investment Recommendations** - made public on July 23, 2020, building on the previous work Policy and Investment Recommendations for Trustworthy AI, however with a focus on health care, e-government, justice, and law enforcement, manufacturing, and industrial internet of things. There is no mention of public discussion in the document.[274]

These deliverables mostly touch on the research objective to evaluate AI's impact on issues such as safety, transparency, accountability, environmental & social well-being, democracy, and human (fundamental) rights.

AI HLEG is not the first expert group that the Commission has set up, additionally there are High-Level Expert Group on the Impact of the Digital Transformation on EU

[271] Ibid.

[272] High-Level Expert Group on AI, *Policy and Investment Recommendations for Trustworthy AI* (Brussels, Belgium: European Commission, 2019), https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence [Accessed on January 2, 2021].

[273] High-Level Expert Group on AI, *Assessment List for Trustworthy AI* (Brussels, Belgium: European Commission, 2020), https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [Accessed on January, 2021].

[274] High-Level Expert Group on AI, *Sectoral Considerations on the Policy and Investment Recommendations* (Brussels, Belgium: European Commission, 2020), https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai [Accessed on January 2, 2021].

Labour Markets[275] or European Group on Ethics in Science and New Technologies[276], and many others. To ensure transparency on what expert groups advise the Commission, and who are the independent members, an exhaustive list of the expert groups has been set up and is available on the online platform of the Commission.[277]

The AI HLEG closed its mandate as a steering group of the European AI Alliance in July 2020, however the Alliance works further on the goals declared in the Strategy, also by organizing the second AI Alliance Assembly in October 2020.[278]

The relevance of the European Commission's participation in this matter is significant, as its primary role is to propose new laws, enforce existing laws, managing EU policies, and allocating EU funding.[279] By appointing the independent experts to share their knowledge on artificial intelligence, the Commission has access to the latest research and relevant (technical) details to make a well-informed decision and to formulate laws that will not only regulate the development of the emerging technologies but also provide frameworks on how to develop these technologies in a compliant way regarding the rights of the citizens of the European Union.

6.1.5 European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment December 3, 2018

To address the issue of AI deployment in the judicial systems mentioned in Section 4.2.3, the European Commission for the Efficiency of Justice (CEPEJ), a judicial body of the Council of Europe, has declared a European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment to prevent violation of citizens' rights and freedoms.[280] CEPEJ has agreed on five fundamental principles of the Charter to be followed when introducing intelligent tools into any judicial system:

- Principle of respect for fundamental rights
- Principle of non-discrimination

[275] European Commission, "High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets," EC Europa, last modified September 29, 2020, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-impact-digital-transformation-eu-labour-markets [Accessed on January 27, 2021].

[276] European Commission, "European Group on Ethics in Science and New Technologies (EGE)," EC Europa, https://ec.europa.eu/info/research-and-innovation/strategy/support-policy-making/scientific-support-eu-policies/ege_en [Accessed on January 28, 2021].

[277] European Commission, "Register of Commission expert groups and other similar entities," EC Europa, https://ec.europa.eu/transparency/regexpert/ [Accessed on January 28, 2021].

[278] European Commission, "High-Level Expert Group on Artificial Intelligence."

[279] European Union, "European Commission," Europa, https://europa.eu/european-union/about-eu/institutions-bodies/european-commission_en [Accessed on January 28, 2021].

[280] European Commission for the Efficiency of Justice, *European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment* (Strasbourg, France: Council of Europe, 2018), p. 7.

- Principle of quality and security
- Principle of transparency, impartiality and fairness
- Principle "under user control"

6.1.6 November 21, 2019: Report on Liability for Artificial Intelligence and other emerging technologies

In the fall of 2019, the Expert Group on Liability and New Technologies has published a report Liability for Artificial Intelligence and other emerging digital technologies to address the new challenges that emerging technologies bring about.[281] The following subsection summarizes the key findings of the Group.

The biggest challenges in the liability are brought about by complexity, opacity, openness, autonomy, predictability, data-drivenness, and vulnerability of the AI.[282] The current legal framework are well established but fail to govern the AI due to the inability to allocate loss fairly and efficiently. This is because it is difficult to identify the person whose behavior caused the damage, who benefitted from the activity, who was in a position to control the risk, and had avoided the insurance costs the most.[283]

Additionally to the aforementioned problems, the part of the problem is also the current legal system, as it does not provide as much security and compensation to those affected by the emerging technologies as to the ones that were victims of conventional technologies' failure. It is also noted that the litigation costs for victims are inappropriately high and onerous.[284]

The Expert Group also concluded that it is not necessary to give the intelligent systems any kind of legal personhood. The reason is that to date, all the damage can be attributed to the neglect or mistake on the side of the producer or the operator, rather than the system itself.[285]

The Group states that strict liability should be applied in non-private environments and scenarios that could cause significant harm. Strict liability is also reported to be an appropriate measure to govern the liability of the person who is in control of risk management that is present in the course of emerging technologies' operation (the operator), not only to the producer.[286] The producer should be strictly liable to any defect of products powered by emerging technologies, as long as the producer is in control of software updates and hardware upgrades.[287]

[281] Expert Group on Liability and New Technologies, *Liability for Artificial Intelligence and other emerging digital technologies* (Brussels, Belgium: Justice and Consumers, European Commission, 2019).
[282] Ibid. p. 32.
[283] Ibid. p. 34.
[284] Ibid. p. 34-35.
[285] Ibid. p. 38.
[286] Ibid. p. 39.
[287] Ibid. p. 42.

The last relevant point to the scope of this work is vicarious liability of the damage caused by an operator of an autonomous system. The Group showcases this in an example of an operator of an autonomous vehicle. As the operator took the risk of activating autopilot mode of such vehicle, which eventually led to a car crash, the operator is acting on behalf of the vehicle's producer and therefore the producer is vicariously liable for the car crash. The producer is in such case responsible for the product maintenance (e.g. the software updates). This example is opposite to the case of conventional vehicles, as the operator of such vehicle has full control of the vehicle and is in control of the potential risk and carries out the maintenance, use and reparation of the vehicle.[288]

6.1.7 Februar 19, 2020: White Paper on AI: Future regulatory framework

In February 2020, the Commission published a white paper in which it announced the next steps and considerations regarding the AI policymaking. The very first consideration was the definition of AI itself and the risk-assessment. When determining a regulatory framework, all terms must be precisely defined and the scope must be defined communicated. The current work of the European Commission works with the definition "products and services that rely on AI".[289] The first draft of the regulatory framework is due in the first quarter of 2021.[290] It is probable that the governance of AI applications will differ between low- and high-risk applications. An AI application is considered a high-risk application if it fulfills one of the following criteria:

- In sector where the system is put into operation, the nature of the activities (also without an AI system employed) poses significant risks. Examples of such sectors are healthcare, transport and energy, migration, or border controls.[291]
- Additionally to the sector, the impact of such AI application also plays a role. This point acknowledges that not only sector is decisive in the course of the risk evaluation, but also to what extent a use of an AI-based system poses risk.[292]
- The applications that directly impact fundamental rights of citizens are considered high-risk, regardless of the sector (e.g. automated recruiting systems with potential bias)
- Last formulated instance of high-risk application is biometric identification and other surveillance methods.

[288] Ibid. p. 35.
[289] European Commission, *White Paper on Artificial Intelligence,* p. 16.
[290] European Commission, "Artificial Intelligence," https://ec.europa.eu/digital-single-market/en/artificial-intelligence [Accessed on January 27, 2021].
[291] European Commission, *White Paper on Artificial Intelligence,* p. 17.
[292] Ibid. p. 17.

If an application fulfills one of the aforementioned criteria, it shall comply with the requirements regarding training data, data record-keeping, information to be provided, robustness and accuracy, human oversight, specific requirements for AI applications in biometric identification, and requirements for applications directly impacting human rights. These requirements are further described in the White Paper.

The future legal framework's goal is to align the policies across Europe, as it is currently the case that some of the countries have already taken the first steps, while the others did not.[293] As an example, the German Data Ethics Commission has proposed a five-level risk-based regulation system that would classify the AI systems in categories that require from no regulation, regulation to some extent, to complete ban.[294]

The European Commission promises to ensure that the European AI ecosystem will be an "ecosystem of trust".[295] The first legislative proposal is due in the first quarter of 2021.[296]

6.2 EU's definition of Responsible AI

What this work refers to as Responsible AI, the High-Level Expert Group defines as Trustworthy AI. To evaluate how the European Commission, through the AI HLEG, plans to govern the Responsible AI, this section summarizes the key principles of Trustworthy AI of the Group's Ethics Guidelines and puts them into relation with the concept of Responsible AI defined in the Chapter 5.

The key principles of the Trustworthy AI are defined as lawfulness, ethics, and robustness in the AI HLEG's Guidelines on Trustworthy AI, whereas lawful and ethical overlap with the concept of Responsible AI defined in the Chapter 5. The Lawful AI is not directly discussed in the Group's Guidelines, as current legal frameworks apply to all the products (powered by AI) with no exception and the frameworks are legally binding. This work provides the basic overview of these legal frameworks in the Chapter 3.

The Group defines the core principles of Ethical AI The Group as: respect for human autonomy, harm prevention, fairness, and explicability. Explicability is a direct subclass of Understandable AI defined in the Chapter 5.

The stakeholders who develop systems powered by AI should comply with the requirements such as human oversight; technical robustness and safety; transparency; privacy and data governance; non-discrimination and fairness; societal and environmental well-being; and accountability. These requirements overlap with the principles of Responsible AI in a larger extent.

[293] AccessNow, *Europe's approach to artificial intelligence: How AI strategy is evolving* (Online: AccessNow, 2020), p. 7, https://www.accessnow.org/cms/assets/uploads/2020/12/Europes-approach-to-AI-How-AI-strategy-is-evolving.pdf [Accessed on January 28, 2021].

[294] Europen Commission, *White Paper on Artificial Intelligence,* p. 10.

[295] Ibid. p. 3.

[296] European Commission, "Artificial Intelligence."

The Group suggests both technical and organizational methods to ensure that the systems comply with the aforementioned requirements.

Technical methods

The requirements for Trustworthy AI should be integrated into the system's architecture, e.g. by defining sets of allowed and prohibited actions and states.[297] Approaches such as privacy-by-design are already in use, and the X-by design (ethics and rule of law by design) shall be incorporated from the very beginning of the system's development.[298] Furthermore, the explainability methods of the black-box systems should be applied to achieve the Trustworthy AI.[299] The Section 5.4 discusses the explanation methods in detail. Additionally, the system's behavior should be well tested and validated beyond the scope of the traditional testing.[300] To communicate the quality of the system in terms of the requirements for Trustworthy AI, different Quality of Service Indicators could be introduced.[301] These indicators could measure the quality of the algorithm's training and training data, functionality, performance, usability, reliability, security and maintainability.[302]

Organisational methods

The Group suggests the application of non-technical methods, such as regulations, certifications, codes of conduct, standardization, dialogue with stakeholders, etc.[303]

6.3 Explainable AI and its legal enforcement in current legislation

if artificial intelligence is employed in the context of automatic decision-making, two legally binding documents in the EU indirectly govern such application. First, it was addressed in the 1995 Data Protection Directive 95/46/EC.[304] Twenty-three years later, the General Data Protection Regulation 2016/679 (GDPR) came into force.[305]

To enlighten the previously stated claim, in the Data Protection Directive from 1995, there is Article 15(1) that orders the Member States to grant the right to the citizens not

[297] High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI,* p. 21.
[298] Ibid.
[299] Ibid.
[300] Ibid. p. 22.
[301] Ibid.
[302] Ibid.
[303] Ibid.
[304] European Union, "Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data" (1995) *Official Journal* L 281.
[305] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)" (2016) *Official Journal of the European Union* L 119.

to be subject to a decision made by a decision support system without any (meaningful) human intervention if such decision could "produce legal effects concerning him [the data subject –Ed.] or significantly affects him."[306]

In the Recital 41 of the Data Protection Directive (95/46 EC), the European Parliament and the Council of the European Union, demand that the Member States implement a regulation to enable the citizens to exercise their right of access to (personal) data that are collected and processed. As Recital 41 notes, the data subject must be in the position to verify the accuracy of the collected data and must be granted the right to information of the logic behind the automatic processing.[307] The EU accentuates that information can be provided to the extent that does not infringe the trade secret of the concerned legal entity.

Before the GDPR came into force in 2018, the first drafts of the Regulation have been examined by several experts especially in the context of *"right to explanation"* of a decision made by any decision support system.[308]

Based on Articles 13, 14, 15, and 22 of the current version of the GDPR, it is obvious that the European Commission, the European Parliament, and the European Council strive for transparency and accountability regarding automatic decision-making. Sandra Watcher et. al. note that in a previous draft of the GDPR, the (Article now known as) Article 22 contained the right to explanation: "The suitable measures to safeguard the data subject's legitimate interests referred to in paragraph 2 shall include the **right** to obtain a human assessment and **an explanation of the decision reached after such assessment**".[309] As seen in the final version of the Regulation, the part with "right to... an explanation of the decision..." has been omitted from the Article 22 but added into the Recital 71, which is not legally binding. With this change, the EU has decided not to legally enforce the obligation of the controllers to ensure that explanation of any decision based on processing must be provided.[310]

To address the potential link between *governance of Explainable AI (XAI)* and the GDPR and discuss the link, this work considers the following Articles of GDPR relevant to the alleged "right to explanation":

- **Article 12 (1)** *Transparent information, communication and modalities for the exercise of the rights of the data subject:* The paragraph requires that the controller provides the requested information to the data subject in an understandable way, so that the data subject (also a child) can understand how the data is processed.

[306] Directive 95/46/EC Art. 15(1).
[307] GDPR Recital 41.
[308] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (2017), https://doi.org/10.1093/idpl/ipx005.
[309] Ibid. p. 6.
[310] Ibid.

- **Article 13(2) point (f)** *Information to be provided where personal data are collected from the data subject:* As soon as the data subject provides their personal data, it is the controller's obligation to inform the data subject that any automatic decision-making is integrated in the process, and also explain the logic behind the eventual decision made by such decision support system, including "the significance and the envisaged consequences" of the decision.

- **Article 14(2) point (g)** *Information to be provided where personal data have not been obtained from the data subject:* same as the Article 13.

- **Article 15(1) point (h)** *Right of access by the data subject:* Similar to Article 12 and Article 13, however noting that the data subject can at any time request access to the collected data, if processed, and also have it confirmed that their data are processed employing automatic decision-making or profiling.

- **Article 22(3)** *Automated individual decision-making, including profiling:* In the third paragraph of this article, it is indirectly implied that the data subject should be provided with the explanation of the decision, based on the fact that in the course of contesting the decision from the side of the data subject (this right is guaranteed in the Article 22(3)), the controller should provide at least fundamental explanation of the output and the factors taken into account.

  – In the currently effective version of the GDPR, the explicit formulation of the "right to explanation" is only mentioned in the corresponding (and not legally enforceable) Recital 71 to the Article 22.

Sandra Wachter et. al. discuss the interpretation of an explanation and also the time when an explanation should be provided.[311] They differentiate between two possibilities: prior (ex-ante) or posterior (ex-post) to the point in time when the decision was made.

The subject of the explanation can either be system functionality and/or a specific decision. System functionality describes a general functionality of the system, requirements, conditions, literally everything that is known about the system before a decision is made and can have an impact on the decision. The explanation of a specific decision is defined in their work as weights of the input features, decision rules, and rationale behind the decision.

In this work, we refer to ex-ante explanations regarding the system's functionality as transparency (Section 5.6) and could be provided by employing Model Cards for Model Reporting. We refer to ex-post explanations of specific decisions as explanator models that approximate a black box model (see Section 5.1).

In the context of the GDPR, it appears that an explanation is required before a decision is made, implying that a data subject must obtain information about the system func-

[311] Ibid. p. 3.

tionality, as written in Article 13 and Article 14 of the GDPR.[312] In the Article 15, it is implied that the explanation can be requested at any time, including ex-post explanation, which could be regarded as specific decision, expressing the need of explanator models discussed in Chapter 5.

As of 2020, with the currently effective version of the GDPR, the right to explanation is not binding and not explicitly stated in the GDPR, although the Recital 71 mentions the *"right to explanation"*.

When regulating the explainability of the intelligent systems, it is necessary to understand and assess the limitations and possibilities of the explanation's accuracy and its quality. The Chapter 5 provides a summary of the state-of-the-art approaches.

6.4 Challenges in policy making of Responsible AI

The topic of AI governance is highly theoretical and the process of its implementation into practice is not straight-forward. The compliance to AI governance guidelines gets more complicated by unclear formulations, vague wording, and lack of mutual understanding between the regulator and the implementer of the concepts in question. Size of the implementing organization or the ecosystem (enterprise, startup, or academia) in which the organization can be a significant factor influencing the interpretation of the generic guidelines, corresponding to the expert resources in the organization.

As said previously, the EU did not yet publish any directive or regulation that would directly govern AI. However, the EU facilitated an extensive public discussion on this topic and brought together the experts in the field of AI to prepare documents and summarize findings of how an ideal legal framework should look like. It is expected that the EU will act upon these findings. As it is not yet possible to analyze new AI governance frameworks and their adoption by the organizations across Europe (as they are due in the first quarter of 2021), this section chooses to analyze the adoption of one of the EU's guidelines, that are currently non-binding - the best practices of the Assessment List for Trustworthy AI (ALTAI). The goal of this analysis is to understand what are the challenges when writing general AI governance guidelines that are meant to be implemented by various stakeholders across Europe and most importantly, that they are effective.

The main assumption in this regard is that the general guidelines are often too abstract to be properly understood by small businesses or founders of start-ups that do not have the expertise in the team. On the other side, for the big enterprises, the wide formulations require cross-functional effort to be able to understand, apply and comply with the guidelines. This phenomenon is visible in the feedback paper of a startup organization and a big enterprise, which this section puts in contrast later in the analysis.

[312] Margot E. Kaminski, "The right to explanation, explained," *Berkeley Technology Law Journal* 34 (2019):199, https://www.doi.org/10.15779/Z38TD9N83H.

6.4.1 Background

The Trustworthy AI assessment list is considered a relevant guideline for the analysis in this work, as it is the first attempt of the High-Level Expert Group on Artificial Intelligence to guide organizations across Europe in the process of design implementation and deployment of their intelligent systems with trustworthiness in the mind. For more insights into the Group's activities refer to the Section 6.1.4.

With the first draft presented to the European Commission in April 2019, the AI HLEG's early version was a basis for a pilot process, where more than 350 stakeholders from the European AI ecosystem participate to provide their expert opinions and share their experience.[313] This feedback was then incorporated and the AI HLEG presented the final version of the list in July 2020.[314]

6.4.2 Main points of the ALTAI

In this self-assessment list, the principles of trustworthy AI are presented as a set of 63 questions, that guide the creators of intelligent systems on their path to implement their product in a compliant way so that the users will not be exposed to unnecessary risks. The Self-Assessment allows organizations developing intelligent systems to self-assess their approach to seven crucial topics concerning trustworthy AI and these are:

- Human agency and oversight (impact on fundamental rights, interference with human capabilities, appropriate level of human control)
- Technical robustness and safety (resilience to attack and security, fallback plan and general safety, reliability and reproducibility)
- Privacy and data governance (respect for privacy and data protection, quality and integrity of data, data governance)
- Transparency (traceability, interpretation of the outcomes, communication of the purpose, reasoning, and other relevant characteristics of the system)
- Diversity, non-discrimination and fairness (unfair bias avoidance, accessibility and universal design, stakeholder participation)
- Societal and environmental well-being (sustainable and environmentally friendly AI, social impact, impact on society and democracy)
- Accountability (auditability, minimizing and reporting negative impact, documenting trade-offs, ability to redress)

[313] European Commission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment."
[314] Ibid.

The Self-Assessment list is available in form of an online questionnaire as well as a document published on the webpage of the European Commission.[315]

The list consists of the majority of closed-answer questions (yes/no) and a couple of open-ended answers that offer more room for reflection.

6.4.3 Evaluation of the ALTAI draft by pilot organizations

To identify the challenges of the design of such document, this chapter summarizes the feedback from 5 different organizations - Microsoft (Corporate, External Legal Affairs team in Brussels),[316] Google,[317] OpenAI (Policy team),[318] UC Berkeley Center for Human-Compatible AI,[319] and Allied for Startups.[320] These organizations represent three different environments - academia, enterprise, and startups. To gain a better understanding of the current subsection and the feedback, the author of this work suggests that the readers first read through the Trustworthy AI assessment list.

This section's research goal is to understand the relevant requirements that the AI organizations have on guidelines and frameworks of this nature. Hence, this section purposely analyzes the feedback on the self-assessment list's initial draft, instead of the final version. The author of this work emphasizes that AI HLEG has implemented most of the later-discussed improvement suggestions in the self-assessment list's final version. The author appreciates the AI HLEG's effort to prepare a guideline to help organizations to design better AI. This analysis is non-exhaustive, as the feedback from only five organizations was analyzed. In-depth analysis of all companies' responses and

[315] Ibid.

[316] Kalyan Ayloo et al., Microsoft, *Microsoft's feedback on the Trustworthy AI Assessment List 2.0* (Online: European Commission, 2019), https://ec.europa.eu/futurium/en/european-ai-alliance/microsofts-feedback-trustworthy-ai-assessment-list-20 [Accessed on December 23, 2020].

[317] Sylwia Giepmans et al., Google, *Google's feedback to the Ethics Guidelines' for Trustworthy AI Assessment List* (Online: European Commission, 2019), https://ec.europa.eu/futurium/en/european-ai-alliance/googles-feedback-ethics-guidelines-trustworthy-ai-assessment-list [Accessed on December 23, 2020].

[318] Policy Team, OpenAI, *Building a More Trustworthy AI Ecosystem: Recommendations from OpenAI* (Online: European Commission, 2019), https://ec.europa.eu/futurium/en/european-ai-alliance/building-more-trustworthy-ai-ecosystem-recommendations-openai [Accessed on December 23, 2020].

[319] Center for Human-Compatible AI UC Berkeley, *Trustworthy AI Assessment List - feedback from UC Berkeley CHAI* (Online: European Commission, 2019), https://ec.europa.eu/futurium/en/european-ai-alliance/trustworthy-ai-assessment-list-feedback-uc-berkeley-chai [Accessed on December 23, 2020].

[320] Allied for Startups, *Allied for Startups's feedback to the Trustworthy AI Assessment List* (Online: European Commission, 2019), https://ec.europa.eu/futurium/en/european-ai-alliance/allied-startupss-feedback-trustworthy-ai-assessment-list-0 [Accessed on December 23, 2020].

evaluating the implementation of the companies' suggestions in the AI HLEG's final list is beyond the scope of this work.

Positive feedback

In general, the reviewing organizations agree that the list is a helpful starting point for these companies, that did not yet take steps to implement and deploy responsible AI. In particular, some authors of the reviews stated that they would appreciate such a list as guidance, rather than a regulatory mechanism.[321] All companies show appreciation towards the European Commission for taking over such a challenging task as to operationalize trustworthy AI and to make an effort to guide the European companies on how to achieve it. The reviewers also understand the difficulty of writing a comprehensive document that should be applicable in different contexts and applications, and therefore offer to share their research and findings from their experience on a quest to achieve trustworthy AI.

General improvement suggestions

Notwithstanding the positive feedback, there were several issues in communicating the purpose of the list and which stakeholders (developers, deployers or designers) the list targets within a reviewed organization. The problem of the unclear purpose was underlined by both, Google and Microsoft. In particular, it was unclear whether the list serves as a best practice guide fostering the AI developers' reflection on the topic or as a part of a mandatory legal framework.

Additionally, the Microsoft's team would appreciate stakeholder distinction in the questions for which such distinction might be relevant. This information would help the employees to better distribute the questionnaire to the responsible experts within the company, as the reviewers stated that the list required a big cross-functional effort and that it is unlikely that a single person has expertise in all of the touched topics (such as human rights, product development or social science)[322]. This point is especially relevant for smaller companies and startups and was confirmed by Allied for Startups, too.

Additionally, the sequence of the questions and the structure of the list was not ideal. As seen in the subsection 6.4.2, the questions in the list are grouped thematically into seven units. Instead, Microsoft's team would find it more useful, if the questions were grouped according to the development phases - design, implementation, and deployment. This point was also mentioned in the Allied for Startups' feedback.

To make this self-assessment list of use for entrepreneurs, questions should be formulated in ways that lead to concrete actionable items to implement to achieve trustworthy AI.[323] This is currently not the case, the questions are mostly close-ended and do not provide many insights into the AI governance of the European companies.

Quality of the answers

[321] Google, Microsoft
[322] Microsoft, Google
[323] Allied for Startups

Abstract sentences and inconsistent wording posed a challenge for the interpretation of the questions and their goal[324]. For the clarity of use, Microsoft would welcome it if the European Commission provided resources such as links to the relevant definitions and practical examples incorporated in the list. Furthermore, the questions are considered overly general, as they also apply to the fields of technology other than artificial intelligence. According to the Microsoft's team, it would be more appropriate to focus only on AI-related questions and cover other topics in documents relevant to them. Additionally, Allied for Startups pointed out that it is of high importance for the entrepreneurs that the questionnaire is designed as specific as possible, and agrees with Microsoft's point on keeping the questions AI-related rather than related to the product design.

Google's team also noted that it was difficult to estimate the expected granularity of the answers. Especially regarding the questions on safety and resilience, Google wrote that the company had set up a department to focus on trust and security issues, however, they were not sure whether to write about concrete the department's measures that ensure compliance to the guidelines or mentioning the department "was enough". Furthermore, the questions do not enforce objective self-assessment. This fact could result in answers based on personal opinions of the employee rather than an objective reflection depicting the state of affairs in the company. This is the case with the question "Does the AI system enhance or augment human capabilities?", among other questions in the questionnaire, different people would provide different judgment.

Findings

Piloting the Ethics Guidelines for Trustworthy AI was a successful initiative, in which over 350 stakeholders took part and shared their opinions and findings. The most relevant takeaways from this analysis are:

- Communicate vision: communicate, what is the purpose of the document, how it is supposed to help and to whom is it addressed.
- Avoid abstract terms: Consistency in wording, concrete examples, and links to further materials and definitions, if a guideline tackles terms, that are not generally well-known.
- Provide action items: the guidelines should lead to concrete action items that the organizations can take to reach trustworthy AI.
- AI-centricity: a guideline of this kind should tackle exclusively artificial intelligence and its development, rather than product design or topics that apply to other technologies. The sequence of questions should be following the development cycle of the AI system, rather than thematically grouped.
- Communicate expectations: it must be clear what is the expected granularity of the answers and what is the goal of each point in the guideline.

[324] Microsoft

- Ensure objectivity: If asking questions that require reflection of the employee on a specific aspect of the internal AI governance they must contain concrete quality indicators to ensure that the answers are not based on opinions and are respondent-dependent.

Summary

To get the best out of the artificial intelligence, it should be developed with responsibility in mind throughout the whole product development life-cycle. This work has reviewed possible use cases of the artificial intelligence and introduced the existing legal frameworks that indirectly govern the development of the AI to date. These are the Charter of Fundamental Rights of the European Union, which is based on the Universal Declaration of Human Rights and has much in common with the European Convention on Human Rights. One of the human rights is the right to privacy, which is addressed by the Convention 108, Directive 97/66/EC, Convention on cybercrime and Data Protection Directive 95/46/EC, which was succeeded by the General Data and Privacy Regulation two years ago.

Society still faces social, environmental, and economic issues, which widen the gap between developed and developing countries. To address these issues, the United Nations has declared Sustainable Development Goals that tackle the aforementioned problems. AI can be a very efficient tool to help reduce the gap and achieve the Goals. However, as with any other technological advancement, AI brings about negative consequences, too. If applied irresponsibly, the technology can cause harm to its users and society as such. The negative harm can be of different severity, from minor discrimination in a search engine's results, to failures to allow criminal defendants to exercise their right to fair trial, effective remedy, or presumption of innocence.

To guide the design, development, and deployment of the artificial intelligence, the concept of Responsible AI has been introduced. Key principles of the Responsible AI development are understandability (incl. transparency, explainability, interpretability), lawfulness, ethics, trustworthiness, and accountability. Understandability is a large research field, also referred to as Explainable or Interpretable AI. The AI systems often represent a black-box to their users, with no explanation about the reasoning behind the system's decision. To provide these explanations, another machine learning model can be used to approximate the black-box's decision, such as LIME or SHAP, which also provide insights into the decision criteria and conditions. To achieve the user's trust in the technology, it must be clear who is liable for the system's failure, and the technology must comply with legal and ethical principles and appear trustworthy in terms of the accuracy of its outputs. The system must be presented transparently. To address transparency, some researchers suggest issuing datasheets or certificates to the machine learning systems

similar to other goods on the market. Studies show that the trust in technology varies across the demographics.

The governance of Responsible AI can be achieved efficiently if done on a large scale, in the form of a new regulatory framework of the European Commission. Some Member States have already taken the first steps to govern the design, development, and deployment of AI, while others did not. Such inconsistency could infringe the single market of the EU. The European Commission is aware of this inconsistency and has already taken the first steps to put together the relevant stakeholders and experts in the field to evaluate the existing legislation acts. By having these knowledgeable stakeholders at its disposal, the European Commission has the necessary resources to make well-informed decisions and extend the current frameworks or propose new high-quality regulatory frameworks on AI.

The current liability frameworks govern AI, but they are not ideal in addressing the uncertainty when brought about by AI. The European Commission acknowledges that changes might be necessary. Legal personhood for AI is considered one of the strategic steps to govern the liability. However, some experts argue that legal personhood is not yet necessary, and further discussion of experts must take place.

Setting up the High-Level Expert Group on AI, among other initiatives, resulted in the creation of guidelines on trustworthy AI that stakeholders can follow throughout the development of their smart systems. The AI HLEG defined the set of requirements for Trustworthy AI as a combination of human oversight; technical robustness and safety; privacy and data governance; transparency; fairness and non-discrimination; societal and environmental well-being; and accountability.

Creating guidelines on AI governance is not an easy task, especially when addressing a wide audience with a different understanding of this topic throughout different ecosystems. General guidelines should communicate their goal, provide equal understanding of discussed concepts to the broad audience, which might come from diverse professional backgrounds, provide concrete action items that guide the audience towards Responsible AI governance with responsibility in mind, as well as provide objective indicators to help stakeholders evaluate their current approach.

If the European Commission takes all these guidelines, feedback from the Expert Groups and AI Alliance, and requirements for Responsible AI into account when creating the new legal frameworks on AI, the European Union might not only be the world pioneer in the data and privacy protection governance, but also become the pioneer in creating an "AI ecosystem of trust and excellence," as set out in its White Paper on Artificial Intelligence.

List of Figures

List of Tables

Bibliography

Articles

Smith, Bryant. "Legal Personality". *Yale Law Journal* 37, no. 3 (1928): 238–299. Visited on 01/23/2021. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=3259&context=ylj.

Lowry, Stella, and Gordon Macpherson. "A blot on the profession". *British medical journal* 296, no. 6623 (1988): 657–658. https://doi.org/10.1038/s41467-019-14108-y.

Chalmers, David J. "The Singularity: A Philosophical Analysis". *Journal of Consciousness Studies* 17, **numbers** 9–10 (2010): 7–65.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination". *Proceedings on Privacy Enhancing Technologies* 2015 (1 2015): 92–112.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". *CoRR* abs/1412.6572 (2015). https://doi.org/abs/1412.6572.

Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". *arXiv preprint arXiv:1702.08608v2* (2017).

"State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing". *Harvard Law Review* 130, no. 5 (2017): 1530–1537.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". *International Data Privacy Law* 7, no. 2 (2017): 76–99. https://doi.org/10.1093/idpl/ipx005.

Oswald, Marion, et al. "Algorithmic risk assessment policing models: lessons from the Durham HARTmodel and 'Experimental' proportionality". *Information Communications Technology Law* 27, no. 2 (2018): 223–250. https://doi.org/10.1080/13600834.2018.1458455.

Guidotti, Riccardo, et al. "A Survey of Methods for Explaining Black Box Models". *ACM Computing Surveys (CSUR)* 51 (2019): 1–42.

Kaminski, Margot E. “The right to explanation, explained”. *Berkeley Technology Law Journal* 34 (2019): 189–218. https://doi.org/10.1093/idpl/ipx005.

Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. *Nature Machine Intelligence* 1 (2019): 206–215. https://doi.org/https://doi.org/10.1038/s42256-019-0048-x.

Parviainen, Jaana, and Mark Coeckelbergh. “The political choreography of the Sophia robot: beyond robot rights and citizenship to political performances for the social robotics market”. *AI & Society* (2020). https://doi.org/10.1007/s00146-020-01104-w.

Rai, Arun. “Explainable AI: from black box to glass box”. *Journal of the Academy of Marketing Science* 48 (2020): 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Vinuesa, Ricardo, et al. “The role of artificial intelligence in achieving the Sustainable Development Goals”. *New Communications* 11, no. 233 (2020). https://doi.org/10.1038/s41467-019-14108-y.

In Proceedings

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestri. “Why Should I Trust You? Explaining the Predictions of Any Classifier”. In *Proceedings of the 2016 Conference of the North American Chapterof the Association for Computational Linguistics: Demonstrations*, 97–101. San Diego, California, USA: ACL, 2016. https://www.doi.org/10.18653/v1/N16-3020.

Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. arXiv:1705.07874, 4768–4777. San Diego, California, USA: Curran Associates Inc., 2017. https://www.doi.org/10.18653/v1/N16-3020.

Doilovi, Filip Karlo, Mario Bri, and Nikica Hlupi. “Explainable artificial intelligence: A survey”. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215.

Milli, Smitha, et al. “Model Reconstruction from Model Explanations”. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 1–9. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287562.

Mitchell, Margaret, et al. “Model Cards for Model Reporting”. In *Proceedings of the Conference onFairness, Accountability, and Transparency (FAT* ’19)*, 220–229. Yew York, NY, USA: ACM, 2019. https://doi.org/10.1145/3287560.3287596.

Raghavan, Pradheepan, and Neamat El Gayar. “Fraud Detection using Machine Learning and Deep Learning”. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 334–339. 2019. https://www.doi.org/10.1109/ICCIKE47802.2019.9004231.

Tubella, Andrea Aler, et al. “Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour”. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5787–5793. Macao, China: International Joint Conferences on Artificial Intelligence Organization, July 2019. https://doi.org/10.24963/ijcai.2019/802.

Xu, Feiyu, et al. “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges”. In Tang et al., vol. 11839.

Slack, Dylan, et al. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods”, 180–186. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3375627.3375830.

Dictionaries and Encyclopedias

Hill, Gerald N., and Kathleen Thompson Hill. “vicarious liability”. In *The People’s Law Dictionary*. New York, NY, USA: MJF Books, 2002. Visited on 01/30/2021. https://archive.org/details/B-001-001-744/page/n427/mode/2up.

Bechtel, Chris Eliasmith William. “Symbolic versus Subsymbolic”. In *Encyclopedia of Cognitive Science*. American Cancer Society, 2006. https://doi.org/10.1002/0470018860.s00022.

Goertzel, Ben. “Artificial General Intelligence”. In *Scholarpedie 10(11):31847*. 2015. Visited on 01/11/2021. http://www.scholarpedia.org/article/Artificial%5C_General%5C_Intelligence.

Copeland, B. J. “Artificial Intelligence”. Chap. Alan Turing and the beginning of AI in *Encyclopedia Britannica*. 2020. Visited on 02/03/2021. https://www.britannica.com/technology/artificial-intelligence.

“autonomous”. In *Cambridge Dictionary*. Visited on 12/07/2020. https://dictionary.cambridge.org/dictionary/english/autonomous.

Books

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford, England, UK: Oxford University Press, 2014.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, Massachusetts, USA: MIT Press, 2016.

Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Harlow, Essex, England: Pearson Education Limited, 2016.

Rainee, Lee, and Janna Anderson. *Code-depentent: Pros and Cons of the Algorithm age*. Chap. Theme 4: Biases exist in algorithmically-organized systems. Washington, DC, USA: Pew Research Center, Internet & Technology, 2017.

Hildebrandt, Mireille. *Law for Computer Scientists*. Chap. Legal Personhood for AI? Oxford University Press, 2019. Visited on 02/07/2021. https://lawforcomputerscientists.pubpub.org/pub/4swyxhx5/release/5.

Molnar, Christoph. *Interpretable Machine Learning*. Online: Leanpub, 2020. https://christophm.github.io/interpretable-ml-book/properties.html#fn8.

Reports

Larson, Jeff, et al. *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica, 2016. Visited on 01/06/2021. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Shook, Ellyn, and Mark Knickrehm. *Reworking the Revolution*. Accenture Strategy, 2017. Visited on 01/23/2021. https://www.accenture.com/_acnmedia/PDF-69/Accenture-Reworking-the-Revolution-Jan-2018-POV.pdf.

Accenture Applied Intelligence. *Realising the economic and societal potential of responsible AI in Europe*. Accenture, 2018.

Committee of experts on internet intermediaries (MSI-NET). *Algorithms And Human Rights - Study On The Human Rights Dimensions Of Automated Data Processing Techniques And Possible Regulatory Implications (DGI (2017)12)*. Strasbourg, France: Council of Europe, 2018. Visited on 12/08/2020. https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html.

Lorenzo, Rocío, et al. *How diverse leadership teams boost innovation*. Boston Consulting Group, 2018. Visited on 01/23/2021. https://image-src.bcg.com/Images/BCG-How-Diverse-Leadership-Teams-Boost-Innovation-Jan-2018_tcm9-207935.pdf.

Matonero, Mark. *Governing Artificial Intelligence: Upholding Human Rights Dignity*. Online: Data Society, 2018. Visited on 02/03/2021. https://datasociety.net/library/governing-artificial-intelligence/.

Cremers, Armin B., et al. *Trustworthy Use of Artificial Intelligence*. Fraunhofer Institute for Intelligent Analy-sis and Information Systems, 2019. Visited on 01/19/2021. https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepape%5Cr_Thrustworthy%5C_AI.pdf.

Expert Group on Liability and New Technologies. *Liability for Artificial Intelligence and other emerging digital technologies*. Brussels, Belgium: Justice and Consumers, European Commission, 2019.

High-Level Expert Group on AI. *Assessment List for Trustworthy AI*. European Commission, 2019. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

— . *Ethics guidelines for trustworthy AI*. European Commission, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

— . *Policy and Investment Recommendations for Trustworthy AI*. European Commission, 2019. https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

Kodiyan, Akhil Alfons. *An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool*. 2019.

BCS, The Chartered Institute for IT. *The exam question: How do we make algorithms for the right thing?* BCS, 2020.

Bertolini, Andrea. *Artificial Intelligence and Civil Liability*. Brussels, Belgium: European Parliament, 2020. Visited on 01/19/2021. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/654178/EPRS%5C_STU(2020)654178%5C_EN.pdf.

Daub, Matthias, et al. *Digital public services: How to achieve fast transformation at scale*. McKinsey & Company, July 2020. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/digital-public-services-how-to-achieve-fast-transformation-at-scale.

Europe's approach to artificial intelligence: How AI strategy is evolving. AccessNow, 2020. Visited on 01/27/2021. https://www.accessnow.org/cms/assets/uploads/2020/12/Europes-approach-to-AI-How-AI-strategy-is-evolving.pdf.

High-Level Expert Group on AI. *Sectoral Considerations on the Policy and Investment Recommendations*. European Commission, 2020. https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai.

Phillips, Jonathon. *Four Principles Of Explainable Artificial Intelligence*. NISTIR 8312. National Institute of Standards and Technology, Aug. 2020. https://doi.org/10.6028/NIST.IR.8312-draft.

Sartor, Giovanni, and Francesca Lagioia. *The impact of the General Datas Protection Regulation(GDPR) on artificial intelligence*. Brussels, Belgium: European Parliament, 2020. Visited on 12/08/2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530%5C_EN.pdf.

Communication from the Commission

European Commission. *"Commission staff working document - Liability for emerging digital technologies*. Brussels, Belgium, 2018.

— . *Communication from the Commission - Artificial Intelligence for Europe COM(2018) 237 final*. Brussels, Belgium, 2018.

— . *Coordinated Plan on Artificial Intelligence*. Brussels, Belgium, 2018.

— . *White Paper on Artificial Intelligence - A European approach to excellenceand trust*. Brussels, Belgium, 2020.

Webpages

Gall, Richard. "Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI", 2018. Visited on 01/21/2021. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html.

S, S. "Difference Between Law and Ethics". Figure 5.2, 2018. Visited on 01/11/2021. https://keydifferences.com/difference-between-law-and-ethics.html.

European Commission. "High-Level Expert Group on Artificial Intelligence", Nov. 2020. Visited on 01/27/2021. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

— . "High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets", Sept. 2020. Visited on 01/27/2021. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-impact-digital-transformation-eu-labour-markets.

— . "Second European AI Alliance Assembly", Dec. 2020. Visited on 01/28/2021. https://ec.europa.eu/digital-single-market/en/news/second-european-ai-alliance-assembly.

— . "The first European AI Alliance Assembly", Aug. 2020. Visited on 01/28/2021. https://ec.europa.eu/digital-single-market/en/news/first-european-ai-alliance-assembly.

Neural Information Processing Systems Conference (NeurIPS). "Getting Started with NeurIPS2020", 2020. Visited on 01/12/2021. https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28.

ACM FAccT Conference. "ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)". Visited on 01/21/2021. https://facctconference.org/index.html.

Adams, Richard, Sally Weale, and Caelainn Barr. “A-level results: almost 40% of teacher assessments in England downgraded”. Visited on 12/20/2020. https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded.

Alexander, Linsey. “Companies providing AI tutoring in Africa”. Visited on 02/03/2021. https://borgenproject.org/tag/daptio/.

American Chemical Society. “Risk Rating Assessment”. Visited on 01/23/2021. https://www.acs.org/content/acs/en/chemical-safety/hazard-assessment/fundamentals/risk-assessment.html.

Argility. “What Is Artificial Intelligence, Machine Learning And Deep Learning?” Visited on 01/11/2021. https://www.argility.com/data-analytics-ai-ml/.

Australian Government, Attorney-General’s Department. “Australian Human Rights Commission”. Visited on 01/18/2021. https://www.ag.gov.au/rights-and-protections/human-rights-and-anti-discrimination/australian-human-rights-commission.

Australian Human Rights Commission. “Human Rights Explained: Fact Sheet 5: The International Bill Of Rights”. Visited on 12/13/2020. https://humanrights.gov.au/our-work/education/human-rights-explained-fact-sheet-5the-international-bill-rights.

Carpenter, Julia. “Google’s algorithm shows prestigious job ads to men, but not to women. Here’s why that should worry you”. Visited on 01/23/2021. https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/.

Congilytica. “The Seven Patterns Of AI”. Visited on 12/07/2020. https://www.cognilytica.com/2019/04/04/the-seven-patterns-of-ai/.

Council of Europe Office in Yerevan. “About the Council of Europe - Overivew”. Visited on 12/14/2020. https://www.coe.int/en/web/yerevan/the-coe/about-coe/overview.

Dag Hammarskjöld. “UN Membership: Founding Members”. Visited on 12/13/2020. https://research.un.org/en/unmembers/founders.

Day Translations Team. “How AI is Helping Undeveloped and Developing Countries”. Visited on 02/07/2021. https://www.daytranslations.com/blog/helping-undeveloped-countries/.

Deep AI. “Narrow AI”. Visited on 12/07/2020. https://deepai.org/machine-learning-glossary-and-terms/narrow-ai.

Department of Economic and Social Affairs Sustainable Development, United Nations. “The 17 Goals”. Visited on 11/17/2020. https://sdgs.un.org/goals.

Dickson, Ben. “What is the AI winter?” Visited on 02/03/2021. https://bdtechtalks.com/2018/11/12/artificial-intelligence-winter-history/.

Doyle, Alison. “How Long Should an Employee Stay at a Job?” Visited on 01/23/2021. https://www.thebalancecareers.com/how-long-should-an-employee-stay-at-a-job-2059796.

European Commission. “Artificial Intelligence”. Visited on 02/03/2021. https://ec.europa.eu/digital-single-market/en/artificial-intelligence.

— . “European Group on Ethics in Science and New Technologies(EGE)”. Visited on 01/28/2021. https://ec.europa.eu/info/research-and-innovation/strategy/support-policy-making/scientific-support-eu-policies/ege_en.

— . “Register of Commission expert groups and other similar entities”. Visited on 01/28/2021. https://ec.europa.eu/transparency/regexpert/.

European Parliament. “European Charter Of Fundamental Rights: Five Things You Need To Know”. Visited on 12/14/2020. https://www.europarl.europa.eu/news/en/headlines/society/20191115STO66607/european-charter-of-fundamental-rights-five-things-you-need-to-know.

European Union. “European Commission”. Visited on 01/28/2021. https://europa.eu/european-union/about-eu/institutions-bodies/european-commission_en.

— . “Regulations, Directives and other acts”. Visited on 01/30/2021. https://europa.eu/european-union/law/legal-acts_en.

— . “The History of the European Union”. Visited on 12/14/2020. https://europa.eu/european-union/about-eu/history_en.

Evgeniou, Theodoros, David R. Hardoon, and Anton Ovchinnikov. “What Happens When AIis Used to Set Grades?” Visited on 12/20/2020. https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades.

Halloran, Tim. “How Atlassian went from 10% female technical graduates to 57% in two years,” visited on 01/23/2021. https://textio.com/blog/how-atlassian-went-from-10-female-technical-graduates-to-57-in-two-years/13035166507.

Hassell, Jonathan. “Netflix captions lawsuit settlement – how the perception of why you’veimproved your accessibility is vital for ROI”. Visited on 01/31/2021. https://www.hassellinclusion.com/blog/netflix-captioning-settlement/.

Lumen Learning. “Introduction to Ethics, Chapter 3: Making Ethical Decisions, Ethics and Law”. Visited on 01/17/2021. https://courses.lumenlearning.com/atd-epcc-introethics-1/chapter/ethics-and-law/.

Maayan, Gilad. “Hyper Personalization: Customizing Services With AI”. Visited on 02/03/2021. https://www.computer.org/publications/tech-news/trends/hyper-personalization-customizing-service-with-ai.

Ministry of Education, Science, Research and Sport of the Slovak Repoublic. “Opatrenia minister-stva školstva - písomné maturity sú zrušené”. Visited on 12/20/2020. https://www.minedu.sk/opatrenia-ministerstva-skolstva-pisomne-maturity-su-zrusene/.

Nimmervoll, Lisa. “Im Corona-Jahr wird Maturanten die mündliche Prüfung erlassen”. Visited on 12/20/2020. https://www.derstandard.de/story/2000116619608/im-corona-jahr-wird-maturanten-die-muendliche-pruefung-erlassen.

Open Access Government. “Are we facing an ’AI Winter’ or is our relationship with AI evolving?” Visited on 02/03/2021. https://www.openaccessgovernment.org/relationship-with-ai/86742/.

Richardson, Hannah. “Ofqual chief Sally Collier steps down after exams chaos”. Visited on 02/03/2021. https://www.bbc.com/news/education-53909487.

Saracco, Roberto. “Computers Keep Getting Better ... Than Us”. Visited on 12/07/2020. https://cmte.ieee.org/futuredirections/2018/01/21/computers-keep-getting-better-than-us/.

The National WWII Museum New Orleans. “75Th Anniversary Of The End Of World War II”. Visited on 12/13/2020. https://www.nationalww2museum.org/war/topics/75th-anniversary-end-world-war-ii.

Townshend, Charles. “History - World Wars: The League Of Nations And The United Nations”. Visited on 12/13/2020. http://www.bbc.co.uk/history/worldwars/wwone/league_nations_01.shtml.

UNESCO. “Literacy”. Visited on 02/06/2021. http://uis.unesco.org/en/topic/literacy.

United Nations. “Universal Declaration of Human Rights”. Visited on 12/13/2020. https://www.un.org/en/universal-declaration-human-rights/index.html.

— . “What We Do”. Visited on 12/13/2020. https://www.un.org/en/sections/what-we-do/index.html.

United Nations Development Programme. “Background on the goals”. Visited on 11/17/2020. https://www.undp.org/content/undp/en/home/sustainable-development-goals/background.html.

United Nations Human Rights Office of the High Commissioner. “Committee On Economic, Social And Cultural Rights”. Visited on 12/13/2020. https://www.ohchr.org/en/hrbodies/cescr/pages/cescrindex.aspx.

— . “Human Rights Committee”. Visited on 12/13/2020. https://www.ohchr.org/en/hrbodies/ccpr/pages/ccprindex.aspx.

Waters, Dustin. “Garry Kasparov vs. Deep Blue: The historic chess match between man and machine”. Visited on 02/03/2021. https://www.washingtonpost.com/history/2020/12/05/kasparov-deep-blue-queens-gambit/.

Welch, Claude. “Universal Declaration Of Human Rights: Why does it matter?” Visited on 12/13/2020. http://www.buffalo.edu/ubnow/stories/2015/12/qa_welch_udhr.html.

Weycer, Mark. “Strict Liability vs Product Liability,” visited on 01/30/2021. https://weycerlawfirm.com/blog/product-liability-vs-strict-liability/.

Zicari, Roberto. “Independent certification working group launched for advancing ethical and responsible AI”. Visited on 01/19/2021. http://www.odbms.org/2020/12/independent-certification-working-group-launched-for-advancing-ethical-and-responsible-ai/.

Press releases

European Commission. *Artificial intelligence: Commission kicks off work onmarrying cutting-edge technology and ethical standards*, Mar. 2018. Visited on 12/25/2020. https://ec.europa.eu/commission/presscorner/detail/en/ip_18_1381.

European Council. *Council approves the EU's legislative priorities for 2018-2019*, Dec. 2018. Visited on 12/25/2020. https://www.consilium.europa.eu/en/press/press-releases/2017/12/12/council-approves-the-eu-s-legislative-priorities-for-2018-2019/.

Fraunhofer Institute for Intelligent Analysis and Information Systems. *Künstliche Intelligenz sicher und vertrauenswürdig gestalten – Nächster großer Schritt Richtung KI-Zertifizierung »made in Germany«*, Nov. 2020. Visited on 01/19/2021. https://www.ki.nrw/en/flagships-en/certified-ai/.

Feedback on ALTAI

Allied for Startups. *Allied for Startups's feedback to the Trustworthy AI Assessment List*, 2019. Visited on 12/23/2020. https://ec.europa.eu/futurium/en/european-ai-alliance/allied-startupss-feedback-trustworthy-ai-assessment-list-0.

Ayloo, Kalyan, et al. *Microsoft's feedback on the Trustworthy AI Assessment List 2.0*, 2019. Visited on 12/23/2020. https://ec.europa.eu/futurium/sites/futurium/files/hleg%5C_response%5C_20191129%5C_final.pdf.

Center for Human-Compatible AI, UC Berkeley. *Trustworthy AI Assessment List - feedback from UC Berkeley CHAI*, 2019. Visited on 12/23/2020. https://ec.europa.eu/futurium/en/european-ai-alliance/trustworthy-ai-assessment-list-feedback-uc-berkeley-chai.

Giepmans, Sylwia et al. *Google's feedback to the Ethics Guidelines' for Trustworthy AI Assessment List*, 2019. Visited on 12/23/2020. https://ec.europa.eu/futurium/en/european-ai-alliance/googles-feedback-ethics-guidelines-trustworthy-ai-assessment-list.

Policy Team OpenAI. *Building a More Trustworthy AI Ecosystem: Recommendations from OpenAI*, 2019. Visited on 12/23/2020. https://ec.europa.eu/futurium/en/european-ai-alliance/building-more-trustworthy-ai-ecosystem-recommendations-openai.

Governance

United Nations. *Charter of the United Nations* 1 UNTS XVI (Oct. 1945). Visited on 12/13/2020. https://www.refworld.org/docid/3ae6b3930.html.

United Nations General Assembly. *Universal Declaration of Human Rights* 217 A (III) (Dec. 1949). Visited on 12/14/2020. https://www.refworld.org/docid/3ae6b3712c.html.

Council of Europe. "European Convention for the Protection of Human Rights and Fundamental Freedoms". *Council of Europe Treaty Series 005* (1950).

European Union. *Official Journal of the European Union* L 281 (1995).

— . *Directive 97/66/EC of 15 December 1997 of the European Parliament and of theCouncil Concerning the Processing of Personal Data and the Protection of Privacy in the Telecommu-nications Sector* FXAL98024ENC/0001/01/00 (Dec. 1997). Visited on 12/15/2020. https://www.refworld.org/docid/3ddcc6364.html.

Council of Europe. *Convention on Cybercrime* (Nov. 2001). Visited on 12/15/2020. https://www.refworld.org/docid/47fdfb202.html.

European Commission. *Joint Declaration on the EU's legislative priorities for 2017*, 2016. Visited on 12/25/2020. https://ec.europa.eu/commission/publications/joint-declaration-eus-legislative-priorities-2017_en.

European Union. *Official Journal of the European Union* L 119 (2016).

European Commission. *Joint Declaration on the EU's legislative priorities for 2018-19*, 2017. Visited on 12/25/2020. https://ec.europa.eu/commission/sites/beta-political/files/joint-declaration-eu-legislative-priorities-2018-19_en.pdf.

General Secretariat of the Council. *European Council meeting (19 October 2017) – Conclusions*, Oct. 2017. Visited on 12/25/2020. https://www.consilium.europa.eu/media/21620/19-euco-final-conclusions-en.pdf.

Declaration of Cooperation on Artificial Intelligence. Digital Day 2018 (Brussels, Belgium), Apr. 2018. Visited on 12/25/2020. https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence.

European Commission for the Efficiency of Justice. *European ethical Charter on the use of ArtificialIntelligence in judicial systems and their environment*. Strasbourg, France, 2018.

Churchill, Winston. *Speech delivered at the University of Zurich, September 19, 1946*. Visited on 12/14/2020. https://rm.coe.int/16806981f3.

State v. Loomis, 881 N.W.2d 749 (Wis. 2016). Visited on 12/14/2020. https://www.leagle.com/decision/inwico20160713i48.