IEAI

FUJITSU

White Paper

Towards an Accountability Framework for Artificial Intelligence Systems

Ellen Hohma[1]
Auxane Boch[1]
Rainer Trauth[2]

AUGUST 2022

[1] Researcher at the Institute for Ethics in Artificial Intelligence, School of Social Sciences and Technology, Technical University of Munich.
[2] Researcher at the Chair of Automotive Technology, Technical University of Munich.

Abstract

This white paper aims to define the steps that need to be taken to achieve accountability in the context of Artificial Intelligence (AI) accelerated systems and builds on previous work by industry and international organizations. We here present the backbone for a risk-based accountability framework for AI systems. To do so, we first review the concept of accountability, the technical risks brought upon by AI and their possible implications for society and organizations. Finally, we conclude on the next step to take in our research project towards an accountability framework for AI systems.

Fundings

This project was financially supported by the TUM IEAI and Fujitsu Limited.

Contents

Introduction

Artificial Intelligence (AI) is one of the major megatrends in the technology sector. Therefore, it is the root of an important change in industry and customer products that causes major impacts on society. Classic products are being replaced by new applications whose capabilities go far beyond what was previously possible. Complex tasks that could only be performed by humans are gradually replaced by AI. Automation in-vehicle systems illustrate this ongoing change. Today, automated braking maneuvers are already performed by intelligent systems in dangerous situations without requiring human reaction. However, the technological developments seen in this sector are accompanied by a shift in responsibilities. Although responsibility frameworks for products and companies already exist, they are no longer applicable to the newly created and implemented technologies. AI and, in particular the sub-field of machine learning (ML), is characterized by decisions that cannot be presented transparently to stakeholders due to their algorithmic complexity (Arrieta, 2020). Important information often remains hidden from the user and developers. For this reason, it must be ensured that the responsibility and accountability for the technology is shared seamlessly and transparently. The far-reaching decisions that AI will make, and is already making in our everyday life, are associated with high risks for humans and society. In recent years, the demand for ethical AI has been increasing and has become thematically more important in society (IEAI, 2020), leading to the involvement of international organizations on the topic. As the number of AI products increases, the need for regulations becomes crucial. Given the growing interest in this topic, legislators and international organizations need to be able to assign ethical and legal responsibility to natural or legal persons for each stage of the AI systems life-cycle, as well as in AI systems-related legal cases. This refers not only to individual monitoring but also to the public supervision of states (UNESCO, 2021). In practical application, this requirement means that an AI system can never replace ultimate human responsibility and accountability (UNESCO, 2021).

Accountability is defined as being responsible for what you do and being able to give a satisfactory reason for it (Cambridge Dictionary, 2022). In other words, there must be a responsible legal or individual person who has a transparent and understandable explanation for the AI's decisions. Accountability, on a higher level, can be defined as the relationship between an actor and the group (e.g., society) to which the actor holds an obligation to justify their conduct (Bovens, 2007; Bovens et al., 2014). It is what allows criticism and praise regarding the performance of a stakeholder and relates to their active choice to give information regarding their behavior (Bovens et al., 2014). Using this definition, the need for explainability of the AI-powered tool implemented and discussion relating to its use and impact is quite clear. Additionally, a judgment entailing formal or informal, positive or negative consequences can be passed onto the actor’s choices and thus on the product proposed by the said actor (Ackerman, 2005; Bovens, 2007; Olson, 2018).

However, for an appropriate application of the concept of accountability in the context of AI, some important questions remain, for example, how to identify the different stakeholders and how to share the responsibilities between them (Gevaert et al., 2021). Existing legal frameworks cannot satisfy the clarification demand regarding these questions for the specific context of AI. This tension between currently existing accountability regimes and their application to AI is also reflected by the European Union (EU). The EU places itself as a world leader in the development of AI and, more specifically, ethical AI (European Council, 2020). With its proposal for a harmonized approach towards AI governance for all member states through the new AI Act (Regulation 2021/0106, 2021), it recognizes and at the same time aims to close the gap between existing regulations and AI application demands. However, targeting a standardized and highly general approach can lead to challenges for its implementation in concrete application scenarios. For example, the AI Act (Regulation 2021/0106) requires transparency but does not further specify the degree of transparency needed to fulfill this obligation. Therefore, while accountability is generally described as a desirable goal or requirement (UNESCO, 2021; HLEG AI 2019), *how* it can be achieved in a standardized but simultaneously practical way is still an open topic (Loi & Spielkamp, 2021).

Targeting such an accountability approach is the final aim of a bigger project led by the IEAI, 'Towards an Accountability Framework for AI Systems', from which this White Paper results[3]. Kicking off this development, the goal of this White Paper is to outline the first steps toward the development of an accountability framework for AI-accelerated products. Based on the definition of accountability and its gaps in the context of AI described above, we propose to focus on four key aspects of accountability:

(1) Who is accountable?

(2) For what is someone accountable and towards whom?

(3) How can the responsible entity ensure compliance with the identified duties of (2)?

(4) And how can a satisfactory explanation of the measures taken for (3), in particular those related to the AI system, be given?

To answer these questions, we are proposing a risk-based approach, as risk is one big challenge in AI and therefore brings with it the obligation to take mitigating actions. On top, linking responsibility determination to risks is an approach that already can be seen in similar studies or frameworks; the AI Act (Regulation 2021/0106), for instance, clearly implements a risk-based idea by classifying AI technologies into risk levels depending on their expected negative consequences due to sector-specific impacts or use-related concerns (i.e., use of AI in education is categorized as high-risk).

The precise conception and design of such a risk-based accountability framework will be the center of future research to be conducted during the project. This Whitepaper is intended to explain the necessary fundamental principles to framing accountability in the context of AI systems risks and outlines the underlying core backbone. Specifically, we will discuss in the following our considerations on Questions 2, the identification of duties, and 3, how to comply with them. We argue that duties can be derived from technical risks linked to the developed AI system and need to be considered for organizations and society. To pave the way for answering Question 3, we review current approaches in industry and policy towards risk management, in particular linked to AI ethics. Finally, we derive major requirements for an accountability framework based on difficulties determined in the analysis, which will serve as key pillars for our future conception.

Accountability for What: Risks of AI and their Implications

Just like its benefits, the risks of AI systems are diverse and manifold, arising from technical deficiencies and unfolding their consequences for organizations and society. Unlike what is usually seen in the literature, technical issues should not be separated from their social consequences (Dahlin, 2021). Instead, the analysis of social and organizational implications should already be included in the design of an AI-enabled product, as proposed within the "ethical by design" approach.

We suggest fulfilling this demand through a risk-based analysis targeting technical risks arising with AI technologies and their implications on two essentially affected actors: organizations and society. The organizational angle focuses mainly on the consequences faced by AI service providers, while the societal aspect refers to risks to individuals and the broader population (UNESCO, 2021). In the workshop "Accountability Requirements for AI Applications" ran in March 2022 in relation to the project[4] presented in this White Paper (IEAI, 2022) on the topic of risk assessment, participants with different backgrounds, interests and expertise highlighted the need for accountability as it relates to the implications of technical

[3] https://www.ieai.sot.tum.de/research/towards-an-accountability-framework-for-ai-systems/

[4] https://www.ieai.sot.tum.de/workshop-accountability-requirements-for-ai-applications/

risks for societies and organizations. In other words, if AI systems themselves are at the origin of the risks by being implemented and used, the implications of bad or worst-case scenarios will be felt mostly by the organizations creating them, selling them, and by the users and their ecosystems. Those layers of consequences are the ones we will build our framework on, ensuring to consider a human-centered approach to the requested accountability. In Figure 1, we present our approach to risk categorization based on the workshop outcomes and the international proposition for regulations (e.g., EU AI Act, 2021; UNESCO, 2021) or principles (e.g., HLEG, 2018), as well as specific research focusing on related topics.

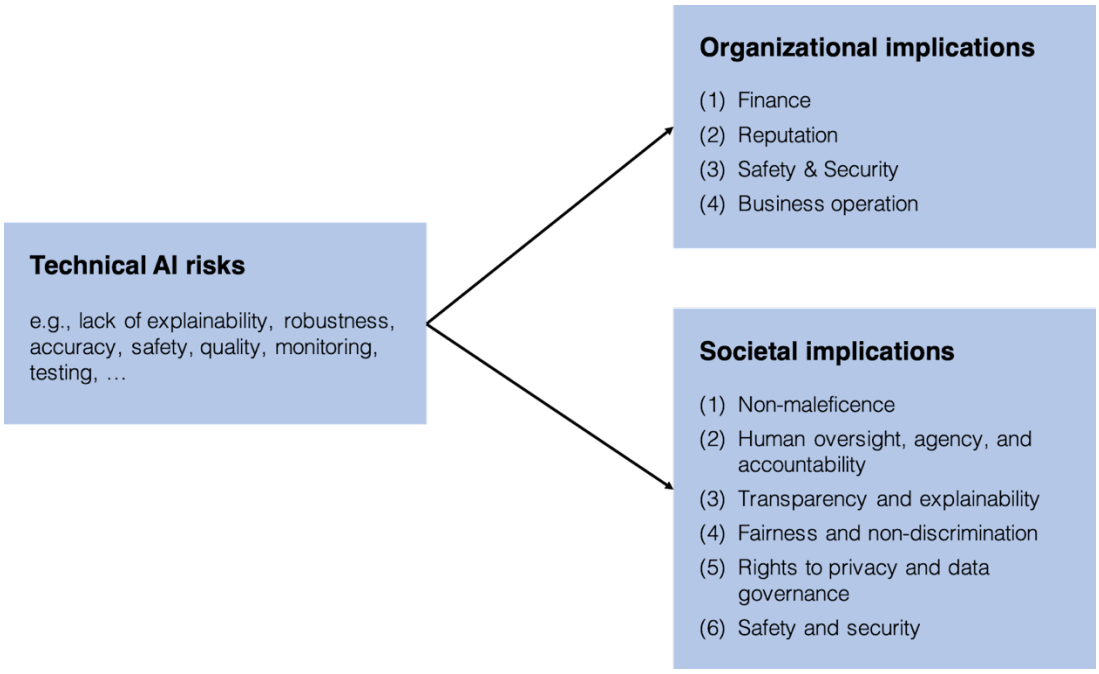

Figure 1: Technical AI risks and their implications for organizations and the society.

The risk-based approach proposed here aims to clearly define the link between all three research angles; technical, societal, and organizational. This means that AI providers should be accountable for the risks associated with their system, whether on an organizational or societal level.

Technical Risks

Technical developments in the field of AI and especially in the sub-field of ML are rapid. The emerging technology enables more powerful products. Many of these products offer opportunities for companies and society but are always associated with risks. Risks increase with the opportunities AI offers as applications have an ever-increasing impact on humans. AI technical risks are the primary reason an accountability framework for AI technologies is imperative. In the following section, we will highlight the critical technical risks of today's AI applications.

AI promises to solve complex problems in the future that traditional analytical applications cannot. A subcategory of AI is ML, which uses a huge amount of data to learn these complex tasks. One of the main problems with this technology is that stakeholders can only see the input-output relationship, not the reason for a decision (Amodei et al., 2016). In general, more complex systems with higher performance lead to opaque behavior, often described as “black-box” behavior. Without additional technologies, a trade-off between performance and explainability must be made (Došilović et al., 2018). Essentially, the technical

risks can be divided into two main problems in addition to the general design of the application. One is the performance of the algorithm, and the other is explainability.

Until a few years ago, AI application developers did not focus on transparent models, as the main goal was to make AI systems more powerful for market introduction (Arrieta, 2020). As the superiority of AI applications in some areas is beyond question, demands from legislators that require explainable algorithms become relevant (Goodman & Flaxman, 2017). Poor traceability of decision-making is a technical risk because application errors are difficult to identify: the risk of losing resources and time increases. Potential risks can arise from adversarial attacks, where weaknesses in the system can be exploited to change the decision-making process of an application (Eykholt et al., 2018). Without a traceable decision-making process, it is impossible to monitor the system's robustness during operation or analyze accidents afterward. The time required to identify errors can be significantly increased as a result. Explainable AI (XAI) techniques can help solve some of these problems.

There are essentially two technical ways to achieve a trustable and understandable model with XAI. Either the development of transparent models from scratch or the post-hoc explainability of ML models (Arrieta, 2020). In transparent models, attention is paid to the requirements of the model already in the design process. Examples of techniques in this area include linear/logistic regression, decision trees, or rule-based learning. Models from this domain are easy to understand but limited in performance. Post-hoc explainability techniques, on the other hand, examine AI models where explainability was not considered during the design process. Some of these algorithms analyze the black-box environment of the ML model to obtain information about the relationship between input and output through perturbations. A potential risk here is that careful thought was not given to what the requirements should be when designing the application. The design phase of the ML application is one of the most important steps where crucial risks can be avoided. ML can only do what it was designed to do. If important features are not properly planned in advance, the product will not have the required capabilities. For example, the testability of models needs to be considered at the design stage and how critical evaluation criteria can be defined to show the application's performance (Riel et al., 2017).

A major risk to the business is the performance of the system. If robustness and accuracy are not high enough or are below stakeholder expectations, there is a risk of product failure. The AI system must operate safely under a variety of conditions to be successful in the market. If the system is not tested sufficiently or if it is not clear to the developers in which situations the system has weaknesses, then there is a great amount of uncertainty in its operation. For this reason, it is necessary to determine in the design phase what the performance of the system should be and where its limits are. If the system is a hybrid one in which the human takes over a monitoring task, it is necessary that the system is handed over to a human in case it cannot cope with the situation. An important task here is to train the system to assess the hazardous situation on its own or to explain to the user transparently why the system acts as it does (Macrae, 2019; Sokol & Flach, 2019).

In supervised learning, data is the central element for the success of an ML model. ML applications are usually driven by a large amount of data. In order to train the system for the desired application, the first step is to generate or acquire a data set. During data acquisition, many risks can arise that can potentially harm stakeholders. First, attention must be paid to which data is captured and it must be ensured that the collected data meets the requirements of the subsequent application (Fischer et al., 2020). If data is collected from one domain in order to introduce the system to another domain, there is a high probability that the data will not meet the system requirements. Second, the data quality must be high enough. Two aspects are important, one is the distribution of the data and the other is the quality itself. If mainly data from female persons are used for human recognition, the uncertainty that male persons are recognized correctly increases (Richardson & Gilbert, 2021). Incorrect data distribution leads to discrimination between people and unfair behavior of the system. Data bias should be considered for this reason because of the risks it creates for the stakeholders (Richardson & Gilbert, 2021).

During data collection, care must also be taken to ensure that the data is anonymized and that there is no unlawful data acquisition (Csányi et al., 2021). The risk posed by ML can also be solved by ML. Therefore, technology can be both the problem and the solution (Langarizadeh et al., 2018; Kushida & al., 2012). Indeed, ML makes it possible to anonymize sensitive elements of the data in an automated way so that no sensitive data is stored in the further processing steps.

The training phase of the ML model is critical to the system's performance. To get an overview of the performance of the system, it is recommended to visualize the training process of the model. The visualization provides an insight into whether the model is overfitted or underfitted for a specific situation (Cai & al., 2016). In statistics, overfitting refers to a model that is specialized in the training data, whereas underfitting means that the system is overgeneralizing. Overfitted models can reach a very high model quality in the training data set. However, if applied to test data, significantly lower values for the model quality are the result.

An important task after the development process of ML models is the monitoring and quality management of products in operation. If system maintenance is neglected, the risk of subsequent costs increases (Poth et al., 2020). Reducing high follow-up costs can also be achieved through an extensive testing phase. In this context, it is essential that the test scenarios match the required operational design domain and that the system is sufficiently tested to avoid misconfigurations (Breck et al., 2016). If a misconfiguration occurs, safeguarding strategies can help to catch system errors. For this purpose, it is advisable to consider worst-case scenarios that could potentially ensue. Especially in security-relevant AI products, it is recommended to work with safeguarding techniques.

Another important feature of AI that should not be neglected is the high energy consumption of the technology. In the next few years, the energy consumption of data centers will increase significantly due to the use of AI systems (Cristian, 2022). The training process of AI systems, in particular, is resource-intensive. The increasingly powerful hardware components enable ever larger and more complex AI applications. The additional energy consumption must be considered financially and in terms of emissions. All the technical risks described have consequences for the management of a company and for society. These consequences are discussed in the following sections.

Organizational implications of technical AI risks

Implications for the system provider resulting from uncontrolled (technical) risks of AI can be examined along four main categories: ***financial***, ***safety***, ***operational*** and ***reputational***. As, essentially, all the above-mentioned technical risks affect all these consequence categories in one way or another, we will focus on the following on outlining some examples of the most influential risks per category and elaborate on why they particularly evoke consequences for the AI provider.

First**,** ***financial consequences*** can result from unlawfully designed or faulty AI systems. The provider's legal liability for products they put on the market is, hence, a key driver. As for any other product, AI providers can be held liable for shortcomings of their software. In this context, certain legal frameworks clearly apply to AI, too. For instance, the General Data Protection Regulation (Regulation 2016/679; GDPR) sets strict guidelines for the use of personal data, specifically, demanding lawful, fair and transparent processing, collection only for specified, explicit and legitimate purposes as well as recommendations for data accuracy and data storage (Art. 5, GDPR). High fines are in place if personal data protection is not ensured properly by the AI provider. Furthermore, requirements for the system's functionality, security, and safety can be generically derived from currently existing legal frameworks. Warranty law defines standards to what extent a user can claim compensation in case of a product defect. In addition, warranty law, as well as competition, consumer, and tort law, lay down specifications for a system's honest and fair design, leading to cease and desist or fines in case of false information on system operation or anti-competitive claims. On the other hand, some facets are less clear and particularly tricky

when applying standard liability frameworks to the specific context of AI. The concrete details of how currently established liability regimes apply or should be adapted for AI are still discussed among scholars and practitioners (e.g., Ebers, 2021; Zech, 2022). Debated open questions include, for instance, to what extent personal data that is used for training is inherent in the resulting AI system itself or how to determine an AI's quality that is customary for goods of the same type, e.g., regarding accuracy, that can define whether the system is free of defects. Additionally, it is still open to what extent risks linked to the ethical design of AI, such as a system's transparency or fairness, will cause direct financial consequences for AI providers. The new Regulation of the European Parliament and the Council laying down harmonized rules on Artificial Intelligence (Regulation 2021/0106, AI Act) clearly mirrors the EU approach demanding lawful, safe, and trustworthy AI applications respecting existing law on fundamental rights and EU values. It provides the first indications that fines are to be expected in this regard. The AI Act still needs to be finalized, and remaining questions for AI liability clarified before we can be certain about all the concrete risks that elicit certain obligations. However, it can be confidently expected that the above-outlined risks will lead to financial consequences for AI providers in whichever tangible way.

Second, (technical) risks can impact the AI provider's reputation if not managed properly, summarized as ***reputational consequences***. Reputational consequences are related to financial consequences, as an organization's bad image can cause corporate performance losses. Still, examples such as Meta or Google, which are still criticized for their data protection approaches (e.g., The New York Times, 2022) yet remain highly successful, have shown that this is not necessarily the case. An organization's reputation can be considered more than its monetary representation. Examples of technical risks that can negatively impact an organization's reputation if unconsidered include data processing, in particular linked to data privacy and protection strategies, system functionality, safety and security, as well as ethical aspects in the AI's design and development. In return, there are risks that can help improve an organization's image if prevented, such as transparency and explainability, through fostering confidence with the technology and, hence, trust and acceptance.

The third category of consequences for AI providers resulting from (technical) risks is ***safety*** *and* ***security consequences*** linked to physical or personal harm. These can follow from unlawful or faulty data processing resulting in violation of data privacy or data security issues, functionality, security and safety problems caused by the AI model, as well as a lack of transparency and post-deployment system monitoring. While safety and security issues ultimately include financial and reputational consequences, we examine them as a separate category due to their potentially serious impact. For example, financial consequences from safety risks arise due to the AI provider's liability for damages caused by a defect of their product. In the EU, the Approximation of the Laws, Regulations and Administrative Provisions of the Member States concerning Liability for Defective Products (Council Directive 85/374/EEC, The Product Liability Directive, PLD) regulates product liability and establishes a 'strict liability'-regime for a product's defects. It declares the producer, respectively the manufacturer of the defective component, liable for any damage caused by a defect product to one of the protected rights. An exhaustive list of protected rights is specified, mainly including death, personal injury, and damage or destruction of property. There is still some debate about the extent to which software, and thus AI, falls under the PLD's product definition and, therefore, whether the PLD is applicable at all. However, as the prevailing majority argues towards regarding AI as a product, liability for damages caused by deficient AI-based systems and therefore payment obligations can be expected (Cabral, 2020; Hacker & Passoth, 2021; Navas, 2020). At the same time, unmanaged safety risks can cause reputational consequences for the AI provider. While their implications are less foreseeable than regulatory obligations, past examples have shown that safety and security problems of AI systems can lead to mistrust in the system potentially causing bad reputation (e.g., crashes of Tesla self-driving cars and the ongoing debate on to what extent autonomous driving can be permitted, or frequent occurrences of data breaches leading to the everlasting demand for high data privacy and security standards). The strength of reputational damage might depend on the type and purpose of the AI within the system. For example, accuracy issues of an AI component within self-driving cars identifying pedestrians resulting in unsafe

driving behavior can arguably lead to higher user mistrust towards the system overall than accuracy problems of a natural language processor within a chatbot application. Therefore, while their concrete strength and ultimate outcome are context-specific and hardly predictable, reputational consequences linked to AI safety and security risks are surely undeniable.

A final category of consequences for the AI provider resulting from the above outlined (technical) risks relates to ***business operations***, referring to enhancing internal company processes and performance. Ultimately, impacts on the company's processes lead to financial consequences as good company performance should at best turn into profit. However, not every organizational operation can directly and exclusively be measured with monetary scales, such as employee well-being or customer satisfaction. Therefore, impacts on the company's ability to run operations smoothly are regarded as their own category. One exemplary risk that can lead to inefficient company performance is transparency. While, on the one hand, transparency is usually mentioned along with user empowerment, ultimately facilitating understanding and trust, the ability to understand an AI's underlying processes can further help AI providers themselves in the development by identifying and targeting problems early on. This can prevent more serious risks, e.g., safety or fairness, or speed up the development phase by avoiding additional feedback loops. Therefore, technical risks can lead to a decline in operation performance if not managed properly.

Societal implications of technical AI risks

When considering risks induced by the creation, implementation, and use of an AI system, it seems of utmost importance to connect the AI technical design analysis to its social implications, thus understanding the technologies' possible impact on society and users (Dahlin, 2021). Indeed, if the intended purpose of a tool is the focus point of ethical evaluations, the AI technology's impact on society might be unpredictable, due to unforeseen or incorrect use or (technical) risks (Regulation 2021/0106). Therefore, the identification of responsible and accountable actors needs to be thoroughly considered with regard to the possible adverse repercussions on society as a whole and particular individuals and groups of users. Building off UNESCO's recommendations for AI Ethics (2021), the EU High Level Expert's Group on AI's 7 key requirements for a trustworthy AI (2019), and the European Commission's proposition of the Artificial Intelligence Act (Regulation 2021/0106), we here propose an overview of the consequences that (technical) risks pose to society.

The first agreed-upon principle of ethics for all AI systems is to **"*Do No Harm*"** and foster societal and environmental well-being. Indeed, suppose the intention behind the implementation of new technology or its use allows for manipulative, exploitative, and social control. In that case, democracy, fundamental and human rights, could be in jeopardy. In the past, AI-powered bots on social media have demonstrated their power in spreading misinformation widely and frequently. If used in an electoral context to support one party or the other, they could pose a serious threat to political stability (Lapowsky, 2017). Moreover, AI systems could provoke physical and psychological harm, which might be a widespread issue if used globally. In the context of mental health support agents or counselors' bots, the risk is high due to their initial designed purpose (The Guardian, 2019). Reducing possible harm also includes considering proportionality in the use of an AI. In other words, "Do No Harm" takes into account the context of use and its actual positive influence on existing situations, with an importance of monitoring and re-adapting in case of misuse or errors. In the context of environmental actions to fight against climate change, ML algorithms are used to estimate the carbon dioxide absorption of a forest from aerial imagery-based analysis (Reiersen et al., 2021). The calculations have shown to evaluate retention of carbon dioxide as higher than it actually is, thus underestimating the amount of forest carbon dioxide credit needing to be bought by emitters to get to a net-zero emission (Reiersen et al., 2021). Overissuing those carbon dioxide credits can have a significant adverse effect on the environment (Han, 2021).

Some of the issues related to the well-being of society and the environment can be identified and managed early on with the implications of humans in the loop, on the loop, or in control, which leads us to our second set of societal implications relating to the risks brought by a lack of ***human oversight, agency, and accountability***. A lack of attribution of ethical or legal responsibility at each stage of the AI life-cycle to legal entities or physical persons can result in the absence of assigned oversight over the tool's decisions and actions. The risks for society as a whole here would be the reduction of human autonomy and possibly impeding fundamental freedoms of individuals. During the pandemic, for example, Stanford Medical Center implemented a logistic AI to decide to whom within the hospital staff should receive the first 5000 doses of vaccine. The AI proposed only seven frontline resident workers within the 1300 first individuals chosen for priority vaccination. Even though made aware early on of the issue, the hospital leadership did not change their method immediately and blamed the algorithm's complexity. Later, Stanford apologized, and each team management was given the authority over vaccination priority within their own team (Guo & Hao, 2021). Humans should be part of the final decision and monitor the AI's process while knowing who is to be held responsible in case of mishap, especially in contexts where the decision relates to highly important matters for one's life such as health, education, or societal and environmental well-being in general.

To reach a clear view of who is responsible for what throughout the AI life-cycle and relate to all other ethical risks presented in these sections, the need for ***transparency and explainability,*** whether to technical or non-technical actors, is paramount. These two principles lie with the requirements for understanding how and why the decision was made by the system (Ribera & Lapedriza, 2019; Miller, 2019; Hoffman et al., 2018; The Alan Turing Institute, 2019), reducing the risk of induced acceptance, or acceptance through ignorance, threatening practical accountability with respect to technology (Loi & Spielkamp, 2021). While in some cases, decisions will have a low impact on society and users, such as price forecasting for raw materials within the industry (*Chai Price Forecasting,* 2022), they might lead to major changes for populations and sectors. For example, trading algorithms are now largely involved in shaping markets and trades worldwide (Dahlin, 2021). The lack of traceability, communication, and education regarding the inside process of AI can have strong consequences on its acceptance within society, proper use, and success.

Some principles of AI ethics, supporting the identification of risk areas for the technology, might relate to more precise groups and individuals. The questions of ***fairness and non-discrimination***, for example, highlights possible risks for specific groups of diverse populations. A practical example of a biased AI-based decision's high impact on lives took place in 2020 in the United Kingdom: An algorithm was put into place to support grading for A-Levels (exams taken prior to entering university, with a high weight in the selection process to enter higher or lower ranked universities), building off school's historical performance, and the ranking of students within their current school. Fee-paying private school students scored better, unfairly scaling the results for lower socio-economic background students at a furthest lower rate. The algorithmically determined results were canceled once the flaws were pointed out by schools, and teachers gave grades based on the students' work throughout the year (Porter, 2020). When considering those issues, context is to be taken into account. In 2021, AI research and development was led mainly by institutions present in the United States, the European Union, and China (Zhang et al., 2021). Additionally, in the same year, women represented only 32% of the roles in AI and data (World Economic Forum, 2021). This unequal repartition entails a responsibility for the leading regions to ensure a universally fair and non-discriminatory design, training, and education surrounding the AI-products proposed with regard to the populations they will be used by, at the risk of creating inequitable access, discrimination towards, and exclusion of specific people (Barocas & Selbst, 2016).

Setting standards for ethical practices within its region, the EU implemented in 2018 the General Data Protection Regulation (Regulation 2016/679), paving the way to high consideration of ***rights to privacy and data governance***, supporting users' awareness and agency regarding personal information control. The

privacy principle, when endangered, can have high repercussions on an individual, a group, or society. In September 2021, researchers discovered a spyware called Pegasus which had infected multiple Apple devices, allowing the perpetrator to access and record messages and calls as well as turn the tools' camera and microphone on without the user's knowledge. The company at the origin of this software initially sold its product to governments, presumably to scrutinize terrorists and criminals. It has been seen that the spyware was also used to collect information on activists, journalists, and politicians (Perlroth, 2021). Moreover, within the scope of an AI system, the recording of human behaviors of individuals and groups might include sensitive information such as political views, names, and sexuality. Thus, more than the need to train the AI system with representative data to reduce discrimination, the protection and confidentiality of such knowledge needs to be ensured.

Finally, unwanted and unexpected harm might occur while using an AI system, whether coming from a voluntary external attack on the system or due to internal accuracy issues and thus mis-judgment from the tool itself. Consequences of such actions can go from a smart thermostat within a home derailing and burning down the house, to a robot making a deadly mistake in surgery (Schütte et al., 2021). When related to ***safety and security*** implications, such mishaps might impact the physical or mental integrity of an individual, or a group of users, highlighting the need for technically robust and secure AIs.

To sum up, the technical risks of an AI system can and do impact both organizations and societies. The considerations presented here give insights into risk categorization and assessment. While aiming at building an accountability framework, risk management needs to be considered to define responsibility sharing between stakeholders. In the next chapter, we present and discuss existing risk management approaches and their limitations brought up by governments, international organizations, and the industry sector.

How to define Accountability: A Risk Management Approach

Accountability is to be defined with regard to the consequences of risks induced by AI systems. In the previous section, we illustrated the different types of technical risks and their possible implications for companies, societies, and users. To reach an agreeable responsibility and liability framework for all AI technologies, it is important to build on existing work allocating tasks to reduce and manage said risks, as they will allow for a layout of stakeholders involved in their prevention and resolution. We here review existing risk management frameworks for AI systems at the international and business level.

International risk management concepts

Governments and international organizations are attempting to create risk management frameworks for AI systems in order to define and clarify best practices.

In 2019, the European Union's High Level Experts Group (EU HLEG) on AI put forward a basis for AI ethics in Europe in presenting their seven key requirements for a trustworthy AI. In the paper, the authors highlight the current need for appropriate tools to mitigate risks proportionally to their magnitude. In other words, current risk management tools are inadequate for all types of AI systems, and do not take into account the importance of a risk, its probability of occurring, or its prevalence as a factor, thus impeding a customized solution for each AI technology independently. The document proposed by the EU HLEG (2019) does not propose a clear solution to the issue, but rather gives a list of questions to be asked by AI-developers and providers throughout the life-cycle of the AI concerning ethical risks, proposing a draft for ethical risk assessment. Nevertheless, this work is at the origin of multiple initiatives that are building risk management frameworks on the international scene.

Building off the prior mentioned work, the OECD (2022) proposed its framework for the classification of AI systems. This framework can be used for risk management, informing relevant work on mitigation, compliance, and enforcement throughout the AI system lifespan, while including a corporate governance angle. Differentiating the "in the lab" and "in the field" contexts for AIs, the management approach highlights the need for monitoring the use and evolution of a system in addition to the conception and development phase. When used by different AI providers and developers, the analysis of their use of the tool allowed them to identify its limitations: the framework is more applicable to specific AI systems than generic ones, and technical characteristics seemed harder to evaluate than social impact, making the tool less accurate than desirable. In addition to this framework, the OECD (2021) proposed a list of tools for trustworthy AI categorized by technical, procedural, and educational topics and usable for risk management of AI systems, paving the way towards a more applicable risk management approach for AI systems on the global level.

In another example, the European Commission suggested four degrees of risk as part of the AI Act (Regulation 2021/0106) proposal. Each level has different requirements to be met in the risk management phase, without giving a clear line of conduct or step-by-step tool to achieve said requirements. It seems important to highlight that the risk-assessment and risk-mitigation framework for the AI-systems approach is considered in individual countries as an acceptable and applicable approach to the problem while not being fully implemented or defined yet (e.g., NIST, 2022; German Data Ethics Commission, 2019).

In general, standard criteria can be found throughout the different approaches of governments and civil society organizations for defining the risk level of an AI application or system, regardless of the number of risk levels or whose organization provides them (OECD, 2022):

- Scale, i.e., the severity of the harmful implications (and probability).
- Scope, i.e., the range of application, such as the number of people who are or will be affected.
- Optionality, i.e., the degree of choice in whether to be exposed to the impacts of an AI system.

However, at this time, no risk management framework applicable to all types of AI systems has yet been presented internationally by governments or organizations.

Risk management concepts proposed by academia and the industry

The private industry has also recognized the validity of risk-based approaches and the need to react to both technical and ethical AI risks. For instance, many companies, such as BMW[5] or Novartis[6], have started to define principles or create codes of conduct for the ethical and responsible use of AI. While this is already a major step towards AI accountability, it predominantly targets clarifying challenges and determining risks and responsibilities. In particular, a second main dimension of accountability, the proving of meeting and complying with responsibility obligations, is currently rarely met.

In order to account for identified risks, technical and methodological tools are increasingly developed by academia and industry following two major strategic approaches: risk prevention and risk detection and mitigation. While risk prevention proactively targets an improvement of software quality, reactive approaches like risk detection and mitigation aim at assessing and reporting risks, as well as pre-defining measures to manage and compensate for potential harm (Clarke, 2019).

The first stream of risk management strategies, risk prevention, is highly targeted through technical tools to adapt software quality and system capabilities. In particular for societal challenges, practically addressing

[5] https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/downloads/ENG_PR_CodeOf EthicsForAI_Short.pdf

[6] https://www.novartis.com/about/strategy/data-and-digital/artificial-intelligence/our-commitment-ethical-and-responsible-use-ai

specific AI ethics principles, among them often fairness and transparency (Ayling & Chapman, 2021), has become a research topic of increased interest. To tackle fairness problems, toolkits, such as IBM's AI Fairness 360 suite[7], have been developed aiming at exhibiting or even removing biases in the used datasets and AI models. Techniques leveraging transparency highly complement such fairness improving methods, as sources of bias can be more easily spotted. Coding libraries, like LIME[8] or SHAP[9], have been proposed to actively increase the system's transparency by design, helping reveal bias along with other hidden risks. This can further serve as a step towards real accountability, as transparency is often a prerequisite for demonstrating that responsibilities have been respected and appropriately addressed.

While risk prevention can help avoid some AI-related risks, not all of them can be controlled early on or during development. Therefore, reactive risk management strategies relying on ongoing monitoring and mitigation are needed to ensure long-term risk governance.

Using standardized procedures to quantify and handle risks is common in many fields; therefore, multiple guidelines and standards have been developed. For example, ISO 31000 outlines a classic risk management process, using risk identification, analysis, evaluation and treatment and aligned with ongoing monitoring, communication, recording, and reporting. These fundamental steps have been transferred to more context-dependent risk management strategies. The ISO/IEC/IEEE 16085:2021 standard, for instance, documents a risk management process model for system and software engineering, including risk management planning, risk analysis, and monitoring, risk treatment, and management process evaluation. More closely adapted to the context of AI, ISO/IEC DIS 23894 lays down management approaches regarding AI risks, involving risk assessment (risk identification, analysis, evaluation), risk treatment (selection, preparation, and implementation), monitoring and reviewing and finally recording and reporting.

Simultaneously, academia has adopted these elementary steps and suggested more practical methodologies for their implementation. Clarke (2019), for instance, examines principles and business processes for responsible AI. He calls for greater inclusion of affected parties, proposing a multi-stakeholder risk assessment and management process that follows a standardized approach along the lines of risk analysis, risk mitigation design and risk treatment and further includes a more detailed assessment of the organization and stakeholders to protect own interests as well as account for those of related actors and the broader society. Felländer et al. (2021) similarly target shortcomings of existing risk management procedures, in particular, the challenge for organizations to practically apply ethical guidelines while missing dedicated and workable tools. They propose the methodology 'Data-driven Risk Assessment for Ethical AI' (DRESS-eAI) which incorporates (1) problem definition and use case scoping, (2) risk scanning and profiling, (3) risk assessment, (4) identification of risk mitigation measures, (5) stakeholder engagement and (6) reviewing and maintaining the risk management process.

While research and academia theoretically determined risk management strategies, their structured, systematic application is still lacking in practical and comprehensive realization. Examples of how the industry starts to discuss risk governance approaches from a practitioner's perspective have been gathered by Ezeani et al. (2021). Several partnerships and initiatives have been launched that recognize and particularly target AI risks, such as the Partnership on AI, initially founded by AI researchers from Apple, Amazon, DeepMind and Google, Meta, IBM and Microsoft, with now more than 100 member organizations, or The Software Alliance (BSA), a trade group representing commercial software developers founded by Microsoft in the 1980s (Ezeani et al., 2021).

However, these examples also show that a uniform risk management procedure has not yet been established in practice. This in turn, poses a problem for standardizing accountability measures, as appropriate risk

[7] https://aif360.mybluemix.net
[8] https://github.com/marcotcr/lime
[9] https://shap.readthedocs.io/en/latest/index.html

management is a prerequisite for defining responsibility obligations and demonstrating compliance with them in a transparent manner, i.e., accountability.

At this time, public and private sectors are developing and trying to bridge the gap between ethical requirements and practical applications of risk management for AI systems, as the interest grows, and regulations are being developed. In our work, we propose to support and participate in this evolution towards applicable ethical risk mitigation for AI systems by defining actors to be held accountable and proposing strategies to be implemented at each step of the AI life-cycle.

Conclusion & Outlook

The takeover of many tasks in our society by AI requires new regulations, as the actions are more far-reaching than for conventional products and processes. Many scholars, therefore, call for a transition of research regarding responsible AI from principal definition to proposition of practical methodologies and frameworks (e.g., Schiff et al., 2020; Morley et al., 2021). While AI ethics research has long focused on the 'what', it is increasingly demanded to now continue by targeting the 'how' (Morley et al., 2021).

In this paper, we defined the most pressing questions to manage accountability for AI-enabled products and suggest addressing them using a risk-based approach. Therefore, we studied the risks posed by AI-accelerated technologies and examined their possible implications from two perspectives: societal and organizational. The investigated angles can help develop a management-oriented framework that companies can apply throughout their product development process. Further, those angles reveal the need for a new way to manage and mitigate negative effects, as standard legal or ethical frameworks cannot be used for AI-enabled applications without further adaptations.

We suggest building such an accountability framework along two major requirements. A practical tool is needed to encourage organizations in charge of transitioning to the more responsible use of AI to easily manage risks and obligations. Thus, an important pillar of our suggested accountability framework is the ease of use and practicality, e.g., ensured through increased stakeholder engagement and expert consultation. A second important pillar is a general applicability in different contexts. The lack of a standardized and coherent approach to risk identification, reporting, and associated accountability hinders universal mitigation. We seek a holistic and generalizable approach to promote the harmonization of currently proposed methodologies. These requirements can enable the design of an accountability framework that leverages risks and their consequences to create a risk-management-based approach and, simultaneously, place applicability and feasibility at the center. Implementing these key requirements as well as the concrete conception of our suggested framework approach will be the purpose of our future project activities. Additionally, extending our current work, we will in the future hold more experts' workshops to learn directly from practitioners their needs and understanding of the current AI industry situation in the EU context.

References

Ackerman, J. M. (2005). *Human rights and social accountability*. Participation and Civic Engagement, Social Development Department, Environmentally and Socially Sustainable Development Network, World Bank.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, 1-25.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104,*671.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, *13*(4), 447-468.

Bovens, M., Goodin, R., & Schillemans, T. (2014). Public Accountability. In *Oxford handbook of public accountability*. Oxford University Press.

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017, December). The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 1123-1132). IEEE.

Cabral, T. S. (2020). Liability and artificial intelligence in the EU: Assessing the adequacy of the current Product Liability Directive. *Maastricht Journal of European and Comparative Law, 27*(5), 615-635.

Cai, S., Breck, E., Nielsen, E., Salib, M., & Sculley, D. (2016*). Tensorflow debugger: Debugging dataflow graphs for machine learning*.

Cambridge Dictionary. (2022, May 11). *Accountability*. https://dictionary.cambridge.org/dictionary/english/accountability?q=Accountability

ChAI Price Forecasting. (2022, March 11). ChAI. https://chaipredict.com/

Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law & Security Review, 35*(4), 410-422.

Council Directive 85/374/EEC. *The approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31985L03 74&from=EN

Cristian, D. (2022, April 21). *STUDY: Data centers energy consumption will double in the coming years*. Business Review. https://business-review.eu/energy/study-data-centers-energy-consumption-will-double-in-the-coming-years-230202

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. *Symmetry*, *13*(8), 1490.

Dahlin, E. (2021). Mind the gap! On the future of AI research. *Humanities and Social Sciences Communications*, *8*(1), 1-4.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.

Ebers, M. (2021). Liability for artificial intelligence and EU consumer law. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, *12*, 204.

European Council. (2020, August). *Special meeting of the European Council – Conclusions* (EUCO 13/20). https://www.consilium.europa.eu/media/45910/021020-euco-final-conclusions.pdf

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition (pp. 1625-1634).
Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). *A survey of artificial intelligence risk assessment methodologies: The global state of play and leading practices identified*. Trilateral Research & EY. https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf
Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2021). Achieving a Data-driven Risk Assessment Methodology for Ethical AI. *arXiv preprint arXiv:2112.01282*.
Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., ... & Moser, B. (2020). Ai system engineering—key challenges and lessons learned. *Machine Learning and Knowledge Extraction*, *3*(1), 56-83.
German Data Ethics Commission. (2019). *Opinion of the Data Ethics Commission*. Data Ethics Commission of the Federal Government. https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3
Gevaert, C. M., Carman, M., Rosman, B., Georgiadou, Y., & Soden, R. (2021). Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns*, *2*(11), 100363.
Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.
Guo, E., & Hao, K. (2021, January 21). *This is the Stanford vaccine algorithm that left out frontline doctors*. MIT Technology Review. https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/amp/
Hacker, P. & Passoth, J., (2022). Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond. In: *Holzinger, Goebel, Fong, Moon, Müller and Samek (eds.), Lecture Notes on Artificial Intelligence 13200*: xxAI - beyond explainable AI, Springer. http://dx.doi.org/10.2139/ssrn.3911324
Han, J. (2021, May 12). *The Climate Solution Actually Adding Millions of Tons of CO2 Into the Atmosphere*. ProPublica. https://www.propublica.org/article/the-climate-solution-actually-adding-millions-of-tons-of-co2-into-the-atmosphere
High-Level Expert Group on Artificial Intelligence (HLEG AI). (2019, April). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html
Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., & Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, *50*(Suppl), S82.
Langarizadeh, M., Orooji, A., Sheikhtaheri, A., & Hayn, D. (2018). Effectiveness of Anonymization Methods in Preserving Patients' Privacy: A Systematic Literature Review. *eHealth*, *248*, 80-87.
Lapowsky, I. (2017, December 30). *Trolls, Bots, and Fake News Made 2017 a Terrible Year for Internet Freedom*. Wired. https://www.wired.com/story/internet-freedom-2017/
Loi, M., & Spielkamp, M. (2021, July). Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 757-766).
Macrae, C. (2019). Governing the safety of artificial intelligence in healthcare. *BMJ quality & safety*, *28*(6), 495-498.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence* (pp. 153-183). Springer, Cham.

Navas, S. (2020). Producer Liability for AI-Based Technologies in the European Union. *International Law Research, 9*(1), 77-84.

National Institute of Standards and Technology (NIST). (2021). *Artificial Intelligence Risk Management Framework.* Federal Register, The Daily Journal of the United States Government. https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework#citation-1-p40811

OECD. (2021). Tools for trustworthy AI : A framework to compare implementation tools for trustworthy AI systems. *Documents de travail de l'OCDE sur l'économie numérique, n° 312*, Éditions OCDE, Paris, https://doi.org/10.1787/008232ec-en.

OECD. (2022, February). OECD Framework for the Classification of AI Systems. *OECD Digital Economy Paper*. https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf?expires=1652269451&id=id&accname=guest&checksum=CFE5F318317EA246D4F21C90EB75BEDC

Olson, R. S. (2018). Establishing public accountability, speaking truth to power and inducing political will for disaster risk reduction:'Gcho Rios+ 25'. In *Environmental Hazards* (pp. 59-68). Routledge.

Perlroth, N. (2021, October 15). *Apple Security Update Closes Spyware Flaw in iPhones, Macs and iWatches*. The New York Times. https://www.nytimes.com/2021/09/13/technology/apple-software-update-spyware-nso-group.html

Porter, J. (2020, August 17). *UK ditches exam results generated by biased algorithm after student protests*. The Verge. https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications

Poth, A., Meyer, B., Schlicht, P., & Riel, A. (2020, December). Quality Assurance for Machine Learning–an approach to function and system safeguarding. In *2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS)* (pp. 22-29). IEEE.

Regulation 2016/679. *General Data Protection Regulation. European Parliament, Council of the European Union*. https://eur- lex.europa.eu/legalcontent/EN/TXT/PDF/?uri =CELEX:32016R0679

Regulation 2021/0106. *Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence*. European Parliament, Council of the European Union. https://eur- lex.europa.eu/legalcontent/EN/TXT/?uri=CEL EX%3A52021PC0206

Reiersen, G., Dao, D., Lütjens, B., Klemmer, K., Zhu, X., & Zhang, C. (2021, July 23). *Tackling the Overestimation of Forest Carbon with Deep Learning and Aerial Imagery*. Climate Change AI. https://www.climatechange.ai/papers/icml2021/79

Richardson, B., & Gilbert, J. E. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *arXiv preprint arXiv:2112.05700*.

Riel, A., Kreiner, C., Macher, G., & Messnarz, R. (2017). Integrated design for tackling safety and security challenges of smart products and digital manufacturing. *CIRP annals*, *66*(1), 177-180.

Ribera, M., & Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. *In IUI Workshops* (Vol. 2327, p. 38).

Schütte, B., Majewski, L., & Havu, K. (2021). Damages Liability for Harm Caused by Artificial Intelligence–EU Law in Flux. *Helsinki Legal Studies Research Paper*, (69).

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to practices for responsible AI: closing the gap. *arXiv preprint arXiv:2006.04707*.

Sokol, K., & Flach, P. A. (2019). Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. *SafeAI@ AAAI*.

The Alan Turing Institute. (2019*). Explaning decisions made with AI: Part 1: The basics of explaning AI.*

Information Commissioner's Office (ICO). https://ico.org.uk/for-organisations/guide-to- data-protection/key-dp-themes/explaining- decisions-made-with-artificial- intelligence/part-1-the-basics-of-explaining-ai/

The Guardian. (2019, January 3). *Could AI counselling be the future of therapy?* https://www.theguardian.com/lifeandstyle/shortcuts/2019/jan/02/woebots-ai-counselling-future-therapy-mental-health

The New York Times (2022, April 22). *As Europe Approves New Tech Laws, the U.S. Falls Further Behind.* https://www.nytimes.com/2022/04/22/technology/tech-regulation-europe-us.html

Institute for Ethics in Artificial Intelligence (IEAI). (2020, October). *AI Ethics: Why does it matter?* https://ieai.mcts.tum.de/wp-content/uploads/2020/10/Research-Brief_WhyAIEthicsMatter_Final-1.pdf

Institute for Ethics in Artificial Intelligence (IEAI). (2022, March). *Workshop – Summary & Outcomes Accountability Requirements for AI Applications* [Slides]. Https://Ieai.Mcts.Tum.De/. https://ieai.mcts.tum.de/wp-content/uploads/2022/04/March-Workshop_Outcomes.pdf

UNESCO. (2021, November). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

World Economic Forum. (2021, March). *Global Gender Gap Report 2021 INSIGHT REPORT MARCH 2021*. https://www3.weforum.org/docs/WEF_GGGR_2021.pdf

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & Perrault, R. (2021). The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

Zech, H. (2022). Haftung für Trainingsdaten Künstlicher Intelligenz. *Neue Juristische Wochenschrift (NJW)*, 502–507.