

November 2023

Responsible Artificial Intelligence: A Structured Literature Review

Sabrina GÖLLNER^{a,1}, Marina TROPMANN-FRICK^a and Boštjan BRUMEN^b

^a*Department of Computer Science, Hamburg University of Applied Sciences*

^b*Faculty of Electrical Engineering and Computer Science, University of Maribor*

Abstract. Our research endeavors to advance the concept of responsible artificial intelligence (AI), a topic of increasing importance within EU policy discussions. The EU has recently issued several publications emphasizing the necessity of trust in AI, underscoring the dual nature of AI as both a beneficial tool and a potential weapon. This dichotomy highlights the urgent need for international regulation. Concurrently, there's a need for frameworks that guide companies in AI development, ensuring compliance with such regulations. Our research aims to assist lawmakers and machine learning practitioners in navigating the evolving landscape of AI regulation, identifying focal areas for future attention. This paper introduces a comprehensive and, to our knowledge, the first unified definition of responsible AI. Through a structured literature review, we elucidate the current understanding of responsible AI. Drawing from this analysis, we propose an approach for developing a future framework centered around this concept. Our findings advocate for a human-centric approach to Responsible AI. This approach encompasses the implementation of AI methods with a strong emphasis on ethics, model explainability, and the pillars of privacy, security, and trust.

Keywords. Artificial Intelligence, Responsible AI, Privacy-preserving AI, Explainable AI, Ethical AI, Trustworthy AI

1. Introduction

In the past years, a lot of research is being conducted to improve Artificial Intelligence (AI) even further, as it is already being used in many aspects of life and industry.

The European Commission published a series of papers [1,2,3] in which they address their strategy for AI. In their white paper on AI from 2020 "A European Approach to Excellence and Trust" the political options for promoting the use of AI while mitigating the risks associated with certain applications of this technology are set out. This proposal aims to establish a legal framework for trustworthy AI in Europe so that the second objective of building an ecosystem for trust can be implemented. The Framework should fully respect the values and rights of EU citizens. It is repeatedly emphasized that AI should be human-centered and that European values have a high priority. The papers also address challenging issues such as ethical issues, privacy, explainability, safety, and sustainability. It is pointed out how important security is in the context of AI and they also present a risk framework in five risk groups for AI systems in short form. The document

¹Corresponding Author: Sabrina Göllner, sabrina.goellner@haw-hamburg.de

authors recognize that "[EU] Member States are pointing at the current absence of a common European framework." This indicates that a common EU framework is missing and it is an important political issue.

The document "Communication on Fostering a European Approach to AI" represents a plan of the EU Commission, where numerous efforts are presented that are intended to advance AI in the EU or have already been undertaken. In the beginning, it is stated that the EU wants to promote the development of a "*human-centric, sustainable, secure, inclusive and trustworthy artificial intelligence (AI) [which] depends on the ability of the European Union*".

The Commission's goal is to ensure that excellence in the field of AI is promoted. Collaborations with stakeholders, building research capacity, environment for developers, and funding opportunities are talked about as well as bringing AI into the play for climate and environment. Part of the discussion on trust led to the question of how to create innovation. It was pointed out that the EU approach should be "*human-centered, risk-based, proportionate, and dynamic*."

The plan also says they want to develop "*cutting-edge, ethical and secure AI, (and) promoting a human-centric approach in the global context*".

At the end of the document there is an important statement: "*The revised plan, therefore, provides a valuable opportunity to strengthen competitiveness, the capacity for innovation, and the responsible use of AI in the EU*". The EC has also published the "Proposal for a Regulation laying down harmonized rules on artificial intelligence" which contains, for example, a list of prohibited AI practices and specific regulations for AI systems that pose a high risk to health and safety as well as some transparency requirements.

It becomes noticeable that terms in the mentioned political documents that are used to describe the goal of trustworthy AI, however, keep changing (are inconsistent), and remain largely undefined. The documents all reflect, on the one hand, the benefits and on the other hand the risks of AI from a political perspective. It becomes clear that AI can improve our lives, solves problems in many ways, and is bringing added value but also can be a deadly weapon. But on the other hand, the papers do not exactly define what trustworthy AI even means in concrete terms. Topics and subtopics are somehow addressed but there is no clear definition of (excellence and) trustworthiness, but more indirectly mentions some aspects which are important, e.g., ethical values, transparency, risks for safety as well as sustainability goals.

Furthermore, we believe that trust as a goal (as defined vaguely in the documents) is also not sufficient to deploy AI. Rather, we need approaches for a "*responsible AI*", which reflects on the EU values. This should of course also be trustworthy, but that concept covers just a part of the responsibility. Therefore, in this paper, our goal is to find out the state-of-the-art from the scientific perspective and whether there is a general definition for "*trustworthy AI*". Furthermore, we want to clarify whether or not there is a definition for "*responsible AI*". The latter should actually be in the core of the political focus if we want to go towards "*excellence*" in AI.

As a step towards responsible AI, we conduct a structured literature review that aims to provide a clear answer to what it means to develop a "*responsible AI*".

During our initial analysis, we found that there is a lot of inconsistency in the terminology overall, not only in the political texts. There is also a lot of overlap in the definitions and principles for responsible AI. In addition, similar/content-wise similar expressions exist that further complicate the understanding of responsible AI as a whole.

There are already many approaches in the analyzed fields, namely trustworthy, ethical, explainable, privacy-preserving, and secure AI, but there are still many open problems that need to be addressed in the future.

Best to our knowledge this is one of the first detailed and structured reviews dealing with responsible AI.

The paper is structured as follows. In the following section, our research methodology is explained. This includes defining our research aims and objectives as well as specifying the databases and research queries we used for searching. The third section is the analysis part in which we first find out which definitions for responsible AI exist in the literature so far. Afterward, we explore content-wise similar expressions and look for their definitions in the literature. These are then compared with each other. As a result, we extract the essence of the analysis to formulate our definition of responsible AI. The subsequent section then summarizes the key findings in the previously defined scopes which are part of our definition of responsible AI. We further conduct a qualitative analysis of every single paper regarding the terms "Trustworthy, Ethics, Explainability, Privacy, and Security" in a structured table and quantitative analysis of the study features. Furthermore, in the discussion part, we do specify the key points and describe the pillars for developing responsible AI. Finally, after mentioning the limitations of our work, we end with our conclusion and future work.

2. Research Methodology

To answer the research questions, a systematic literature review (SLR) was performed based on the guidelines developed in [4]. The process of doing the structured literature review in our research is described in detail in the following subsections and summarized in the Systematic Review Protocol.

2.1. Research Aims and Objectives

In the present research, we aim to understand the role of "Responsible AI" from different perspectives, such as privacy, explainability, trust, and ethics. Firstly, our aim is to understand what constitutes the umbrella term "responsible AI", and secondly, to get an overview of the state of the art in the field. Finally, we seek to identify the open problems, challenges, and opportunities where further research is needed.

In summary, we provide the following contributions:

1. Specify a concise Definition of "Responsible AI"
2. Analyze the state of the art in the field of "Responsible AI"

2.2. Research Questions Formulation

Based on the aims of the research, we state the following research questions:

- RQ1: What is a general or agreed on definition of "Responsible AI" and what are the associated terms defining it?
- RQ2: What does "Responsible AI" encompass?

2.3. Databases

In order to get the best results when searching for the relevant studies, we used the indexing data sources. These sources enabled us a wide search of publications that would otherwise be overlooked. The following databases were searched:

- ACM Digital Library (ACM)
- IEEE Explore (IEEE)
- SpringerLink (SL)
- Elsevier ScienceDirect (SD)

The reason for selecting these databases was to limit our search to peer-reviewed research papers only.

2.4. Studies Selection

To search for documents, the following search query was used in the different databases:
("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning"
OR "Neural Network" OR "AI" OR "ML") AND (Ethic* OR Explain* OR Trust*)
AND (Privacy*).

Considering that inconsistent terminology is used for "Artificial Intelligence", the terms "Machine Learning", "Deep Learning" and "Neural Network" were added, which should be considered synonyms. Because there are already many papers using the abbreviations AI and ML, these were included to the set of synonyms.

The phrases "Ethic", "Trust" and "Explain" as well as "Privacy" was included with an asterisk (*), for all combinations of the terms following the asterisk, are included in the results (e.g. explain*ability). The search strings were combined using the Boolean operator OR for inclusiveness and the operator AND for the intersection of all sets of search strings. These sets of search strings were put within parentheses.

The selection of the period of publication was set to two years: 2020 and 2021 to get all of the state-of-the-art papers. The search was performed in December 2021.

The results were sorted by relevance prior to the inspection, which was important because the lack of advanced options in some search engines returned many non-relevant results.

To exclude irrelevant papers, the authors followed a set of guidelines during the screening stage. Papers did not pass the screening if:

1. They mention AI in the context of cyber-security, embedded systems, robotics, autonomous driving or internet of things, or alike.
2. They are not related to the defined terms of responsible AI.
3. They belong to general AI studies.
4. They only consist of an abstract.
5. They are published as posters.

These defined guidelines were used to greatly decrease the number of full-text papers to be evaluated in subsequent stages, allowing the examiners to focus only on potentially relevant papers.

The initial search produced 10.313 papers of which 4.121 were retrieved from ACM, 1064 from IEEE, 1.487 from Elsevier Science Direct, and 3.641 from Springer Link.

The screening using the title, abstract, and keywords removed 6.507 papers. During the check of the remaining papers for eligibility, we excluded 77 irrelevant studies and 9 inaccessible papers. We ended up with 254 papers that we included for the qualitative and quantitative analysis (see Figure 1).

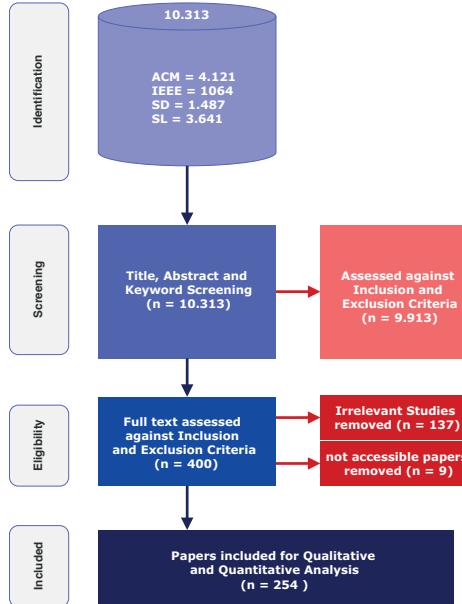


Figure 1. Structured review flow chart: the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart detailing the records identified and screened, the number of full-text articles retrieved and assessed for eligibility, and the number of studies included in the review.

3. Analysis

This section includes the analysis part in which we first find out which definitions for 'responsible AI' existed in the literature so far. Afterward, we explore content-wise similar expressions and look for their definitions in the literature. These definitions are then compared with each other and searched for overlaps. As a result, we extract the essence of the analysis to formulate our definition of responsible AI.

3.1. Responsible AI

In this subsection, we answer the first research question: What is a general or agreed on definition of 'Responsible AI', and what are the associated terms defining it?

3.1.1. Terms defining Responsible AI

Out of all 254 analyzed papers, we only found 5 papers that explicitly introduce aspects for defining "responsible" AI. The papers use the following terms in connection with 'responsible AI':

- Fairness, Privacy, Accountability, Transparency and Soundness [5]
- Fairness, Privacy, Accountability, Transparency, Ethics, Security & Safety [6]
- Fairness, Privacy, Accountability, Transparency, Explainability [7]
- Fairness, Accountability, Transparency, and Explainability [8]
- Fairness, Privacy, Sustainability, Inclusiveness, Safety, Social Good, Dignity, Performance, Accountability, Transparency, Human Autonomy, Solidarity [9]

However, after reading all 254 analyzed papers we strongly believe, that the terms that are included in those definitions can be mostly treated as subterms or ambiguous terms.

- 'Fairness'[5] and 'Accountability' [5,6,7], as well as the terms 'Inclusiveness, Sustainability, Social Good, Dignity, Human Autonomy, Solidarity' [9] according to our definition, are subterms of Ethics.
- 'Soundness'[5], interpreted as 'Reliability' or 'Stability', is included within Security and Safety.
- Transparency [5,6,7] is often used as a synonym for explainability in the whole literature.

Therefore we summarize these terms of the above definitions to: "Ethics, Trustworthiness, Security, Privacy, and Explainability". However, only the terms alone are not enough to get a picture of responsible AI. Therefore, we will analyze and discuss what the *meaning* of the five terms "Ethics, Trustworthiness, Security, Privacy, and Explainability" in the context of AI is, and how they *depend* on each other. During the analysis, we found also content-wise similar expressions to the concept of "responsible AI" which we want to include in the findings. This topic will be dealt with in the next section.

3.1.2. Content-wise similar expressions for Responsible AI

During the analysis, we found that the term "Responsible AI" is often used interchangeably with the terms "Ethical AI" or "Trustworthy" AI, and "Human-Centered AI" is a content-wise similar expression.

Therefore, we treat the terms:

- "Trustworthy AI", found in [10,11,12,13,14,15,16], and [17] as cited in [18]
- "Ethical AI", found in [19,20,21,22,23], and [24] as cited in [25]
- "Human-Centered AI", found in [26] as cited in [23]

as the *content-wise similar expressions* for "Responsible AI" hereinafter.

3.2. Collection of definitions

The resulting collection of definitions from 'responsible AI' and 'content-wise similar expressions for responsible AI' from the papers results in the following Venn diagram:

Analysis: We compared the definitions in the Venn diagram and determine the following findings:

- From all four sets there is an overlap of 24% of the terms: Explainability, Safety, Fairness, Accountability, Ethics, Security Privacy, Transparency.

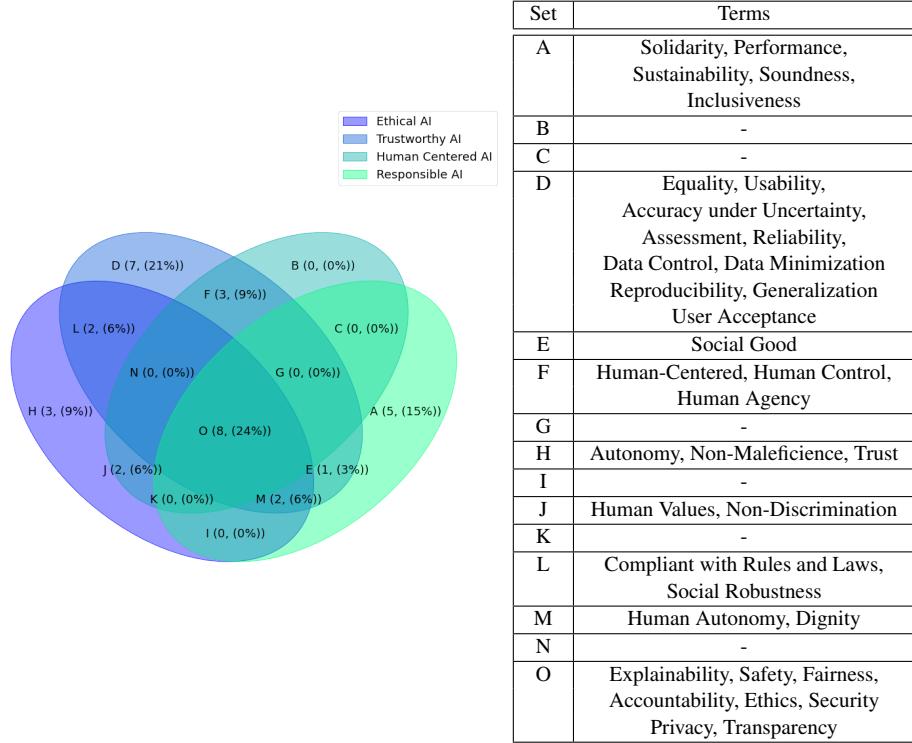


Figure 2. Venn diagram

- The terms occurring in the set of the definition for 'trust' only occurred in these, which is why this makes up the second largest set in the diagram. This is due to the fact that most of the terms actually come from definitions for trustworthy AI.
- There are also 6 null sets.

To tie in with the summary from the previous section, it should be pointed out once again that the terms 'Explainability, Safety, Fairness, Accountability, Ethics, Security Privacy, Transparency' can be grouped into generic terms as follows: Ethics, Security, Privacy, and Explainability.

We also strongly claim that 'trust/trustworthiness' should be seen as an outcome of a responsible AI system, and therefore we determine, that it belongs to the set of requirements. And each responsible AI should be built in a 'human-centered' manner, which makes it therefore another important subterm.

On top of these findings we specify our definition of Responsible AI in order to answer the first research question:

Responsible AI is **human-centered** and ensures users' **trust** through **ethical** ways of decision making. The decision-making must be fair, accountable, not biased, with good intentions, non-discriminating, and consistent with societal laws and norms. Responsible AI ensures, that automated decisions are **explainable** to users while always preserving users **privacy** through a **secure** implementation.

As mentioned in the sections before, the terms defining "responsible AI" result from the analysis of the terms in sections 3.1.1 and 3.1.2. We presented a figure depicting the

overlapping of the terms of content-wise similar expressions of Responsible AI, namely "Ethical AI, Trustworthy AI, and Human-Centered AI", and extracted the main terms of it. Also by summarizing the terms Fairness and Accountability into Ethics, and clarifying the synonyms (e.g., explainability instead of transparency), we finally redefined the terms defining "responsible AI" as "**Human-centered, Trustworthy, Ethical, Explainable, Privacy(-preserving) and Secure AI**".

3.3. Aspects of Responsible AI

According to our analysis of the literature, we have identified several categories in section 3 in connection to responsible AI, namely "Human-centered, Trustworthy, Ethical, Explainable, Privacy-preserving and Secure AI" which should ensure the development and use of it.

To answer the second research question (RQ2), we analyze the state-of-the-art of topics "Trustworthy, Ethical, Explainable, Privacy-preserving and Secure AI" in the following subsections. We have decided to deal with the topic of 'Human-Centered AI' in a separate paper so as not to go beyond the scope of this work.

To find out the state of the art of the mentioned topics in AI, all 118 papers were assigned to one of the categories "Trustworthy AI, Ethical AI, Explainable AI, Privacy-preserving AI, and Secure AI", based on the prevailing content of the paper compared to each of the topic. These papers were then analyzed and we highlight their most important features of them in the following subsections.

3.3.1. Trustworthy AI

A concise statement for trust in AI is as follows:

"Trust is an attitude that an agent will behave as expected and can be relied upon to reach its goal. Trust breaks down after an error or a misunderstanding between the agent and the trusting individual. The psychological state of trust in AI is an emergent property of a complex system, usually involving many cycles of design, training, deployment, measurement of performance, regulation, redesign, and retraining." [27]

Trustworthy AI is about delivering the promise of AI's benefits while addressing the scenarios that have vital consequences for people and society.

In this subsection, we summarize which are the aspects covered by the papers in the category "Trustworthy AI" and what are the issues to engender users' trust in AI.

Surveys and Reviews The following papers analyze trustworthy AI in their survey or review: [11,13,14,17,28,29]. The most important insights were the following:

- According to [13] "*Formal verification is a way to provide provable guarantees and thus increase one's trust that the system will behave as desired.*" However, this is more difficult with AI because of the inherently probabilistic nature of machine-learned models and the critical role of data in training, testing, and deploying a machine-learned model.

- The study of [29] observes that implementation projects of Trustworthy AI from which best practices can be derived can only be found in the research contexts and not in the industry, with only a few exceptions. It is further suggested to break down existing implementation guidelines to the requirements of software engineers, computer scientists, and managers while embedding also social scientists or ethicists in the implementation process.
- The 'best practices' for Trusted AI formulated by [14] are Data and model transparency, data governance, data minimization, assessment methods (for fairness), and access requirements.
- The review of [30] has revolved around trustworthy AI and discusses its need and importance and requirements as well as testing techniques for verification.

Perception of trust The following publications deal with how humans perceive Trust in AI: [31,32,33]. The interesting findings herein were as follows:

- The study [31] deals with analyzing users' trust in AI. Therefore, the authors examine the extent to which personal characteristics can be associated with perceptions of automated decision-making (ADM) through AI. The insight of the study was that Privacy can be seen as a central aspect as well as a human agency because people who felt they had more control over their own online information were more likely to view ADM as fair and useful.
- In the study of [32] the authors found out that "*The general public are not users of AI; they are subject to AI.*" There is a need for regulatory structures for trustworthiness.
- The study of [33] deals with Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. The key findings highlight that research around human experiences of AI should consider critical differences in social groups.

Frameworks Frameworks for "how developing Trustworthy AI can be achieved" were presented in [26,34,35]. The most important finding herein as the "chain of trust":

- The authors in the study [34] use various interrelated stages of a system life cycle within the development process for their concept. They then describe this process as forming the "Chain of Trust".
- [36] introduce a "new instrument to measure teachers' trust in AI-based EdTech, provides evidence of its internal structure validity, and uses it to portray secondary-level school teachers' attitudes toward AI."
- [37] develop a conceptual model called MATCH, which describes how trustworthiness is communicated in AI systems. They highlight transparency and interaction as AI systems' affordances that present a wide range of trustworthiness cues to users.
- [38] consider the challenge of verified AI from the perspective of formal methods for making AI more trustworthy.
- [15] we provide AI practitioners with a comprehensive guide for building trustworthy AI systems.
- [39] propose a framework and outlined case studies for applying modern data science to health care using a participatory design loop in which data scientists, clinicians, and patients work together.

- [40] describes an architecture to support scalable trustworthy ML and describes the features that have to be incorporated into the ML techniques to ensure that they are trustworthy.
- [41] conceptualize trust in AI in a multidimensional, multilevel way and examine the relationship between trust and ethics.

Miscellaneous In other papers [42,43] related to "Trustworthy AI", we found the following:

- Trust can be improved if the user can successfully understand the true reasoning process of the system (called intrinsic trust) [42]
- in the paper of [44] is about information fusion as an integrative cross-sectional topic to gain more trustworthiness, robustness, and explainability.
- In the paper of [45] reviews the state of the art in Trustworthy ML (TML) research and shed light on the multilateral tradeoffs, which are defined as the trade-offs among the four desiderata for TML they define as 'accuracy, robustness, fairness, and privacy' in the presence of adversarial attacks.
- [46] showed how to assess Trustworthy AI (based on the EU guidelines) in practice in times of pandemic based on a deep-learning-based solution deployed at a public hospital.
- The article of [16] discusses the tradeoffs between data privacy and fairness, robustness as well as explainability in the scope of trustworthy ML.
- [47] introduced the federated trustworthy Artificial Intelligence (FTAI) architecture.

Some papers we did not primarily categorize as "Trustworthy AI" (they rather belong to explainable AI) also mention important points dealing with trustworthiness:

- According to [12], understanding AI is another important factor to achieve trust. Understanding means how AI-led decisions are made and what determining factors were included that are crucial to understanding.
- Understanding is directly linked to the confidence if a model will act as intended when facing a given problem [6].
- According to [48], in addition to understanding, also knowing the prediction model's strengths and weaknesses is important for gaining trust.

We conclude that trust must be an essential goal of an AI application in order to be accepted in society and that every effort must be made to maintain and measure it at all times and in every stage of development. However, trustworthy AI still remains as a big challenge as it is not addressed (yet) holistically.

3.3.2. Ethical AI

In this subsection, we list the findings in the field of ethical AI. In our opinion, the definition found in [49] best describes ethics in conjunction with AI:

"AI ethics is the attempt to guide human conduct in the design and use of artificial automata or artificial machines, aka computers, in particular, by rationally formulating and following principles or rules that reflect our basic individual and social commitments and our leading ideals and values [49]."

Now we come to summarize the most important key points that came up while analyzing the literature.

Reviews and Surveys

- [50] reviews the ethical and human rights challenges and proposed mitigation strategies to discuss how a regulatory body could be designed to address these challenges.
- [51] gives a comprehensive overview of the field of AI ethics, including a summary and analysis of AI ethical issues, ethical guidelines, and principles, approaches to addressing AI ethical issues, and methods for evaluating AI ethics.
- In the survey of [52] an overview of the technical and procedural challenges involved in creating medical machine learning systems responsibly and in conformity with existing regulations, as well as possible solutions to address these challenges, are discussed.
- [53] conducted a scientometric analysis of publications on the ethical, legal, social, and economic (ELSE) implications of artificial intelligence.
- [54] provide a systematic scoping review to identify the ethical issues of AI application in healthcare.
- [55] conduct a semi-systematic literature review and thematic analysis to determine the extent to which the ethics of AI business practices are addressed in a wide range of guidelines.
- The review of [56] contributes to the debate on the identification and analysis of the ethical implications of algorithms which aims to analyze epistemic and normative concerns and offer actionable guidance for the governance of the design, development, and deployment of algorithms.

Frameworks Implementing Ethical AI is often discussed and structured in frameworks because the difficulty in moving from principles to practice presents a significant challenge to the implementation of ethical guidelines. As also stated in [22], there is still a significant gap. The following papers deal with solutions in the form of frameworks on this topic.

- In [57] the authors present a systematic framework for "socially responsible AI algorithms." The topics of AI indifference and the need to investigate socially responsible AI algorithms are addressed.
- [21] provided theoretical grounding of a concept named 'Ethics as a Service'.
- [58] developed a choices framework for the responsible use of AI for organizations which should help them to make better decisions toward the ethical use of AI. They distinguish between AI-specific technical choices e.g. continuous learning and generic digital technical choices, e.g., privacy, security, and safety.
- [59] proposed the "Ethics by design" framework which can be used to guide the development of AI systems. The "I" is based on three main aspects, "intelligibility, fairness, auditability" with the prototyping phase being crucial to establishing a solid ethical foundation for these systems.
- The article [7] also discusses about the possibility of developing ethical AI in a company by means of a framework. Different steps that lead through the whole development phase are discussed. It is also emphasized that rigorous testing and continuous measurement are of high importance to ensure that the system remains ethical and effective throughout its life cycle.

- In [60] two frameworks are presented including one for a responsible design process and another for better resolution of technology experience to help address the difficulty of moving from principle to practice in ethical impact assessment.
- [61] presents "ECCOLA", a method using some kind of gamification method for implementing ethically aligned AI systems.
- [62] advocates the use of licensing to enable legally enforceable behavioral use conditions on software and code and provides several case studies that demonstrate the feasibility of behavioral use licensing. It's envisioned how licensing may be implemented in accordance with existing responsible AI guidelines.
- [63] present TEDS as a new ethical concept, which focuses on the application of phenomenological methods to detect ethical errors in digital systems.
- [64] present a framework for assessing AI ethics and show applications in the field of cybersecurity.
- [65] propose a framework for developing and designing AI components within the Manufacturing sector under responsible AI scrutiny (i.e. framework for developing ethics in/by design).
- [66] presents a novel approach for the assessment of the impact of bias to raise awareness of bias and its causes within an ethical framework of action.
- [67] relates the literature about AI ethics to the ethics of systemic risks, proposes a theoretical framework based on the ethics of complexity as well as applies this framework to discuss implications for AI ethics.
- [68] propose an extension to FMEA, the "Failure mode and effects analysis", which is a popular safety engineering method, called "FMEA-AI" to support the conducting of "AI fairness impact assessments" in organizations.
- [69] map AI ethical principles onto the lifecycle of an AI-based digital service and combining it with an explicit governance model to clarify responsibilities in operationalization.
- [70] summarise normative ethical theories to a set of "principles for writing algorithms for the manufacture and marketing of artificially intelligent machines".
- [71] offer a solution-based framework for operationalizing ethics in AI for healthcare.
- [72] provide a holistic maturity framework in the form of an AI ethics maturity model that includes six critical dimensions for operationalizing AI ethics within an organization.

Tools

- [73] present a tool called: REvealing VIusal biaSEs (REVISE), that assists in the investigation of a visual dataset, surfacing potential biases along three dimensions: (1) object-based, (2) person-based, and (3) geography-based.

Ethical issues of AI The following papers discuss many of the existing ethical issues of AI:

- [74] identify the gaps in current AI ethics tools in auditing and risk assessment that should be considered.
- In [21] the *explainability problem* deals with the fact that an AI black-box model is difficult to make understandable, and the *public reason deficit*, the translation of code into a set of justifications in natural language.

- The work of [75] looks more closely at the concept of ethical debt in AI and its consequences. The authors point out that the biggest challenge is seen here as the discrepancy between those who incur debt and those who ultimately pay for it. There is concern that the AI industry does little to address the complex sociotechnical challenges and that the industry is predominantly composed of individuals least likely to be affected by ethical debt.
- The contribution of [76] to the literature on ethical AI concentrates on the work required to configure AI systems while addressing the AI engineer's responsibility and refers to situations in which an AI engineer has to evaluate, decide and act in a specific way during the development.
- [77] presents the findings of the "SHERPA project", which used case studies and a Delphi study to identify what people perceived to be ethical issues. The primary and frequent concern is privacy and data protection, which points to more general questions about the reliability of AI systems. Lack of transparency makes it more difficult to recognize and address questions of bias and discrimination. Safety is also a key ethical issue; mostly this involves autonomous driving or systems for critical health services. It then addresses ethical issues arising from general artificial intelligence and sociotechnical systems that incorporate AI.
- [78] points out different dilemmas like "Human Alienation", "Privacy Disclosure" and "Responsibility Issues". First, the author goes into various points such as human alienation (replacing human work with machines) which leads to higher unemployment rates; relying on smart technologies can lead to a decrease in independence; and the weakening of interpersonal relationships because of a closer relationship between man and machine. Second, the author addresses the issue of privacy leakage. He claims that service providers such as Google and Amazon are not complying with the General Data Protection Regulations in terms of completeness of information, clarity of language, and fairness of processing. Third, it will inevitably bring moral decision-making risks.
- [79] shows up several ethical issues an AI-Ethicist should consider when making decisions and especially the dilemma when an AI ethicist must weigh the extent to which his or her own success in communicating a recognized problem involves a high risk of reducing the chances of successfully solving the problem. This is then resolved through different ethical theories (such as virtue ethics, deontological ethics, and consequentialist ethics).
- [80] reviewed Nissenbaum's "four barriers" to accountability, addressing the current situation in which data-driven algorithmic systems have become ubiquitous in decision contexts.
- The key finding of [81] is, that there is indeed a notable gap between the practices of the analyzed companies and the key requirements for ethical/trustworthy AI.
- [82] assesses and compares existing critiques of current fairness-enhancing technical interventions in machine learning that draw from a range of non-computing disciplines e.g. philosophy.
- [83] explores the ethical issues of AI in environmental protection.
- The work of [84] outlines the ethical implications of AI from a climate perspective.
- The authors of [85] make the case for the emergence of novel kinds of bias with the use of algorithmic decision-making systems.

- [86] discusses blind spots regarding to topics that hold significant ethical importance but are hardly or not discussed at all in AI ethics.
- The critical discussion of [87] argues for the application of cognitive architectures for ethical AI and
- [88] provide an overview of some of the ethical issues that both researchers and end users may face during data collection and development of AI systems, as well as an introduction to the current state of transparency, interpretability, and explainability of systems in radiological applications.
- [89] contributes critically to the ethical discussion of AI principles, arguing that they are useless because they cannot in any way mitigate the racial, social, and environmental harms of AI technologies, and seeks to suggest alternatives, thinking more broadly about systems of oppression and more narrowly about accuracy and auditing.

Miscellaneous The papers of [90,19,91,92,93,94,22,21,95,96,97,98] deal with ethical AI in their studies whereas they handled miscellaneous topics.

- [90,94] evaluates existing ethical frameworks.
- [19] gives a review of the documents that were published about ethical principles and guidelines and the lessons learned from them.
- [92] surveyed the ethical principles and also their implementations. The paper suggested checklist-style questionnaires as benchmarks for the implementation of ethical principles of AI.
- [93] collected insights from a survey of machine learning researchers.
- The study [96] reports the use of an interdisciplinary AI ethics program for high school students. Using short stories during the study was effective in raising awareness, focusing discussion, and helping students develop a more nuanced understanding of AI ethical issues, such as fairness, bias, and privacy.
- [97] provides an empirical study to investigate the discrepancies between the intended design of fairness mitigation tools and their practice and use in context. The focus is on: disaggregated assessments of AI systems designed to reveal performance differences across demographic groups.
- [98] compare AI and Human Expert Collaboration in Ethical Decision Making and investigate how the expert type (human vs. AI) and level of expert autonomy (adviser vs. decider) influence trust, perceived responsibility, and reliance.
- [99] created a field guide for ethical mitigation strategies in machine learning through a web application.
- [8] adds 'data provenance' as an important prerequisite to the table for mitigating biases stemming from the data's origins and pre-processing to realize responsible AI-based systems.
- [100] aims to provide a multi-disciplinary assessment of how fairness for machine learning fits into the context of clinical trials research and practice.
- [101] perform an empirical study involving interviews with 21 scientists and engineers to understand the practitioners' views on AI ethics principles and their implementation.
- The work of [102] explores the design of interpretable and interactive human-in-the-loop interfaces that enable ordinary end-users with no technical or domain knowledge background to identify and potentially address potential fairness issues.

- The article of [103] is an attempt to outline ethical aspects linked to iHealth by focussing on three crucial elements that have been defined in the literature: self-monitoring, ecological momentary assessment (EMA), and data mining.
- [104] have surveyed hundreds of datasets used in the fair ML and algorithmic equity literature to help the research community reduce its documentation debt, improve the utilization of existing datasets, and the curation of novel ones.
- [105] surveyed the major ethical guidelines using content analysis and analyzed the accessible information regarding their methodology and stakeholder engagement.
- [106] introduce a business ethics perspective based on the normative theory of contractualism and conceptualize ethical implications as conflicts between the values of different interest groups.
- [107] proposes a comparative analysis of the AI ethical guidelines endorsed by China and by the EU.
- The empirical study of [108] deals with the ethics of using ML in psychiatric settings.
- [109] compares the discourses of computer ethics with AI ethics and discusses their similarities, differences, issues, and social impact.
- The work of [110] is moving from the AI practice towards principles: Ethical insights are generated from the lived experiences of AI designers working on tangible human problems, and then cycled upward to influence theoretical debates.
- The main aim of [111] has been to outline a new approach for AI ethics in heavy industry.
- [112] deals with research on pro-social rule breaking (PSRB) for AI.
- [113] aims to provide an ethical analysis of AI recruiting from a human rights perspective.
- [114] identify and discuss a set of advantages and ethical concerns related to incorporating recommender systems into the digital mental health ecosystem.
- The article of [115] focuses on the design and policy-oriented computer ethics while investigating new challenges and opportunities.
- The main goal of [116] was to shed philosophical light on how the responsibility for guiding the development of AI in a desirable direction should be distributed between individuals and between individuals and other actors.

We also generally found during our analysis that Ethical AI deals often with fairness, therefore this should be mentioned here. Fair AI can be understood as "*AI systems [which] should not lead to any kind of discrimination against individuals or collectives in relation to race, religion, gender, sexual orientation, disability, ethnicity, origin or any other personal condition. Thus, fundamental criteria to consider while optimizing the results of an AI system is not only their outputs in terms of error optimization but also how the system deals with those groups.*"[6]

In any case, the development of ethical artificial intelligence should be also subject to proper oversight within the framework of robust laws and regulations.

It is also stated, that transparency is widely considered also as one of the central AI ethical principles [61].

In the state-of-the-art overview of [117] the authors deal with the relations between explanation and AI fairness and examine, that fair decision-making requires extensive contextual understanding, and AI explanations help identify potential variables that are driving the unfair outcomes.

Mostly, transparency and explainability are achieved using so-called explainability (XAI) methods. Therefore, it is discussed separately in the following/next section.

3.3.3. Explainable AI

Decisions made by AI systems or by humans using AI can have a direct impact on the well-being, rights, and opportunities of those affected by the decisions. This is what makes the problem of the explainability of AI such a significant ethical problem. This subsection deals with the analysis of the literature in the field explainable AI (XAI).

We found an interesting definition in [6] which is quite suitable for defining explainable AI:

Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.[6]

In the following subsections, we highlight the most interesting aspects of XAI.

Black-box models problem According to [6] there is a trade-off between model explainability and performance. The higher accuracy comes at the cost of opacity: it is generally not possible to understand the reasons that explain why an AI system has decided the way it did, that it is the correct decision, or course of action was taken properly. This is what, according to the literature, is often called interchangeably, AI's "black box," "explainability," "transparency," "interpretability," or "intelligibility" problem [118] or "black box model syndrome" [119].

Another point to mention here is, that 'Explainable AI', which aims to open the black box of machine learning, might also be a Pandora's Box according to [120]. This means that opening the black box might undermine trust in an organization and its decision-making processes by revealing potential limitations of the data or model defects.

[121] claim also, that there is also a need for an explanation of how ML tools have been built, which requires documenting and justifying the technical choices that practitioners have made in designing such tools.

Synonyms for XAI These papers deal with the synonyms in context with XAI:

- There is not yet consensus within the research community on the distinction between the terms interpretability, intelligibility, and explainability, and they are often, though not always, used interchangeably [122,6,123].
- [48] says that usually, interpretability is used in the sense of understanding how the predictive model works as a whole. Explainability, on the other hand, is often used when explanations are given by predictive models that are themselves incomprehensible.
- [124] mentioned 36 more notions related to the concept of explainability in their systematic review (e.g., 'Actionability', 'Causality', 'Completeness', 'Comprehensibility', 'Cognitive relief', etc.) They also provided a description of each of these notions.

- According to [125] the lack of consistent terminology hinders the dialog about XAI.

Motivation for XAI The following papers address the motivation for XAI:

- The key motivation of XAI is to "(1) increase the trustworthiness of the AI, (2) increase the trust of the user in a trustworthy AI, or (3) increase the distrust of the user in a non-trustworthy AI" [42].
- Explainability should be also considered as a bridge to avoid the unfair or unethical use of the algorithm's outputs.[6]
- According to [48] other motivating aspects are causality, transferability, informativeness, fair and ethical decision-making, accountability, making adjustments, and proxy functionality.
- It should also help end-users to build a complete and correct mental model of the inferential process of either a learning algorithm or a knowledge-based system and to promote users' trust for its outputs, [124] and reliance on the AI system [126].

Reviews and Surveys

- [127] have reviewed explainable and interpretable ML techniques for various healthcare applications while also highlighting security, safety, and robustness challenges along with ethical issues.
- [128] provides a survey, that attempts to provide a comprehensive review of global interpretation methods that completely explain the behavior of the AI models along with their strengths and weaknesses.
- [129] presents an extensive systematic literature review of the use of knowledge graphs in the context of Explainable Machine Learning.
- [130] present a mini-review on explainable AI in health care, introducing solutions for XAI leveraging multi-modal and multi-center data fusion followed by two showcases of real clinical scenarios.
- [131] proposed survey explicitly details the requirements of XAI in Healthcare 5.0, the operational and data collection process.
- The review of [132] aims to provide a unified and comprehensive review of the latest XAI progress by discovering the critical perspectives of the rapidly growing body of research associated with XAI.

XAI Techniques There are many different XAI techniques discussed in the literature. [6] as well as [48] give a detailed overview of the known techniques and their strengths and weaknesses, therefore we will only cover this topic in short.

First, the models can be distinguished into two different approaches to XAI, the intrinsically transparent models and the Post-hoc explainability target models that are not readily interpretable by design. These so-called "black-box models" are the more problematic ones, because they are way more difficult to understand. The post-hoc explainability methods can then be distinguished further into model-specific and model-agnostic techniques.

We can also distinguish generally between data-independent and data-independent mechanisms for gaining interpretability as well as global and local interpretability methods.

- [133] highlight issues in explanation faithfulness when CNN models explain their predictions on images that are biased with systematic error and address this by developing Debiased-CAM to improve the truthfulness of explanations.

- In the work of [134] a comprehensive analysis of the explainability of Neural Network models in the context of power Side-Channel Analysis (SCA) is presented, to gain insight into which features or Points of Interest (PoI) contribute to the most to the classification decision
- [135] investigates the explainability of Generative AI for Code.
- In the work of [136] provides a formal framework for achieving and analyzing differential privacy in model explanations and highlights the possible tradeoffs between fidelity of explanations and data privacy.
- [137] deal with a transformation technique between black box models and explainable (as well as interoperable) classifiers on the basis of semantic rules via automatic recreation of the training datasets and retraining the decision trees (explainable models) in between.
- This research of [138] presented an architecture that supports the creation of semantically enhanced explanations for demand forecasting AI models.
- [139] introduced LEGIT, a model-agnostic framework that incorporates the benefits of locally interpretable explanations into graph sampling methods.
- [140] presents six different saliency maps that can be used to explain any face verification algorithm with no manipulation inside of the face recognition model.
- [141] focus on explaining by means of model surrogates the (mis)behavior of black-box models trained via federated learning.

Frameworks [142] presents a single framework for analyzing the robustness, fairness, and explainability of a classifier based on counterfactual explanations through a genetic algorithm.

Application areas of explainability methods Application areas of explainability methods are, for example, medicine and health care, where these methods have a great influence. As discussed in [119], the authors paid specific attention to the methods of data and model visualization and concluded that involving the medical experts in the analytical process helped improve the interpretability and explainability of ML models even more.

Evaluation of explainability methods Not only the explainability methods but also the evaluation of those is of great relevance.

- According to [124] two main ways are objective evaluations and the other is human-centered evaluations.
- [143] presented an explainability Fact Sheet Framework for guiding the development of new explainability approaches by aiding in their critical evaluation along the five proposed dimensions, namely: functional, operational, usability, safety, and validation.
- [144] presents a categorization of XAI design goals and evaluation methods. The authors used a mapping between design goals for different XAI user groups, namely: Novice Users, Data Experts, and AI Experts, and their evaluation methods. Further, they present a framework through a model and a series of guidelines to provide a high-level guideline for a multidisciplinary effort to build XAI systems. [145] proposed the "XAI test", an application-based evaluation method tailored to isolate the effects of providing the end user with different levels of information. It became clear that the evaluation of XAI methods is still in the early stages and has to be very specific due to different end-user requirements.

- [126] analyzes XAI tools in the public sector. The case study based on a goal-question-metric analysis of explainability aims to quantitatively measure three state-of-the-art XAI tools. The results show that experts welcome new insights and more complex explanations with multiple causes. They also point out that the different levels of complexity may be appropriate for different stakeholders depending on their backgrounds.
- [146] show that the generated explanations are volatile when the model training changes, which is not consistent with the classification task and model structure. This raises further questions about confidence in deep learning models for healthcare.
- [147] propose two new measures to evaluate explanations borrowed from the field of algorithmic stability: mean generalizability MeGe and relative consistency ReCo.

Stakeholders of XAI The target groups receiving the explanations need to be analyzed and their individual requirements are of great importance, as well as the usability of the software presenting these explanations:

- XAI researchers often develop explanations based on their own intuition rather than the situated needs of their intended audience [148].
- According to [6] it is very important that the generated explanations take into account the profile of the user who receives these explanations, the so-called audience.
- There are different user personalities to be considered, which probably require different explanation strategies, and these are not evenly covered by the current XAI tools [125].
- There are significant opportunities for UI/UX designers to contribute through new design patterns that make aspects of AI more accessible to different audiences [125].
- [48] suggests building blocks into "What to explain" (content type), "How to explain" (communication), and "to Whom is the explanation addressed" (target group).
- According to the study of [149] explanation interfaces with User-Centric Explanation ("why", "when", and "how" different users need transparency information) as well as "Interactive Explanation" methods, which increases awareness of how AI agents make decisions, play an important role.
- [148] introduced Social Transparency (ST), a sociotechnically informed perspective that incorporates socio-organizational context in explaining AI-mediated decision-making. ST makes visible the socially situated technological context: the trajectory of AI decision outcomes to date, as well as people's interactions with those technological outcomes. Such contextual information could help people calibrate trust in AI, not only by tracking AI performance but also by incorporating human elements into AI that could elicit socially based perception and heuristics.
- In the study of [150] proposes a three-tiered typology of stakeholder needs. It consists of long-term goals (understanding, building trust), shorter-term goals that work toward those goals (e.g., checking a model or questioning a decision), and immediate tasks that stakeholders can perform to achieve their goals (e.g., evaluating the reliability of predictions and detecting errors).

Miscellaneous

- [151] discuss the potential benefits of XAI and Human-In-The-Loop (HITL) methods in future work practices and illustrate how such methods can create new interactions and dynamics between human users and AI.
- The position paper of [152] brings together different roles and perspectives on XAI to explore the concept in-depth and offers a functional definition and framework for considering XAI in a medical context.
- [153] focus on the problem of explainable medical image retrieval using neural networks and different explanation methods.
- The article of [154] article seeks to provide technical explanations that can be given by XAI and to show how suitable explanations for liability can be reached in court.

The general public needs more transparency about how ML/AI systems can fail and what is at stake if they fail. Ideally, they should clearly communicate the outcomes and focus on the downsides to help people think about the trade-offs and risks of different choices (for example, the costs associated with different outcomes). But in addition to the general public also Data Scientists and ML Practitioners represent another key stakeholder group. In the study by [122] the effectiveness and interpretability of two existing tools: the InterpretML implementation of GAMs and the SHAP Python package were investigated. Their results indicate that data scientists over-trust and misuse interpretability tools.

There is a “right to explanation” in the context of AI systems that directly affect individuals through their decisions, especially in legal and financial terms, which is one of the themes of the General Data Protection Regulation (GDPR) [123,119]. Therefore we need to protect data through secure and privacy-preserving AI-methods. We will analyze this in the next section.

3.3.4. Privacy-preserving and Secure AI

As it was noted before, privacy and security are seen as central aspects of building trust in AI. However, the fuel for the good performance of ML models is data, especially sensitive data. This has led to growing privacy concerns, such as unlawful use of private data and disclosure of sensitive data[57,155]. We, therefore, need comprehensive privacy protection through holistic approaches to privacy protection that can also take into account the specific use of data and the transactions and activities of users [156] .

Privacy-preserving and Secure AI methods can help mitigate those risks. We define ”Secure AI” as protecting data from malicious threats, which means protecting personal data from any unauthorized third-party access or malicious attacks and exploitation of data. It is set up to protect personal data using different methods and techniques to ensure data privacy. Data privacy is about using data responsibly. This means proper handling, processing, storage, and usage of personal information. It is all about the rights of individuals with respect to their personal information. Therefore data security is a prerequisite for data privacy.

Security and privacy threats There are a lot of security threats in the branch of machine learning like stealing the model or sensitive information from the user, reconstruction attacks, poisoning attacks, and membership inference attacks, while the latter is a rapidly evolving research branch [157]. Selected papers deal with the security threats:

- [158] provides a brief review of these threats as well as the defense methods on security and privacy issues in such models while maintaining their performance and accuracy. Therefore they classify three defense methods: gradient-level, function-level, and label-level, which are based on the differential privacy theory.
- In the paper of [157] several of these membership inference attacks across a large number of different datasets were evaluated as well as the Differential Private Stochastic Gradient Descent (DP-SGD) method as a defense.
- [159] contribute to this topic while providing an evaluation and critical reflection upon why prominent robustness methods fail to deliver a secure system despite living up to their promises of adding robustness in the light of facial authentication.
- [160] provide an empirical Evaluation of Adversarial Examples Defences, Combinations, and Robustness Scores.
- [161] argue, that a language model's privacy can be hardly preserved by such methods as for example Differential Privacy and conclude that the language model should be trained on text data that was explicitly produced for public use.
- [162] study adversarial attacks on graph-level embedded methods.

The security threats need to be mitigated through techniques such as presented in the works of [163], where a privacy-preserving detection of poisoning attacks in Federated Learning is presented and in the paper of [164], which is about mitigating model poisoning in privacy-preserving Federated Learning.

Surveys and reviews on privacy-preserving techniques The following papers give an overview of privacy-preserving machine learning (PPML)-techniques through a survey:

- [165] evaluate privacy-preserving techniques in a comprehensive survey and propose a multi-level taxonomy, which categorizes the current state-of-the-art privacy-preserving deep learning techniques: (1) model training or learning, (2) PP inference or analysis, and (3) release a PP model.
- [166] summarize infrastructure support for privacy-preserving machine learning (PPML) techniques at both the software and hardware levels. The authors emphasize that the software/hardware co-design principle plays an important role in the development of a fully optimized PPML system.
- [167] provides a systematic review of deep learning methods for privacy-preserving natural language processing.
- [168] have discussed visual privacy attacks and defenses in the context of deep learning in their survey.

The different PPML- techniques will be presented in the next few sections:

Differential Privacy Differential Privacy (DP) is a strict mathematical definition of privacy in the context of statistical and ML analyses [166] and many procedures are based mainly on this concept. "Differentially private" means to design query responses in such a way that it is impossible to detect the presence or absence of information about a par-

ticular individual in the database [169]. In the context of PPML, DP typically works by adding noise to the training database. The challenge is the trade-off between privacy and precision for a dataset. This means the amount of noise added to the data is what allows the quantification of privacy of the dataset [170].

- DP approaches are described and used in [171,172,173].
- The study of [171] deals with the effects of differential privacy in the healthcare sector finding that DP-SDG is not well-suited for that kind of usage.
- The study of [173] deals with prescriptive analytics using DP using synthetic and real datasets and a new evaluation measure.
- The study of [172] focuses on a practical method for private deep learning in computer vision based on a k-nearest neighbor.
- A novel perturbed iterative gradient descent optimization (PIGDO) algorithm is proposed by [174].
- [175] propose a novel Local Differential Privacy (LDP)-based feature selection system, called LDP-FS, that estimates the importance of features over securely protected data while protecting the confidentiality of individual data before it leaves the user's device.
- [176] used a deep privacy-preserving CTG data classification model by adopting the Differential Privacy (DP) framework.
- The major contribution of [177] is adding differential privacy (DP) into continual learning (CL) procedures, aimed at protecting against adversarial examples.
- The paper of [178] proposes a novel model based on differential privacy named DA-PMLM that protects the sensitive data and classification model outsourced by multiple owners in a real cloud environment.
- In [179] a new differential privacy decision tree building algorithm is proposed and secondly, this is used for developing a two-phase differential privacy random forest method which increases the complementarity among decision trees.
- [180] propose a novel correlated differential privacy of the multiparty data release (MPCRDP).
- [181] provides a comparative evaluation of differentially private DL models in both input and gradient perturbation settings for predicting multivariate aggregate mobility time series data.

DP is often used in combination with other techniques like Homomorphic Encryption, Federated Learning, and Secure Multiparty Computation.

Homomorphic Encryption Homomorphic Encryption (HE) allows performing computations directly with encrypted data (ciphertext) without the need to decrypt them. The method is typically used as follows: First, the owner of the data encrypts it using a homomorphic function and passes the result to a third party tasked with performing a specific calculation; the third party then performs the computation using the encrypted data and returns the result, which is encrypted because the input data is encrypted. The owner of the data then decrypts the result and receives the result of the calculation with the original plaintext data. HE schemes support two types of computation: HE addition and HE multiplication [166]. Some noise is typically added to the input data during the encryption process. In order to get the expected result when decrypting, the noise must be kept below a certain threshold. This threshold affects the number of computations that can be

performed on encrypted data. The technique is used in the study of [182]. The following papers also deal with HE:

- In the study of [183] a privacy-preserving training algorithm for a fair support vector machine classifier based on Homomorphic Encryption (HE) is proposed, where the privacy of both sensitive information and model secrecy can be preserved.
- [184] propose a privacy-preserving logistic regression scheme based on CKKS, a leveled fully homomorphic encryption with the assistance of trusted hardware.
- [185] proposed a privacy-preserving ridge regression algorithm with homomorphic encryption of multiple private variables and suggested an adversarial perturbation method that can defend attribute inference attacks on the private variables.

Secure Multiparty Computation Secure Multiparty Computation (MPC / SMPC) is a cryptographic protocol that distributes computation among multiple parties, where no single party can see the other parties' data. The parties are independent and do not trust each other. The main idea is to allow to perform computation on private data while keeping the data secret. MPC guarantees that all participants learn nothing more than what they can learn from the output and their own input.

In [186] this approach is used for a framework named "Secure Decentralized Training Framework" which is able to operate in a decentralized network that does not require a trusted third-party server while ensuring the privacy of local data with low communication bandwidth costs.

In the study of [187] the authors use this technique for the development of a novel Privacy-preserving Speech Recognition framework using the Bidirectional Long short-term memory neural network based on SMC.

Federated Learning Federated Learning (FL) is a popular framework for decentralized learning and FL is the most common method for preserving privacy found in this analysis.

The central idea is to have a base model first shared and then trained with each client node. The ML provider then creates a global model and sends it to the selected clients. The local models are then updated and improved via backpropagation using the local dataset. The global model is updated by aggregating the local updates (through federated averaging) only using the minimum necessary information [188]. FL ensures the privacy of the local participants since the client's data never leaves its local platform and no updates from individual users are stored in the cloud. Two different federated learning settings exist: cross-device (very large number of mobile or IoT devices) and cross-silo (a small number of clients) [188].

In the state-of-the-art analysis of [189] the authors classify FL into different segmentations and algorithms used. They also show current scenarios where FL is used including the Google GBoard System, Smart Medical Care, Smart Finance, Smart Transportation, and Smart Educational Systems.

Also, [190] provides a brief introduction to key concepts in federated learning and analytics with an emphasis on how privacy technologies may be combined in real-world systems. In the article of [191] the authors specify their systematic literature on FL in the context of electronic health records for healthcare applications, whereas the survey of [192] is specifically about FL for smart healthcare. In [193] the authors present an ex-

tensive literature review to identify state-of-the-art Federated Learning applications for cancer research and clinical oncology analysis.

The following papers use special FL-Approaches in their study:

- In [194] a user privacy preservation model for cross-silo Federated Learning systems (CrossPriv) is proposed.
- The study of [195] a federated deep learning algorithm was developed for biomedical data collected from wearable IoT devices.
- In the literature FL is also used in [196] to create a federated parallel data platform (FPDP) including end-to-end data analytics pipeline.
- [197] use FL for the design of SAFELearn, a generic private federated learning design that enables efficient thwarting of strong inference attacks that require access to clients' individual model updates.
- [198] presents an efficient, private and byzantine-robust FL (SecureFL) framework considering the communication and computation costs are reduced without sacrificing robustness and privacy protection.
- [199] have been working on implementing SA for Python users in the context of the 'Flower FL Framework'.
- [200] introduce an approach for vertically partitioned FL setup achieving reduced training time and data-transfer time and enabling a changing sets of parties. (supports linear and regression models and SVM)
- [201] present FedAT, a novel Federated learning system with Asynchronous Tiers under Non-IID training data.
- [202] proposing a scalable privacy-preserving federated learning (SPPFL) against poisoning attacks. The main contribution is crossing the chasm between these two contrary issues of poisoning defense and privacy protection.
- [203] introduces a new problem set in a multi-device context called Federated Learning in Multi-Device Local Networks (FL-MDLN) as well as highlighting the challenges of the proposed setting.
- [204] presents a distributed FL framework in Trusted Execution Environment (TEE) to protect gradients from the perspective of hardware. The authors present the usage of trusted Software Guard eXtensions (SGX) as an instance to implement the FL as well as the proposal of an SGX-FL framework.
- The authors of [205] did not train a single global model, instead, the clients specialize in their local data in their approach and use other clients' model updates depending on the similarity of their respective data (based on a directed acyclic graph). The advantage of this approach are achieving more accuracy and less variance than through federated averaging.
- In this article of [206], the authors address the challenges of the standard FL techniques such as the vulnerability to data corruptions from outliers, systematic mislabeling, or even adversaries by proposing Auto-weighted Robust Federated Learning (ARFL), a novel approach that jointly learns the global model and the weights of local updates to provide robustness against corrupted data sources.
- The work of [207] deals with the challenge, that aggregating the data from different wearable devices to a central server introduces privacy concerns. Therefore they propose an architecture, CloudyFL, by deploying cloudlets close to wearable devices.

- [208] designed a federated decision tree-based random forest algorithm using FL and conducted our experiments using different datasets. The test set was considering a small number of corporate companies for collaborative machine learning.
- [209] propose FedNKD, which utilizes knowledge distillation and random noise, to enable federated learning to work dependably in the real world with complex data environments.
- [210] propose a robust model aggregation mechanism called FLARE for FL, which is designed for defending against state-of-the-art model poisoning attacks.
- In the paper of [211] a novel scheme based on blockchain architecture for Federated Learning data sharing is proposed.
- [212] present a novel decentralized Federated Learning algorithm, DECFEDAVG, obtained as a direct decentralization of the original Federated Learning algorithm, FEDAVG.
- [213] propose a privacy-preserving and verifiable decentralized federated learning framework, named PVD-FL.
- [214] present a blockchain-based trustworthy federated learning architecture to enhance the accountability and fairness of federated learning systems.
- In [215] proposed trust as a metric to enable secure federated learning through a mathematical framework for trust evaluation and propagation within a networked system.
- [216] developed a Blockchain-FL architecture to ensure security and privacy, which utilizes secure global aggregation and blockchain techniques to resist attacks from malicious edge devices and servers.
- In [217] the authors design a new framework, called HealthFed, that leverages Federated Learning and blockchain technologies to enable privacy-preserving and distributed learning among multiple clinician collaborators.
- [218] propose "VFL-R", a novel Vertical FL framework combined with a ring architecture for multi-party cooperative modeling.
- [219] presents a privacy-preserving federated learning-based approach, PriCell, for complex models such as convolutional neural networks.

Although Federated learning is a promising candidate for developing powerful models while preserving individual privacy and complying with the GDPR the following papers show the challenges and vulnerabilities of FL:

- The study of [220] highlights some vulnerabilities of this approach. Attacks in a federated setup can be classified as poisoning attacks (preventing the model to learn at all) or inference attacks (attacking the private data of the target participants).
- The study of [198] points out the weakness of destroying the integrity of the constructed model through byzantine attacks.
- The authors of [220] also point out that extensive communication is a big challenge too as well as system heterogeneity.
- Additionally to detecting these attacks and also identifying the attackers, [189] highlights the challenges of reducing communication overhead in the encryption process and solving the noise threshold of different scenarios.

According to the survey of [189] an 'ideal state of FL' can be considered if the FL model can be fully decentralized and the current development makes it clear, that there

are still many barriers on the way to this ideal state.

All in all, FL is still not enough to guarantee privacy, therefore it is often combined in hybrid mechanisms with other techniques as we will discuss in the section below.

Hybrid PPML-approaches There are many new papers looking at hybridizing approaches as these could be promising solutions for the future. A hybrid PPML approach can take advantage of each component, providing an optimal tradeoff between ML task performance and privacy overhead[156,165,166]. The following papers deal with hybrid approaches:

- The study of [221] proposes an FL approach based on Gaussian differential privacy, called Noisy-FL, which can more accurately track the changes in privacy loss during model training.
- [222] presents "PRICURE", a system that combines the complementary strengths of secure multiparty computation (SMPC) and differential privacy (DP) to enable privacy-compliant collaborative predictions between multiple model owners. SMPC is relevant for protecting data before inference, while DP targets post-inference protection to avoid attacks such as membership inference.
- The study of [223] proposes a deep learning framework building upon distributed differential privacy and a homomorphic argmax operator specifically designed to maintain low communication loads and efficiency.
- [224] present a privacy-preserving DNN model known as Multi-Scheme Differential Privacy (MSDP) depending on the fusion of Secure Multi-party Computation (SMC) and ϵ -differential privacy. The method reduces communication and computational cost at a minimal level.
- [225] presents the Sherpa.ai Federated Learning framework that is built upon a holistic view of federated learning and differential privacy. The study results both from exploring how the machine learning paradigm can be adapted to federated learning and from defining methodological guidelines for the development of artificial intelligence services based on federated learning and differential privacy.
- [226] proposes a privacy-preserving federated learning algorithm for medical data using homomorphic encryption in a secure multi-party computation setting for protecting the DL model from adversaries.
- [227] study the privacy protection strategy of enterprise information data based on consortium blockchain and federated learning.

Miscellaneous PPML-approaches There are a few interesting PPML approaches, as outlined below.

- In [228] the Split-learning method is used on 1D CNN models. Unfortunately, the results of the analysis show that it is possible to reconstruct the raw data from the activation of the split intermediate layer and this method needs further research.
- In [229] a fully decentralized peer-to-peer (P2P) approach to multi-party ML, which uses blockchain and cryptographic primitives to coordinate a privacy-preserving ML process between peering clients is proposed which produces a final model that is similar in utility to federated learning and has the ability to withstand poisoning attacks.
- In [230] the authors propose a decentralized secure ML-training platform called Secular for based on using a private blockchain and InterPlanetary File System (IPFS) networks.

- In [231] the authors use an automatic face de-identification algorithm that generates a new face from a face image that retains the emotions and non-biometric facial attributes of a target face based on the StyleGAN technique.
- In [232] the authors focus on designing an effective human-in-the-loop-aided (HitL-aided) scheme to preserve privacy in smart healthcare.
- [233] focuses on designing a human-in-the-loop-aided (HitL-aided) scheme to preserve privacy in smart healthcare.
- [234] investigates and analyzes machine learning privacy risks to understand the relationship between training data properties and privacy leakage and propose a privacy risk assessment scheme based on the clustering distance of training data.
- [235] propose a comprehensive approach for face recognition techniques in a privacy preserving manner, i.e., without compromising the privacy of individuals in exchanged data while considering together the concepts of privacy and accuracy.
- [236] contribute towards the development of more rigorous privacy-preserving methodologies capable of anonymizing case-based explanations without compromising their explanatory value.
- [237] designs a secure and efficient classification scheme based on SVM to protect the privacy of private data and support vectors in the calculation and transmission process.
- The authors of [238] introduce PrivPAS (A real time Privacy-Preserving AI System) as a novel framework to identify sensitive content.
- The authors of [239] present Sphinx, an efficient and privacy-preserving online deep learning system without any trusted third parties.
- The survey [240] explores the domain of personalized FL (PFL) to address the fundamental challenges of FL on heterogeneous data, a universal characteristic inherent in all real-world datasets.

PPML-Measurement techniques Developing good techniques to preserve privacy is one thing but good techniques to measure it is needed as well. In this regard, the paper by [165] suggests these measurement techniques: effectiveness, which is typically evaluated in terms of accuracy; efficiency, which primarily includes communication or computational overhead and execution time; and privacy, which is primarily evaluated in terms of direct and indirect guarantees against leakage.

We conclude that there is a lot of research related to privacy and security in the field of AI and there is no approach yet to achieve perfectly privacy-preserving and secure AI and many challenges are left open.

3.4. Quantitative analysis

The final set of 254 high-quality studies was selected for an in-depth analysis to aid in answering the presented research questions.

Our choice of features is based on their content in each of the following categories, "Trustworthy AI, Ethical AI, Explainable AI, Privacy-preserving AI, and Secure AI", as derived from section 3.2. We analyzed the papers quantitatively. Table 1 presents study features along with their absolute and percentile representations in the reviewed literature as well as their sources.

The distribution of the paper is as follows: most papers covered the topic "Privacy-Preserving and Secure AI", followed by "Ethical AI" and then "Explainable AI" and Trustworthy AI.

Within the topic "Privacy-Preserving and Secure AI", most papers belong to "Federated learning", obviously being a very emerging research field in the time frame.

There were also many different papers that were not assigned to any specific category (see "Miscellaneous") since the topic is very multifaceted.

In the topic area of "Ethical AI", the most common category was 'Miscellaneous', since the authors of the ethical AI field handle very different topics. In addition, second most of them could be assigned to the category 'ethical issues' since this is a hot topic in the field of ethics. The rest of the papers dealt with ethical frameworks that try to integrate ethical AI in context of a development process.

Most studies in the field of XAI deal with coming up with new XAI approaches to solve different explainability problems with new AI models. There were also a few that presented stakeholder analyses specifically in the context of explainability of AI models. Few of them presented miscellaneous topics that could not be assigned to any specific category or frameworks to integrate explainable AI.

In Trustworthy AI, we saw that most presented a review or survey on the current state of Trustworthy AI in research. There were also papers presented frameworks specially for trustworthiness or papers that reported on how Trust is perceived and described by different users.

Feature	Repr.	Perc.	Sources
Trustworthy AI (28/254, 11%) *			
Reviews and Surveys	9/28	32%	[11,17,28,13,29,14,30,130,45]
Perceptions of trust	4/28	14%	[31,32,33,27]
Frameworks	9/28	32%	[26,34,35,37,38,15,39,40,41]
Miscellaneous	6/28	28%	[42,43,44,46,16,47]
Ethical AI (85/254,34%) *			
Frameworks	19/85	22%	[57,58,59,7,20,60,61,24,62,63] [64,65,66,67,68,69,70,71,72]
Ethical issues	22/85	26%	[74,20,118,241,78,242,243,84] [76,244,77,49,79,155,75,82] [80,81,85,86,87,89]
Miscellaneous	33/85	39%	[90,19,91,92,245,93,94,22,21,95,96] [97,98,99,100,9,101,103] [114,116,102,104,105,106,107] [108,109,110,111,112,113,115,8]
Reviews and Surveys	10/85	12%	[50,127,51,83,52,53,54,55,56,117]
Tools	1/85	1%	[73]
Explainable AI (46/254 , 18%) *			
Reviews and Surveys	10/46	22%	[6,48,123,12,149,119] [124,128,131,132]
Stakeholders	7/46	15%	[125,148,246,145] [122,150,126]
XAI Approaches	14/46	30%	[247,5,143,182,248,133] [134,135,137,129,139,140,141,138]
Frameworks	4/46	9%	[144,142,249,36]
Miscellaneous	11/46	24%	[250,251,136,151,152] [146,147,153,154,121,120]
Privacy-preserving and Secure AI (95/254 , 38%) *			
Reviews and Surveys	10/95	10%	[165,166,252,253,254,156] [169,188,167,168]
Differential Privacy	12/95	13%	[173,171,172,170,174,175] [176,177,178,179,180,181]
Secure Multi-Party Computation	2/95	2%	[186,187]
Homomorphic Encryption	4/95	4%	[182,183,184,185]
Federated learning	35/95	37%	[195,196,194,197,220,255,229,189] [207,208,213,205,198] [199,200,204,201,203,202,256,206,190] [191,192,209,211,210,212,214] [215,216,217,193,219,164,218]
Hybrid Approaches	8/95	xx%	[221,223,222,224,225,226,227,240]
Security Threats	7/95	8%	[157,158,159,160,161,163,162]
Miscellaneous	16/95	17%	[231,257,258,259,232,260,261,228,230,233] [234,235,236,237,238,239]

Table 1. Quantitative Analysis

*percentage does not add up to 100 due to rounding.

3.5. Qualitative analysis

The main categories of "Responsible AI", namely "Trustworthy AI, Ethical AI, Explainable AI, Privacy-preserving AI, and Secure AI", were defined in Section 3.2. The aspects of responsible AI were presented and discussed in detail in Section 3.3. Here, table 2 summarizes section 3.2. and 3.3. by presenting the qualitative analysis of the literature regarding the categories for responsible AI. Each of the papers was content-wise analyzed and checked for membership in the defined categories.

The legend in the table reads as follows: • = meets criteria (i.e., the focus of the paper covers the topic of the category); ○ = partially meets criteria (i.e., the paper covers the topic of the category but its focus is elsewhere); no circle = does not meet criteria (i.e., the paper does not deal with the topic of the category). Abbreviations in the table heads are defined as follows: Trustworthy AI = Tr. AI, Ethical AI = Eth. AI, Explainable AI = XAI, Privacy-preserving and Secure AI = PP & Sec. AI

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
1	Abbasi et al. 2022	[235]				•
2	Abou El Houda et al. 2022	[217]				•
3	Abolfazlian 2020	[155]	•	•	○	○
4	Abuadbba et al. 2020	[228]				•
5	Agarwal et al. 2020	[231]				•
6	Allahabadi et al. 2022	[46]	•	○	○	○
7	Alishahi et al. 2022	[175]				•
8	Anderson and Fort 2022	[111]		•		
9	Antunes et al. 2022	[191]				•
10	Aminifar et al. 2021	[257]				•
11	Araujo et al. 2020	[31]	•	•		○
12	Arcolezzi et al. 2022	[181]				•
13	Arrieta et al. 2020	[6]	•	•	•	•
14	Attard-Frost et al. 2022	[55]		•		

Table 2. Qualitative Analysis 1/8

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
15	Ayling and Chapman 2021	[74]	•	○	○	
16	Bacciu and Numeroso 2022	[139]		•		
17	Banerjee et al. 2022	[39]	•			
18	Bai et al. 2022	[234]			•	
19	Beckert 2021	[29]	•	•		
20	Belenguer 2022	[66]		•		
21	Bélisle-Pipon 2022	[105]		•		
22	Beilharz et al. 2021	[205]				•
23	Benefo et al. 2022	[53]		•		
24	Benjamins 2021	[58]		•	•	•
25	Bertino 2020	[156]	•			•
26	Bickley and Torgler 2021	[87]		•		
27	Biswas 2021	[253]				•
28	Boenisch et al. 2021	[261]				•
29	Bonawitz et al. 2022	[190]				•
30	Boulemtafes et al. 2020	[165]				•
31	Bourgais and Ibnouhsein 2021	[59]		•		○
32	Boyd 2022	[99]		•		
33	Brennen 2020	[125]	○		•	
34	Brown et al. 2022	[161]				•
35	Brusseau 2022	[110]		•		
36	Bruschi and Diamede 2022	[64]		•		
37	Burkart and Huber 2021	[48]	○	•	•	○
38	Byun et al. 2022	[185]				•
39	Can und Ersoy 2021	[195]				•
40	Chai et al. 2021	[201]				•
41	Chang and Shokri 2021	[254]				•
42	Chen et al. 2020	[166]				•
43	Chen et al. 2021	[196]	○			•
44	Chien et al. 2022	[100]		•		
45	Cheng et al. 2021	[57]	•	•	•	•
46	Cho et al. 2021	[201]				•
47	Choraś et al. 2020	[123]		•	•	
48	Choung et al. 2022	[41]		•		
49	Chowdhury et al. 2022	[193]				•

Table 3. Qualitative Analysis 2/8

4. Discussion

Several key points have emerged from the analysis. It has become clear that AI will have an ever-increasing impact on our daily lives, from delivery robots to e-health, smart nutrition and digital assistants, and the list is growing every day. AI should be viewed as a tool, not a system that has infinite control over everything. It should therefore not replace humans or make them useless, nor should it lead to humans no longer using their own intelligence and only letting AI decide. We need a system that we can truly call

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
50	Chuanxin et al. 2020	[221]				•
51	Colaner et al. 2021	[251]	•	•	•	○
52	Combi et al. 2022	[152]			•	
53	Contractor et al. 2022	[251]		•		
54	Cooper et al. 2022	[80]		•		
55	Diddee and Kansra 2020	[194]				•
56	Ding et al. 2022	[174]				•
57	Ehsan et al. 2021	[148]	•		•	
58	Ehsan et al. 2021b	[246]	•		•	
59	Eitel-Porter 2021	[7]	○	•	•	○
60	Fabris et al. 2022	[104]		•		
61	Fel et al. 2022	[147]		•		
62	Fereidooni et al. 2021	[197]				•
63	Fernandez-Quillez 2022	[88]		•		
64	Feng and Chen 2022	[227]				•
65	Forbes 2021	[94]	○	•		
66	Forsyth et al. 2021	[96]		•		○
67	Fung and Etienne 2022	[107]		•		
68	Gambelin 2021	[241]		•		
69	Ghamry et al. 2021	[230]				•
70	Gholami et al. 2022	[215]				•
71	Gill 2021	[242]	○	•	○	
72	Giordano at al. 2022	[162]				•
73	Girka et al. 2021	[258]				•
74	Gittens et al. 2022	[45]	•	•		•
75	Giorgieva et al. 2022	[69]		•		
76	Giuseppi et al. 2022	[212]				•
77	Golder et al. 2022	[134]			•	○
78	Goldsteen et al. 2021	[260]		•	•	•
79	Gong et al. 2022	[207]				•
80	Grivet Sébert et al. 2021	[223]				•
81	Guevara et al. 2021	[170]				•
82	Gupta and Singh et al. 2022	[178]				•
83	Ha et al. 2020	[158]				•
84	Haffar et al. 2022	[141]			•	

Table 4. Qualitative Analysis 3/8

“responsible” AI. The analysis has clearly shown that the elements of ethics, privacy, security and explainability are the true pillars of responsible AI, which should lead to a basis of trust.

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
85	Hagendorff 2020	[90]	○	●	●	●
86	Hagendorff 2022	[86]		●		
87	Hailemariam et al. 2020	[250]			●	
88	Hanna and Kazim 2021	[49]		●		○
89	Häußermann and Lütge 2022	[106]		●		
90	Hao et al. 2021	[198]				●
91	Harichandana et al. 2022	[238]				●
92	Harikumar et al. 2021	[173]	●			●
93	Hassanpour et al. 2022	[177]				●
94	He et al. 2020	[259]				●
95	Heuillet et al. 2021	[247]	○	○	●	
96	Hickok 2021	[19]		●		
97	Holzinger et al. 2022	[44]	○	○	○	○
98	Hu et al. 2021	[43]	●			
99	Hu et al. 2022	[153]		●		
100	Huang et al. 2022	[51]		●		
101	Hunkenschroer & Kriebitz 2022	[113]		●		
102	Ibáñez und Olmeda 2021	[22]	○	●	○	○
103	Jacovi et al. 2021	[42]	●	○	●	○
104	Jacobs and Simon 2022	[115]		●		
105	Jakesch et al. 2022	[9]		●		
106	Jain et al. 2020	[11]	●	●	●	●
107	Jancovic & Mayer 2022	[160]				●
108	Jarin and Eshete 2021	[222]				●
109	Jatain et al. 2021	[220]				●
110	Jesus et al. 2021	[145]	○		●	
111	Joisten et al., 2022	[63]		●		
112	Joos et al., 2022	[159]				●
113	Kalloori and Klingler 2022	[208]				●
114	Karimian et al. 2022	[54]		●		
115	Kaur et al. 2020	[122]	○		●	
116	Kaur et al. 2022	[30]	●			
117	Kiemde and Kora 2021	[91]		●		
118	Knowles and Richards 2021	[32]	●	○		
119	Krijger 2022	[72]		●		

Table 5. Qualitative Analysis 4/8

4.1. Pillars of Responsible AI

Here we highlight the most important criteria that a responsible AI should fulfill. These are also the points that a developer should consider if he wants to develop responsible AI. Therefore, they also form the pillars for the future framework.

Key-requirements for the Ethical AI are as follows:

- fair: non-biased and non-discriminating in every way,

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
120	Kumar et al. 2020	[17]	•	•	○	○
121	Kumar and Chowdhury 2022	[70]		•		
122	Lal and Kartikeyan 2022	[176]				•
123	Lee and Rich 2021	[33]	•			
124	Li et al. 2021	[199]				•
125	Li et al. 2021	[202]				•
126	Li et al. 2022	[206]				•
127	Li et al. 2022	[15]	•			
128	Li et al. 2022	[68]		•		
129	Li et al. 2022	[218]				•
130	Liao and Sundar 2022	[37]	•			
131	Lin et al. 2022	[83]		•		
132	Liu et al. 2021	[188]				•
133	Liu et al. 2022	[184]				•
134	Liu et al. 2022	[179]				•
135	Lo et al. 2022	[214]	○	○	○	•
136	Loi et al. 2020	[20]	•	•	○	
137	Lu et al. 2022	[101]	○	•	○	○
138	MacLure 2021	[118]		•	•	
139	Madaio et al. 2022	[97]		•		
140	Maltbie et al. 2021	[126]	○		•	
141	Mao et al. 2022	[237]				•
142	Ma et al. 2022	[164]				•
143	Maree et al. 2020	[5]			•	
144	Mercier et al. 2021	[252]				•
145	Mery and Morris 2022	[140]			•	
146	Middleton et al. 2022	[27]	•			
147	Milossi et al. 2021	[24]	○	•		○
148	Minh et al. 2021	[132]			•	
149	Mohseni et al. 2021	[144]	•		•	
150	Montenegro et al. 2022	[236]				•
151	Morley et al. 2021	[21]	•	○	○	
152	Mothukuri et al. 2021	[255]				•
153	Muhr et al. 2021	[163]				•
154	Mulligan & Elaluf-Calderwood 2022	[84]		•		

Table 6. Qualitative Analysis 5/8

- accountability: justifying the decisions and actions,
- sustainable: built with long-term consequences in mind, satisfying the Sustainable Development Goals,
- compliant: with robust laws and regulations.

Key-requirements for the privacy and security techniques are identified as follows:

- need to comply with regulations: HIPAA, COPPA, and more recently the GDPR (like, for example, the Federated Learning),

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
155	Munn 2022	[89]	•			
156	Nakao et al., 2022	[102]	•			
157	Nazaretsky et al., 2022	[36]	•			
158	Nguyen et al., 2022	[192]				•
159	Owusu-Agyemeng et al. 2021	[224]				•
160	Padovan et al. 2022	[154]			•	
161	Park et al. 2022	[183]	○			•
162	Patel et al. 2022	[136]			•	•
163	Persson & Hedlund 2022	[116]	•			
164	Peters et al. 2020	[60]	•	•	○	
165	Petersen et al. 2022	[52]	○	○	○	○
166	Petrozzino 2021	[75]	•			
167	Prunkl and Whittlestone 2020	[245]	•			
168	Raab 2020	[244]	•	•	○	•
169	Rahimian et al. 2021	[157]			•	•
170	Ramanayake et al. 2021	[112]	•			
171	Rasheed et al. 2022	[127]	○	○	•	○
172	Ratti et al. 2022	[121]			•	
173	Rochel and Evéquoz 2020	[76]	•			
174	Rodríguez-Barroso et al. 2020	[225]				•
175	Rozanec et al. 2022	[138]			•	
176	Rubeis 2022	[103]	•			
177	Saetra et al. 2021	[79]	•			
178	Saleem et al. 2022	[128]				•
179	Saraswat et al. 2022	[131]			•	
180	Sav et al. 2022	[219]				•
181	Sharma et al. 2020	[142]	•	•		
182	Shayan et al. 2021	[229]				•
183	Seshia et al. 2022	[38]	•			○
184	Sheth et al. 2021	[12]	•		•	
185	Shneiderman et al. 2020	[26]	•	•	○	○
186	Singh et al. 2021	[28]	•	•	•	•
187	Sokol and Flach 2020	[143]			•	○
188	Sokol and Flach 2020b	[249]	•		•	

Table 7. Qualitative Analysis 6/8

- need to be complemented by proper organizational processes,
- must be used depending on tasks to be executed on the data and on specific transactions a user is executing,
- use hybrid PPML-approaches because they can take advantage of each component, providing an optimal trade-off between ML task performance and privacy overhead,
- use techniques that reduce communication and computational cost (especially in distributed approaches).

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
189	Solanki 2022	[71]		•		
190	Sousa and Kern	[167]				•
191	Stahl 2021	[77]		•	•	•
192	Stahl et al. 2021	[243]	•	•		○
193	Stahl et al. 2022	[50]		•		
194	Stahl et al. 2022	[109]		•		
195	Starke et al. 2022	[108]		•		
196	Storey et al. 2022	[120]			•	
197	Strobel and Shokri 2022	[16]	•	○	○	○
198	Sun et al. 2021	[149]	•			•
199	Sun et al. 2022	[135]			•	
200	Suresh et al. 2021	[150]	•		•	
201	Suriyakumar et al. 2021	[171]	○			•
202	Svetlova et al. 2021	[67]		•		
203	Tan et al. 2022	[240]				•
204	Tartaglione and Grassetto 2020	[95]	•	•		○
205	Terziyan & Vitko 2022	[137]			•	
206	Thuraisingham 2022	[40]	•			
207	Tolmejer et al. 2022	[98]		•		
208	Toreini et al. 2020	[34]	•	•	•	•
209	Tian 2022	[239]				•
210	Tiddi and Schlobach 2022	[129]			•	
211	Tran et al. 2021	[186]				•
212	Tsamados et al. 2022	[56]		•		
213	Tsiakis and Murray 2022	[151]			•	
214	Utomo et al. 2022	[47]	○			•
215	Valentine 2022	[114]		•		
216	Vakkuri 2021	[61]	○	•		
217	Vakkuri et al. 2022	[81]		•		
218	Vellido et al. 2020	[119]			•	
219	Vilone and Logo 2021	[124]	•		•	
220	Waller and Waller 2022	[85]		•		
221	Wang et al. 2020	[187]				•
222	Wang et al. 2022	[210]				•
223	Wang et al. 2022	[211]				•

Table 8. Qualitative Analysis 7/8

Key-requirements for Explainable AI are the following:

- Human-Centered: the user interaction plays a important role and how he understands and interacts with the system,
- Explanations must be tailored to the user needs and target group
- Intuitive User interface/experience: the results need to be presented in a understandable visual language,
- Explainable is also feature to say how well the system does its work (non func-

	Author	Ref.	Tr. AI	Eth. AI	XAI	PP & Sec. AI
224	Wang et al. 2022	[73]		•		
225	Wang and Moulden 2021	[35]	•			
226	Watson 2022	[146]			•	
227	Weinberg 2022	[82]		•		
228	Werder et al. 2022	[8]	○	•	○	○
229	Wibawa 2022	[226]				•
230	Wing 2021	[13]	•	○	•	○
231	Wyhmeister et al. 2022	[65]		•		
232	Xiaoling et al. 2021	[262]	•	•		
233	Xu et al. 2021	[200]				•
234	Xu et al. 2021	[204]				•
235	Yang et al. 2021	[189]				•
236	Yang et al. 2022	[130]		•		
237	Yang et al. 2022	[216]				•
238	Yuan and Shen 2020	[182]				•
239	Yuan et al. 2020	[263]	○		•	
240	Zapechnikov et al.	[169]				•
241	Zhang et al. 2021	[93]	•	•		○
242	Zhang et al. 2021	[14]	•	○	○	○
243	Zhang et al. 2020	[256]				•
244	Zhang et al. 2022	[133]		○	•	
245	Zhang et al. 2022	[168]				•
246	Zhao et al. 2022	[213]				•
247	Zhao et al. 2022	[180]				•
248	Zhou et al. 2020	[232]	•	•		○
249	Zhou et al. 2020	[92]				•
250	Zhou et al. 2022	[233]				•
251	Zhou et al. 2022	[117]		•		
252	Zhu et al. 2020	[172]				•
253	Zhu et al. 2022	[209]				•
254	Zytek et al. 2021	[248]	•		•	

Table 9. Qualitative Analysis 8/8

tional requirement),

- Impact of explanations on decision making process,

Key-Perceptions of trustworthy AI are as follows:

- ensure user data is protected,
- probabilistic accuracy under uncertainty,
- provides an understandable, transparent, explainable reasoning process to the user,
- usability,
- act "as intended" when facing a given problem,
- perception as fair and useful,
- reliability.

We define Responsible AI as an interdisciplinary and dynamic process: it goes beyond technology and includes laws (compliance and regulations) and society standards such as ethics guidelines and the Sustainable Development Goals.

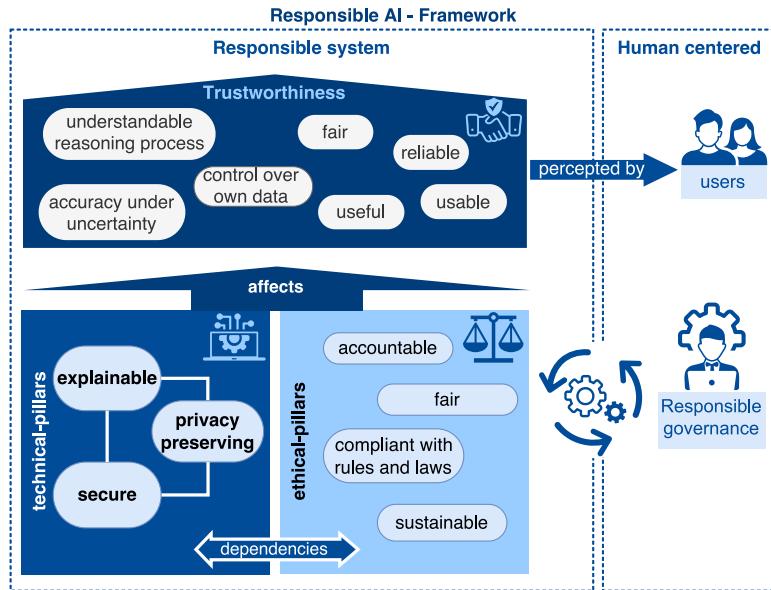


Figure 3. Pillars of the Responsible AI framework

Figure 3 shows that on the one hand there are social/ethical requirements/pillars and on the other hand the technical requirements/pillars. All of them are dependent on each other. If the technical and ethical side is satisfied the user trust is maintained. Trust can be seen as the perception of the users of AI.

There are also "sub-modules" present in each of the pillars, like accountability, fairness, sustainability and compliance in the field of ethics. They are crucial that we can say the AI meets ethical requirements.

Furthermore, the explainability methods must value privacy, meaning they must not have that much access to a model so that it results in a privacy breach. Privacy is dependent on security, because security is a prerequisite for it.

With each "responsible system" there are the humans that care for the system. The people who take care of the system must also handle it responsibly and constantly carry out maintenance work and check by metrics whether the responsibility is fulfilled. This can be ensured by special metrics which are considered as a kind of continuous check as standard. This means responsible AI encompasses the system-side and the developer-side.

Human-Centered AI (mentioned in 3.3) needs to be considered as a very important part of responsible AI and it is closely connected to the approach "Human-in-the-loop". The human in the loop here is very important because this is the person who checks and improves the system during the life cycle. so the whole responsible AI system needs to be Human-Centred, too. This topic will not be dealt with in detail in this study, but is a

part of the future work.

Therefore, responsible AI is interdisciplinary, and it is not a static but it is a dynamic process that needs to be taken care of in the whole system lifecycle.

4.2. Trade-offs

To fulfill all aspects comes with tradeoffs as discussed for example in [16] and comes for example at cost of data privacy. For example the methods that make model more robust against attacks or methots that try to explain a models behaviour and could leak some information. But we have fo find a way to manage that AI Systems that are accurate, fair, private, robust and explainable at the same time, which will be a very challenging task. We think that one approach to start with would be to create a benchmark for the different requirements that can determine to which proportion a certain requirement is fulfilled, or not.

5. Research Limitations

In the current study, we have included the literature available through various journals and provided a comprehensive and detailed survey on the literature in the field of responsible AI.

In conducting the study, we unfortunately had the limitation that some journals were not freely accessible despite a comprehensive access provided by our institutions. Although we made a good effort to obtain the information needed for the study on responsible AI from various international journals, accessibility was still a problem. It is also possible that some of the relevant research publications are not listed in the databases we used for searching. Additional limitation is the time frame of searched articles; this was carefully addressed to include only the state-of-the-art in the field. However, some older yet still current developement might have been missed out.

6. Conclusion

The field of AI is such a fast changing area and a legal framework for responsible AI is strongly necessary. From the series of EU-Papers on Artificial Intelligence of the last 2 years we noticed that "trustworthy AI" and "responsible AI" are not clearly defined, and as such a legal framework could not be efficiently established. Hence, the trust as a goal to define a framework/regulation for AI is not sufficient. Regulations for 'responsible AI' need to be defined instead. As the EU is a leading authority when it comes to setting standards (like the GDPR) we find it is absolutely necessary to help the politicians to really know what they are talking about. On the other hand, helping practitioners to prepare for what is coming next in both research and legal regulations is also of great importance.

The present research made important contributions to the concept of responsible AI. It is the first contribution to wholly address the "responsible AI" by conducting a structured literature research, and an overarching definition is presented as a result. The structured literature review covered 118 most recent high quality works on the topic. We have included a qualitative and quantitative analysis of the papers covered.

By defining "responsible AI" and further analyzing the state of the art of its components

(i.e., Human-centered, Trustworthy, Ethical, Explainable, Privacy(-preserving) and Secure AI), we have shown which are the most important parts to consider when developing AI products and setting up legal frameworks to regulate their development and use. In the discussion section we have outlined an idea for developing a future framework in the context of Responsible AI based on the knowledge and insights gained in the analysis part.

In future research the topic of Human-Centred AI and "Human-in-the-loop" should be developed further in the context responsible AI. Other important topics to be worked upon are the benchmarking approaches for responsible AI and a holistic framework for Responsible AI as the overarching goal.

References

- [1] European Commission. White Paper on Artificial Intelligence A European approach to excellence and trust. European Commission,; 2020. Available from: <https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence>.
- [2] European Commission. Coordinated Plan on Artificial Intelligence 2021 Review. European Commission,; 2021. Available from: <https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review>.
- [3] European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. European Commission,; 2021. Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>.
- [4] Kitchenham B, Brereton OP, Budgen D, Turne M, Bailey J, Linkman S. Systematic literature reviews in software engineering – A systematic literature review. Information and Software Technology. 2009;51:7-15.
- [5] Maree C, Modal JE, Omlin CW. Towards Responsible AI for Financial Transactions. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI); 2020. p. 16-21.
- [6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115. Available from: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- [7] Eitel-Porter R. Beyond the promise: implementing ethical AI. AI and Ethics. 2021;1(1):73-80.
- [8] Werder K, Ramesh B, Zhang RS. Establishing Data Provenance for Responsible Artificial Intelligence Systems. ACM Transactions on Management Information Systems. 2022 Jun;13(2):1-23. Available from: <https://dl.acm.org/doi/10.1145/3503488>.
- [9] Jakesch M, Buçinca Z, Amershi S, Olteanu A. How Different Groups Prioritize Ethical Values for Responsible AI. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 310-23. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533097>.
- [10] European Commission. Ethics guidelines for trustworthy AI e. European Commission,; 2019. Available from: <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>.
- [11] Jain S, Luthra M, Sharma S, Fatima M. Trustworthiness of Artificial Intelligence. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS); 2020. p. 907-12.
- [12] Sheth A, Gaur M, Roy K, Faldu K. Knowledge-Intensive Language Understanding for Explainable AI. IEEE Internet Computing. 2021;25(5):19-24.
- [13] Wing JM. Trustworthy AI. Commun ACM. 2021;64(10):64-71.
- [14] Zhang T, Qin Y, Li Q. Trusted Artificial Intelligence: Technique Requirements and Best Practices. In: 2021 International Conference on Cyberworlds (CW); 2021. p. 303-6. ISSN: 2642-3596.

- [15] Li B, Qi P, Liu B, Di S, Liu J, Pei J, et al. Trustworthy AI: From Principles to Practices. ACM Computing Surveys. 2022 Aug;3555803. Available from: <https://dl.acm.org/doi/10.1145/3555803>.
- [16] Strobel M, Shokri R. Data Privacy and Trustworthy Machine Learning. IEEE Security & Privacy. 2022 Sep;20(5):44-9. Available from: <https://ieeexplore.ieee.org/document/9802763/>.
- [17] Kumar A, Braud T, Tarkoma S, Hui P. Trustworthy AI in the Age of Pervasive Computing and Big Data. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops); 2020. p. 1-6.
- [18] Floridi L, Taddeo M. What is data ethics? Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences. 2016 12;374:20160360.
- [19] Hickok M. Lessons learned from AI ethics principles for future actions. AI and Ethics. 2021;1(1):41-7.
- [20] Loi M, Heitz C, Christen M. A Comparative Assessment and Synthesis of Twenty Ethics Codes on AI and Big Data. In: 2020 7th Swiss Conference on Data Science (SDS); 2020. p. 41-6.
- [21] Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds and Machines. 2021.
- [22] Ibáñez JC, Olmeda MV. Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI & SOCIETY. 2021.
- [23] Fjeld J, Achtern N, Hilligoss H, Nagy A, Srikanth M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication. 2020;(2020-1).
- [24] Milossi M, Alexandropoulou-Egyptiadou E, Psannis KE. AI Ethics: Algorithmic Determinism or Self-Determination? The GDPR Approach. IEEE Access. 2021;9:58455-66.
- [25] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines. 2018;28(4):689-707.
- [26] Shneiderman B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems. ACM Trans Interact Intell Syst. 2020;10(4).
- [27] Middleton SE, Letouzé E, Hossaini A, Chapman A. Trust, regulation, and human-in-the-loop AI: within the European region. Communications of the ACM. 2022 Apr;65(4):64-8. Available from: <https://dl.acm.org/doi/10.1145/3511597>.
- [28] Singh R, Vatsa M, Ratha N. Trustworthy AI. In: 8th ACM IKDD CODS and 26th COMAD. CODS COMAD 2021. New York, NY, USA: Association for Computing Machinery; 2021. p. 449-53.
- [29] Beckert B. The European way of doing Artificial Intelligence: The state of play implementing Trustworthy AI. In: 2021 60th FITCE Communication Days Congress for ICT Professionals: Industrial Data – Cloud, Low Latency and Privacy (FITCE); 2021. p. 1-8.
- [30] Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy Artificial Intelligence: A Review. ACM Computing Surveys. 2023 Mar;55(2):1-38. Available from: <https://dl.acm.org/doi/10.1145/3491209>.
- [31] Araujo T, Helberger N, Kruikemeier S, de Vreese CH. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & SOCIETY. 2020;35(3):611-23.
- [32] Knowles B, Richards JT. The Sanction of Authority: Promoting Public Trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 262-71.
- [33] Lee MK, Rich K. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2021. .
- [34] Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG, van Moorsel A. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 272-83.
- [35] Wang J, Moulden A. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21. New York, NY, USA: Association for Computing Machinery; 2021. .
- [36] Nazaretsky T, Cukurova M, Alexandron G. An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology. In: LAK22: 12th International Learning Analytics and Knowledge Conference. Online USA: ACM; 2022. p. 56-66. Available from: <https://dl.acm.org/doi/10.1145/3555803>

3506860.3506866.

- [37] Liao QV, Sundar SS. Designing for Responsible Trust in AI Systems: A Communication Perspective. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 1257-68. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533182>.
- [38] Seshia SA, Sadigh D, Sastry SS. Toward verified artificial intelligence. Communications of the ACM. 2022 Jul;65(7):46-55. Available from: <https://dl.acm.org/doi/10.1145/3503914>.
- [39] Banerjee S, Alsop P, Jones L, Cardinal RN. Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies. Patterns. 2022 Jun;3(6):100506. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2666389922000988>.
- [40] Thuraisingham B. Trustworthy Machine Learning. IEEE Intelligent Systems. 2022 Jan;37(1):21-4. Available from: <https://ieeexplore.ieee.org/document/9756264/>.
- [41] Choung H, David P, Ross A. Trust and ethics in AI. AI & SOCIETY. 2022 May. Available from: <https://link.springer.com/10.1007/s00146-022-01473-4>.
- [42] Jacovi A, Marasović A, Miller T, Goldberg Y. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 624-35.
- [43] Peng Hu, Yaobin Lu, Yeming (Yale) Gong. Dual humanness and trust in conversational AI: A person-centered approach. Computers in Human Behavior. 2021;119:106727. Available from: <https://www.sciencedirect.com/science/article/pii/S0747563221000492>.
- [44] Holzinger A, Dehmer M, Emmert-Streib F, Cucchiara R, Augenstein I, Ser JD, et al. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion. 2022 Mar;79:263-78. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1566253521002050>.
- [45] Gittens A, Yener B, Yung M. An Adversarial Perspective on Accuracy, Robustness, Fairness, and Privacy: Multilateral-Tradeoffs in Trustworthy ML. IEEE Access. 2022:1-1. Available from: <https://ieeexplore.ieee.org/document/9933776/>.
- [46] Allahabadi H, Amann J, Balot I, Beretta A, Binkley C, Bozenhard J, et al. Assessing Trustworthy AI in times of COVID-19. Deep Learning for predicting a multi-regional score conveying the degree of lung compromise in COVID-19 patients. IEEE. 2022:32.
- [47] Utomo S, John A, Rouniyar A, Hsu HC, Hsiung PA. Federated Trustworthy AI Architecture for Smart Cities. In: 2022 IEEE International Smart Cities Conference (ISC2). Pafos, Cyprus: IEEE; 2022. p. 1-7. Available from: <https://ieeexplore.ieee.org/document/9922069/>.
- [48] Burkart N, Huber MF. A Survey on the Explainability of Supervised Machine Learning. J Artif Int Res. 2021;70:245-317.
- [49] Hanna R, Kazim E. Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach. AI and Ethics. 2021.
- [50] Stahl BC, Rodrigues R, Santiago N, Macnish K. A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. Computer Law & Security Review. 2022 Jul;45:105661. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0267364922000097>.
- [51] Huang C, Zhang Z, Mao B, Yao X. An Overview of Artificial Intelligence Ethics. IEEE Transactions on Artificial Intelligence. 2022:1-21. Available from: <https://ieeexplore.ieee.org/document/9844014/>.
- [52] Petersen E, Potdevin Y, Mohammadi E, Zidowitz S, Breyer S, Nowotka D, et al. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions. IEEE Access. 2022;10:58375-418. Available from: <https://ieeexplore.ieee.org/document/9783196/>.
- [53] Benefo EO, Tingler A, White M, Cover J, Torres L, Broussard C, et al. Ethical, legal, social, and economic (ELSE) implications of artificial intelligence at a global level: a scientometrics approach. AI and Ethics. 2022 Jan. Available from: <https://link.springer.com/10.1007/s43681-021-00124-6>.
- [54] Karimian G, Petelos E, Evers SMAA. The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. AI and Ethics. 2022 Mar. Available from: <https://link.springer.com/10.1007/s43681-021-00131-7>.
- [55] Attard-Frost B. The ethics of AI business practices: a review of 47 AI ethics guidelines. AI and Ethics. 2022:18.

- [56] Tsamados A, Aggarwal N, Cowls J, Morley J, Roberts H, Taddeo M, et al. The ethics of algorithms: key problems and solutions. *AI & SOCIETY*. 2022 Mar;37(1):215-30. Available from: <https://link.springer.com/10.1007/s00146-021-01154-8>.
- [57] Cheng L, Varshney KR, Liu H. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *J Artif Int Res*. 2021;71:1137-81.
- [58] Benjamins R. A choices framework for the responsible use of AI. *AI and Ethics*. 2021;1(1):49-53.
- [59] Bourgais A, Ibnouhsein I. Ethics-by-design: the next frontier of industrialization. *AI and Ethics*. 2021.
- [60] Peters D, Vold K, Robinson D, Calvo RA. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*. 2020;1(1):34-47.
- [61] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, Pekka Abrahamsson. EC-COLA — A method for implementing ethically aligned AI systems. *Journal of Systems and Software*. 2021;182:111067. Available from: <https://www.sciencedirect.com/science/article/pii/S0164121221001643>.
- [62] Contractor D, McDuff D, Haines JK, Lee J, Hines C, Hecht B, et al. Behavioral Use Licensing for Responsible AI. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 778-88. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533143>.
- [63] Joisten K, Thiemer N, Renner T, Janssen A, Scheffler A. Focusing on the Ethical Challenges of Data Breaches and Applications. In: 2022 IEEE International Conference on Assured Autonomy (ICAA). Fajardo, PR, USA: IEEE; 2022. p. 74-82. Available from: <https://ieeexplore.ieee.org/document/9763591>.
- [64] Bruschi D, Diomede N. A framework for assessing AI ethics with applications to cybersecurity. *AI and Ethics*. 2022 May. Available from: <https://link.springer.com/10.1007/s43681-022-00162-8>.
- [65] Vyhmeister E, Castane G, Östberg PO, Thevenin S. A responsible AI framework: pipeline contextualisation. *AI and Ethics*. 2022 Apr. Available from: <https://link.springer.com/10.1007/s43681-022-00154-8>.
- [66] Belenguer L. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*. 2022 Feb. Available from: <https://link.springer.com/10.1007/s43681-022-00138-8>.
- [67] Svetlova E. AI ethics and systemic risks in finance. *AI and Ethics*. 2022 Nov;2(4):713-25. Available from: <https://link.springer.com/10.1007/s43681-021-00129-1>.
- [68] Li J, Chignell M. FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics*. 2022 Mar. Available from: <https://link.springer.com/10.1007/s43681-022-00145-9>.
- [69] Georgieva I, Lazo C, Timan T, van Veenstra AF. From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics*. 2022 Jan. Available from: <https://link.springer.com/10.1007/s43681-021-00127-3>.
- [70] Kumar S, Choudhury S. Normative ethics, human rights, and artificial intelligence. *AI and Ethics*. 2022 May. Available from: <https://link.springer.com/10.1007/s43681-022-00170-8>.
- [71] Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI and Ethics*. 2022 Jul. Available from: <https://link.springer.com/10.1007/s43681-022-00195-z>.
- [72] Krijger J, Thuis T, de Ruiter M, Ligthart E, Broekman I. The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations. *AI and Ethics*. 2022 Oct. Available from: <https://link.springer.com/10.1007/s43681-022-00228-7>.
- [73] Wang A, Liu A, Zhang R, Kleiman A, Kim L, Zhao D, et al. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision*. 2022 Jul;130(7):1790-810. Available from: <https://link.springer.com/10.1007/s11263-022-01625-5>.
- [74] Ayling J, Chapman A. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*. 2021.
- [75] Petrozzino C. Who pays for ethical debt in AI? *AI and Ethics*. 2021.
- [76] Rochel J, Evéquoz F. Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics. *AI & SOCIETY*. 2020.
- [77] Stahl BC, Antoniou J, Ryan M, Macnish K, Jiya T. Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY*. 2021.
- [78] Xiaoling P. Discussion on Ethical Dilemma Caused by Artificial Intelligence and Countermeasures. In:

- 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC); 2021. p. 453-7.
- [79] Sætra HS, Coeckelbergh M, Danaher J. The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. *AI and Ethics*. 2021 Dec. Available from: <https://doi.org/10.1007/s43681-021-00123-7>.
- [80] Cooper AF, Moss E, Laufer B, Nissenbaum H. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 864-76. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533150>.
- [81] Vakkuri V, Kemell KK, Tolvanen J, Jantunen M, Halme E, Abrahamsson P. How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis. In: The International Conference on Evaluation and Assessment in Software Engineering 2022. Gothenburg Sweden: ACM; 2022. p. 100-9. Available from: <https://dl.acm.org/doi/10.1145/3530019.3530030>.
- [82] Weinberg L. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research*. 2022 May;74:75-109. Available from: <https://jair.org/index.php/jair/article/view/13196>.
- [83] Lin H, Zhang Y, Chen X, Zhai R, Kuai Z. Artificial Intelligence Ethical in Environmental Protection. In: 2022 International Seminar on Computer Science and Engineering Technology (SCSET). Indianapolis, IN, USA: IEEE; 2022. p. 137-40. Available from: <https://ieeexplore.ieee.org/document/9700880/>.
- [84] Mulligan C, Elaluf-Calderwood S. AI ethics: A framework for measuring embodied carbon in AI systems. *AI and Ethics*. 2022 Aug;2(3):363-75. Available from: <https://link.springer.com/10.1007/s43681-021-00071-2>.
- [85] Waller RR, Waller RL. Assembled Bias: Beyond Transparent Algorithmic Bias. *Minds and Machines*. 2022 Sep;32(3):533-62. Available from: <https://link.springer.com/10.1007/s11023-022-09605-x>.
- [86] Hagendorff T. Blind spots in AI ethics. *AI and Ethics*. 2022 Nov;2(4):851-67. Available from: <https://link.springer.com/10.1007/s43681-021-00122-8>.
- [87] Bickley SJ, Torgler B. Cognitive architectures for artificial intelligence ethics. *AI & SOCIETY*. 2022 Jun. Available from: <https://link.springer.com/10.1007/s00146-022-01452-9>.
- [88] Fernandez-Quilez A. Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. *AI and Ethics*. 2022 Apr. Available from: <https://link.springer.com/10.1007/s43681-022-00161-9>.
- [89] Munn L. The uselessness of AI ethics. *AI and Ethics*. 2022 Aug. Available from: <https://link.springer.com/10.1007/s43681-022-00209-w>.
- [90] Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*. 2020;30(1):99-120.
- [91] Kiemde SMA, Kora AD. Towards an ethics of AI in Africa: rule of education. *AI and Ethics*. 2021.
- [92] Zhou J, Chen F, Berry A, Reed M, Zhang S, Savage S. A Survey on Ethical Principles of AI and Implementations. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI); 2020. p. 3010-7.
- [93] Zhang B, Anderljung M, Kahn L, Dreksler N, Horowitz MC, Dafoe A. Ethics and Governance of Artificial Intelligence: Evidence from a Survey of Machine Learning Researchers. *J Artif Int Res*. 2021;71:591-666.
- [94] Forbes K. Opening the path to ethics in artificial intelligence. *AI and Ethics*. 2021.
- [95] Tartaglione E, Grangetto M. A non-Discriminatory Approach to Ethical Deep Learning. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom); 2020. p. 943-50.
- [96] Forsyth S, Dalton B, Foster EH, Walsh B, Smilack J, Yeh T. Imagine a More Ethical AI: Using Stories to Develop Teens' Awareness and Understanding of Artificial Intelligence and its Societal Impacts. In: 2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT); 2021. p. 1-2.
- [97] Madaio M, Egede L, Subramonyam H, Wortman Vaughan J, Wallach H. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*. 2022 Mar;6(CSCW1):1-26. Available from: <https://dl.acm.org/doi/10.1145/3512899>.

- [98] Tolmeijer S, Christen M, Kandul S, Kneer M, Bernstein A. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In: CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM; 2022. p. 1-17. Available from: <https://dl.acm.org/doi/10.1145/3491102.3517732>.
- [99] Boyd K. Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 2069-82. Available from: <https://dl.acm.org/doi/10.1145/3531146.3534626>.
- [100] Chien I, Deliu N, Turner R, Weller A, Villar S, Kilbertus N. Multi-disciplinary fairness considerations in machine learning for clinical trials. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 906-24. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533154>.
- [101] Lu Q, Zhu L, Xu X, Whittle J, Douglas D, Sanderson C. Software engineering for responsible AI: an empirical study and operationalised patterns. In: Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice. Pittsburgh Pennsylvania: ACM; 2022. p. 241-2. Available from: <https://dl.acm.org/doi/10.1145/3510457.3513063>.
- [102] Nakao Y, Stumpf S, Ahmed S, Naseer A, Strappelli L. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. *ACM Transactions on Interactive Intelligent Systems*. 2022 Sep;12(3):1-30. Available from: <https://dl.acm.org/doi/10.1145/3514258>.
- [103] Rubeis G. iHealth: The ethics of artificial intelligence and big data in mental healthcare. *Internet Interventions*. 2022 Apr;28:100518. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2214782922000252>.
- [104] Fabris A, Messina S, Silvello G, Susto GA. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*. 2022 Sep. Available from: <https://link.springer.com/10.1007/s10618-022-00854-z>.
- [105] Bélisle-Pipon JC. Artificial intelligence ethics has a black box problem. *AI and Society*. 2022:16.
- [106] Häußermann JJ, Lütge C. Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics. *AI and Ethics*. 2022 May;2(2):341-62. Available from: <https://link.springer.com/10.1007/s43681-021-00047-2>.
- [107] Fung P, Etienne H. Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU. *AI and Ethics*. 2022 Jun. Available from: <https://link.springer.com/10.1007/s43681-022-00180-6>.
- [108] Starke G, Schmidt B, De Clercq E, Elger BS. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI and Ethics*. 2022 Jun. Available from: <https://link.springer.com/10.1007/s43681-022-00177-1>.
- [109] Stahl BC. From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI and Ethics*. 2022 Feb;2(1):65-77. Available from: <https://link.springer.com/10.1007/s43681-021-00080-1>.
- [110] Brusseau J. From the ground truth up: doing AI ethics from practice to principles. *AI & SOCIETY*. 2022 Jan. Available from: <https://link.springer.com/10.1007/s00146-021-01336-4>.
- [111] Anderson MM, Fort K. From the ground up: developing a practical ethical methodology for integrating AI into industry. *AI & SOCIETY*. 2022 Jul. Available from: <https://link.springer.com/10.1007/s00146-022-01531-x>.
- [112] Ramanayake R. Immune moral models? Pro-social rule breaking as a moral enhancement approach for ethical AI. *AI & SOCIETY*. 2022:13.
- [113] Hunkenschoer AL, Kriebitz A. Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring. *AI and Ethics*. 2022 Jul. Available from: <https://link.springer.com/10.1007/s43681-022-00166-4>.
- [114] Valentine L, D'Alfonso S, Lederman R. Recommender systems for mental health apps: advantages and ethical challenges. *AI & SOCIETY*. 2022 Jan. Available from: <https://link.springer.com/10.1007/s00146-021-01322-w>.
- [115] Jacobs M, Simon J. Reexamining computer ethics in light of AI systems and AI regulation. *AI and Ethics*. 2022 Oct. Available from: <https://link.springer.com/10.1007/s43681-022-00229-6>.
- [116] Persson E, Hedlund M. The future of AI in our hands? To what extent are we as individuals morally responsible for guiding the development of AI in a desirable direction? *AI*

- and Ethics. 2022 Nov;2(4):683-95. Available from: <https://link.springer.com/10.1007/s43681-021-00125-5>.
- [117] Zhou J, Chen F, Holzinger A. Towards Explainability for AI Fairness. In: Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W, editors. *xxAI - Beyond Explainable AI*. vol. 13200. Cham: Springer International Publishing; 2022. p. 375-86. Series Title: Lecture Notes in Computer Science. Available from: https://link.springer.com/10.1007/978-3-031-04083-2_18.
 - [118] Maclure J. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines*. 2021;31(3):421-38.
 - [119] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*. 2020;32(24):18069-83.
 - [120] Storey VC, Lukyanenko R, Maass W, Parsons J. Explainable AI. *Communications of the ACM*. 2022 Apr;65(4):27-9. Available from: <https://dl.acm.org/doi/10.1145/3490699>.
 - [121] Ratti E, Graves M. Explainable machine learning practices: opening another black box for reliable medical AI. *AI and Ethics*. 2022 Feb. Available from: <https://link.springer.com/10.1007/s43681-022-00141-z>.
 - [122] Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1-14.
 - [123] Chorąś M, Pawlicki M, Puchalski D, Kozik R. Machine Learning – The Results Are Not the only Thing that Matters! What About Security, Explainability and Fairness? In: Krzhizhanovskaya VV, Závodszky G, Lees MH, Dongarra JJ, Sloot PMA, Brissos S, et al., editors. *Computational Science – ICCS 2020*. vol. 12140. Cham: Springer International Publishing; 2020. p. 615-28.
 - [124] Giulia Vilone, Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*. 2021;76:89-106. Available from: <https://www.sciencedirect.com/science/article/pii/S1566253521001093>.
 - [125] Brennen A. What Do People Really Want When They Say They Want Explainable AI? We Asked 60 Stakeholders. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. CHI EA '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1-7.
 - [126] Maltbie N, Niu N, van Doren M, Johnson R. XAI Tools in the Public Sector: A Case Study on Predicting Combined Sewer Overflows. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2021. New York, NY, USA: Association for Computing Machinery; 2021. p. 1032-44.
 - [127] Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*. 2022 Oct;149:106043. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0010482522007569>.
 - [128] Saleem R, Yuan B, Kurugollu F, Anjum A, Liu L. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*. 2022 Nov;513:165-80. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0925231222012218>.
 - [129] Tiddi I, Schlobach S. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*. 2022 Jan;302:103627. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0004370221001788>.
 - [130] Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*. 2022 Jan;77:29-52. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1566253521001597>.
 - [131] Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, et al. Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access*. 2022;10:84486-517. Available from: <https://ieeexplore.ieee.org/document/9852458/>.
 - [132] Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*. 2022 Jun;55(5):3503-68. Available from: <https://link.springer.com/10.1007/s10462-021-10088-y>.
 - [133] Zhang W, Dimiccoli M, Lim BY. Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning. In: CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM; 2022. p. 1-32. Available from: <https://dl.acm.org/doi/10.1145/3491102.3517522>.
 - [134] Golder A, Bhat A, Raychowdhury A. Exploration into the Explainability of Neural Network Models for

- Power Side-Channel Analysis. In: Proceedings of the Great Lakes Symposium on VLSI 2022. Irvine CA USA: ACM; 2022. p. 59-64. Available from: <https://dl.acm.org/doi/10.1145/3526241.3530346>.
- [135] Sun J, Liao QV, Muller M, Agarwal M, Houde S, Talamadupula K, et al. Investigating Explainability of Generative AI for Code through Scenario-based Design. In: 27th International Conference on Intelligent User Interfaces. Helsinki Finland: ACM; 2022. p. 212-28. Available from: <https://dl.acm.org/doi/10.1145/3490099.3511119>.
- [136] Patel N, Shokri R, Zick Y. Model Explanations with Differential Privacy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 1895-904. Available from: <https://dl.acm.org/doi/10.1145/3531146.3533235>.
- [137] Terziyan V, Vitko O. Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models. Procedia Computer Science. 2022;200:216-26. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1877050922002290>.
- [138] Rožanec JM, Fortuna B, Mladenć D. Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). Information Fusion. 2022 May;81:91-102. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1566253521002414>.
- [139] Bacciu D, Numeroso D. Explaining Deep Graph Networks via Input Perturbation. IEEE Transactions on Neural Networks and Learning Systems. 2022;1-12. Available from: <https://ieeexplore.ieee.org/document/9761788/>.
- [140] Mery D, Morris B. On Black-Box Explanation for Face Verification. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE; 2022. p. 1194-203. Available from: <https://ieeexplore.ieee.org/document/9706895/>.
- [141] Haffar R, Sánchez D, Domingo-Ferrer J. Explaining predictions and attacks in federated learning via random forests. Applied Intelligence. 2022 Apr. Available from: <https://link.springer.com/10.1007/s10489-022-03435-1>.
- [142] Sharma S, Henderson J, Ghosh J. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 166-72.
- [143] Sokol K, Flach P. One Explanation Does Not Fit All. KI - Künstliche Intelligenz. 2020;34(2):235-50.
- [144] Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans Interact Intell Syst. 2021;11(3-4).
- [145] Jesus S, Belém C, Balayan V, Bento J, Saleiro P, Bizarro P, et al. How Can I Choose an Explainer? An Application-Grounded Evaluation of Post-Hoc Explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 805-15.
- [146] Watson M, Shiekh Hasan BA, Moubayed NA. Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE; 2022. p. 1524-33. Available from: <https://ieeexplore.ieee.org/document/9706847/>.
- [147] Fel T, Vigouroux D, Cadene R, Serre T. How Good is your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE; 2022. p. 1565-75. Available from: <https://ieeexplore.ieee.org/document/9706798/>.
- [148] Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding Explainability: Towards Social Transparency in AI Systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2021..
- [149] Sun L, Li Z, Zhang Y, Liu Y, Lou S, Zhou Z. Capturing the Trends, Applications, Issues, and Potential Strategies of Designing Transparent AI Agents. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21. New York, NY, USA: Association for Computing Machinery; 2021..
- [150] Suresh H, Gomez SR, Nam KK, Satyanarayan A. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2021..
- [151] Tsikas K, Murray-Rust D. Using human-in-the-loop and explainable AI to envisage new future work

- practices. In: The 15th International Conference on PErvasive Technologies Related to Assistive Environments. Corfu Greece: ACM; 2022. p. 588-94. Available from: <https://dl.acm.org/doi/10.1145/3529190.3534779>.
- [152] Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*. 2022 Nov;133:102423. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0933365722001750>.
 - [153] Hu B, Vasu B, Hoogs A. X-MIR: EXplainable Medical Image Retrieval. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE; 2022. p. 1544-54. Available from: <https://ieeexplore.ieee.org/document/9706900/>.
 - [154] Padovan PH, Martins CM, Reed C. Black is the new orange: how to determine AI liability. *Artificial Intelligence and Law*. 2022 Jan. Available from: <https://link.springer.com/10.1007/s10506-022-09308-9>.
 - [155] Abolfazlian K. Trustworthy AI Needs Unbiased Dictators! In: Maglogiannis I, Iliadis L, Pimenidis E, editors. *Artificial Intelligence Applications and Innovations*. Cham: Springer International Publishing; 2020. p. 15-23.
 - [156] Bertino E. Privacy in the Era of 5G, IoT, Big Data and Machine Learning. In: 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA); 2020. p. 134-7.
 - [157] Rahimian S, Orekondy T, Fritz M. Differential Privacy Defenses and Sampling Attacks for Membership Inference. In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. AISec '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 193-202. Available from: <https://doi.org/10.1145/3474369.3486876>.
 - [158] Ha T, Dang TK, Le H, Truong TA. Security and Privacy Issues in Deep Learning: A Brief Review. *SN Computer Science*. 2020;1(5):253.
 - [159] Joos S, Van hamme T, Preuveneers D, Joosen W. Adversarial Robustness is Not Enough: Practical Limitations for Securing Facial Authentication. In: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics. Baltimore MD USA: ACM; 2022. p. 2-12. Available from: <https://dl.acm.org/doi/10.1145/3510548.3519369>.
 - [160] Jankovic A, Mayer R. An Empirical Evaluation of Adversarial Examples Defences, Combinations and Robustness Scores. In: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics. Baltimore MD USA: ACM; 2022. p. 86-92. Available from: <https://dl.acm.org/doi/10.1145/3510548.3519370>.
 - [161] Brown H, Lee K, Mireshghallah F, Shokri R, Tramèr F. What Does it Mean for a Language Model to Preserve Privacy? In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 2280-92. Available from: <https://dl.acm.org/doi/10.1145/3531146.3534642>.
 - [162] Giordano M, Maddalena L, Manzo M, Guaracino MR. Adversarial attacks on graph-level embedding methods: a case study. *Annals of Mathematics and Artificial Intelligence*. 2022 Oct. Available from: <https://link.springer.com/10.1007/s10472-022-09811-4>.
 - [163] Muhr T, Zhang W. Privacy-Preserving Detection of Poisoning Attacks in Federated Learning. In: 2022 19th Annual International Conference on Privacy, Security & Trust (PST). Fredericton, NB, Canada: IEEE; 2022. p. 1-10. Available from: <https://ieeexplore.ieee.org/document/9851993/>.
 - [164] Ma Z, Ma J, Miao Y, Li Y, Deng RH. ShieldFL: Mitigating Model Poisoning Attacks in Privacy-Preserving Federated Learning. *IEEE Transactions on Information Forensics and Security*. 2022;17:1639-54. Available from: <https://ieeexplore.ieee.org/document/9762272/>.
 - [165] Boulemtafes A, Derhab A, Challal Y. A review of privacy-preserving techniques for deep learning. *Neurocomputing*. 2020;384:21-45. Available from: <https://www.sciencedirect.com/science/article/pii/S0925231219316431>.
 - [166] Chen H, Hussain SU, Boemer F, Stapf E, Sadeghi AR, Koushanfar F, et al. Developing Privacy-preserving AI Systems: The Lessons learned. In: 2020 57th ACM/IEEE Design Automation Conference (DAC); 2020. p. 1-4.
 - [167] Sousa S, Kern R. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*. 2022 May. Available from: <https://link.springer.com/10.1007/s10462-022-10204-6>.
 - [168] Zhang G, Liu B, Zhu T, Zhou A, Zhou W. Visual privacy attacks and defenses in deep learning: a survey. *Artificial Intelligence Review*. 2022 Aug;55(6):4347-401. Available from: <https://link.springer.com/10.1007/s10462-022-10204-6>.

- springer.com/10.1007/s10462-021-10123-y.
- [169] Sergey Zaepchikov. Privacy-Preserving Machine Learning as a Tool for Secure Personalized Information Services. *Procedia Computer Science*. 2020;169:393-9. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050920303598>.
- [170] Guevara M, Desfontaines D, Waldo J, Coatta T. Differential Privacy: The Pursuit of Protections by Default. *Commun ACM*. 2021;64(2):36-43.
- [171] Suriyakumar VM, Papernot N, Goldenberg A, Ghassemi M. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 723-34.
- [172] Zhu Y, Yu X, Chandraker M, Wang YX. Private-kNN: Practical Differential Privacy for Computer Vision. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. p. 11851-9.
- [173] Harikumar H, Rana S, Gupta S, Nguyen T, Kaimal R, Venkatesh S. Prescriptive analytics with differential privacy. *International Journal of Data Science and Analytics*. 2021.
- [174] Ding X, Chen L, Zhou P, Jiang W, Jin H. Differentially Private Deep Learning with Iterative Gradient Descent Optimization. *ACM/IMS Transactions on Data Science*. 2021 Nov;2(4):1-27. Available from: <https://dl.acm.org/doi/10.1145/3491254>.
- [175] Alishahi M, Moghtadaee V, Navidan H. Add noise to remove noise: Local differential privacy for feature selection. *Computers & Security*. 2022 Dec;123:102934. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0167404822003261>.
- [176] Lal AK, Karthikeyan S. Deep Learning Classification of Fetal Cardiotocography Data with Differential Privacy. In: 2022 International Conference on Connected Systems & Intelligence (CSI). Trivandrum, India: IEEE; 2022. p. 1-5. Available from: <https://ieeexplore.ieee.org/document/9924087/>.
- [177] Hassanpour A, Moradikia M, Yang B, Abdelhadi A, Busch C, Fierrez J. Differential Privacy Preservation in Robust Continual Learning. *IEEE Access*. 2022;10:24273-87. Available from: <https://ieeexplore.ieee.org/document/9721905/>.
- [178] Gupta R, Singh AK. A Differential Approach for Data and Classification Service-Based Privacy-Preserving Machine Learning Model in Cloud Environment. *New Generation Computing*. 2022 Jul. Available from: <https://link.springer.com/10.1007/s00354-022-00185-z>.
- [179] Liu J, Li X, Wei Q, Liu S, Liu Z, Wang J. A two-phase random forest with differential privacy. *Applied Intelligence*. 2022 Oct. Available from: <https://link.springer.com/10.1007/s10489-022-04119-6>.
- [180] Zhao JZ, Wang XW, Mao KM, Huang CX, Su YK, Li YC. Correlated Differential Privacy of Multiparty Data Release in Machine Learning. *Journal of Computer Science and Technology*. 2022 Feb;37(1):231-51. Available from: <https://link.springer.com/10.1007/s11390-021-1754-5>.
- [181] Arcolezi HH, Couchot JF, Renaud D, Al Bouna B, Xiao X. Differentially private multivariate time series forecasting of aggregated human mobility with deep learning: Input or gradient perturbation? *Neural Computing and Applications*. 2022 Aug;34(16):13355-69. Available from: <https://link.springer.com/10.1007/s00521-022-07393-0>.
- [182] Yuan L, Shen G. A Training Scheme of Deep Neural Networks on Encrypted Data. In: Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies. CIAT 2020. New York, NY, USA: Association for Computing Machinery; 2020. p. 490-5.
- [183] Park S, Byun J, Lee J. Privacy-Preserving Fair Learning of Support Vector Machine with Homomorphic Encryption. In: Proceedings of the ACM Web Conference 2022. Virtual Event, Lyon France: ACM; 2022. p. 3572-83. Available from: <https://dl.acm.org/doi/10.1145/3485447.3512252>.
- [184] Liu C, Jiang ZL, Zhao X, Chen Q, Fang J, He D, et al. Efficient and Privacy-Preserving Logistic Regression Scheme based on Leveled Fully Homomorphic Encryption. In: IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). New York, NY, USA: IEEE; 2022. p. 1-6. Available from: <https://ieeexplore.ieee.org/document/9797933/>.
- [185] Byun J, Park S, Choi Y, Lee J. Efficient homomorphic encryption framework for privacy-preserving regression. *Applied Intelligence*. 2022 Aug. Available from: <https://link.springer.com/10.1007/s10489-022-04015-z>.
- [186] Anh-Tu Tran, The-Dung Luong, Jessada Karnjana, Van-Nam Huynh. An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation. Neu-

- rocomputing. 2021;422:245-62. Available from: <https://www.sciencedirect.com/science/article/pii/S0925231220315095>.
- [187] Wang Q, Feng C, Xu Y, Zhong H, Sheng VS. A novel privacy-preserving speech recognition framework using bidirectional LSTM. *Journal of Cloud Computing*. 2020;9(1):36.
- [188] Liu B, Ding M, Shaham S, Rahayu W, Farokhi F, Lin Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Computing Surveys*. 2021;54(2).
- [189] Yang M, He Y, Qiao J. Federated Learning-Based Privacy-Preserving and Security: Survey. In: 2021 Computing, Communications and IoT Applications (ComComAp); 2021. p. 312-7.
- [190] Bonawitz K, Kairouz P, McMahan B, Ramage D. Federated learning and privacy. *Communications of the ACM*. 2022 Apr;65(4):90-7. Available from: <https://dl.acm.org/doi/10.1145/3500240>.
- [191] Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Transactions on Intelligent Systems and Technology*. 2022 Aug;13(4):1-23. Available from: <https://dl.acm.org/doi/10.1145/3501813>.
- [192] Nguyen DC, Pham QV, Pathirana PN, Ding M, Seneviratne A, Lin Z, et al. Federated Learning for Smart Healthcare: A Survey. *ACM Computing Surveys*. 2023 Apr;55(3):1-37. Available from: <https://dl.acm.org/doi/10.1145/3501296>.
- [193] Chowdhury A, Kassem H, Padov N, Umeton R, Karargyris A. A Review of Medical Federated Learning: Applications in Oncology and Cancer Research. In: Crimi A, Bakas S, editors. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Cham: Springer International Publishing; 2022. p. 3-24.
- [194] Diddee H, Kansra B. CrossPriv: User Privacy Preservation Model for Cross-Silo Federated Software. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering. ASE '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1370-2.
- [195] Can YS, Ersoy C. Privacy-Preserving Federated Deep Learning for Wearable IoT-Based Biomedical Monitoring. *ACM Trans Internet Technol*. 2021;21(1).
- [196] Chen L, Zhang W, Xu L, Zeng X, Lu Q, Zhao H, et al. A Federated Parallel Data Platform for Trustworthy AI. In: 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI); 2021. p. 344-7.
- [197] Fereidooni H, Marchal S, Miettinen M, Mirhoseini A, Möllering H, Nguyen TD, et al. SAFELearn: Secure Aggregation for private FEderated Learning. In: 2021 IEEE Security and Privacy Workshops (SPW); 2021. p. 56-62.
- [198] Hao M, Li H, Xu G, Chen H, Zhang T. Efficient, Private and Robust Federated Learning. In: Annual Computer Security Applications Conference. ACSAC. New York, NY, USA: Association for Computing Machinery; 2021. p. 45-60. Available from: <https://doi.org/10.1145/3485832.3488014>.
- [199] Li KH, de Gusmão PPB, Beutel DJ, Lane ND. Secure aggregation for federated learning in flower. In: Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning, DistributedML '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 8-14. Available from: <https://doi.org/10.1145/3488659.3493776>.
- [200] Xu R, Baracaldo N, Zhou Y, Anwar A, Joshi J, Ludwig H. FedV: Privacy-Preserving Federated Learning over Vertically Partitioned Data. In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. AISec '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 181-92. Available from: <https://doi.org/10.1145/3474369.3486872>.
- [201] Chai Z, Chen Y, Anwar A, Zhao L, Cheng Y, Rangwala H. FedAT: a high-performance and communication-efficient federated learning system with asynchronous tiers. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. SC '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 1-16. Available from: <https://doi.org/10.1145/3458817.3476211>.
- [202] Li Y, Hu G, Liu X, Ying Z. Cross the Chasm: Scalable Privacy-Preserving Federated Learning against Poisoning Attack. In: 2021 18th International Conference on Privacy, Security and Trust (PST); 2021. p. 1-5.
- [203] Cho H, Mathur A, Kawsar F. Device or User: Rethinking Federated Learning in Personal-Scale Multi-Device Environments. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. SenSys '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 446-52. Available from: <https://doi.org/10.1145/3485730.3493449>.
- [204] Xu T, Zhu K, Andrzejak A, Zhang L. Distributed Learning in Trusted Execution Environment: A Case Study of Federated Learning in SGX. In: 2021 7th IEEE International Conference on Network

- Intelligence and Digital Content (IC-NIDC); 2021. p. 450-4. ISSN: 2575-4955.
- [205] Beilharz J, Pfitzner B, Schmid R, Geppert P, Arnrich B, Polze A. Implicit model specialization through dag-based decentralized federated learning. In: Proceedings of the 22nd International Middleware Conference. Middleware '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 310-22. Available from: <https://doi.org/10.1145/3464298.3493403>.
- [206] Li S, Ngai E, Ye F, Voigt T. Auto-weighted Robust Federated Learning with Corrupted Data Sources. ACM Transactions on Intelligent Systems and Technology. 2022 Oct;13(5):1-20. Available from: <https://dl.acm.org/doi/10.1145/3517821>.
- [207] Gong Q, Ruan H, Chen Y, Su X. CloudyFL: a cloudlet-based federated learning framework for sensing user behavior using wearable devices. In: Proceedings of the 6th International Workshop on Embedded and Mobile Deep Learning. Portland Oregon: ACM; 2022. p. 13-8. Available from: <https://dl.acm.org/doi/10.1145/3539491.3539592>.
- [208] Kalloori S, Klingler S. Cross-silo federated learning based decision trees. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. Virtual Event: ACM; 2022. p. 1117-24. Available from: <https://dl.acm.org/doi/10.1145/3477314.3507149>.
- [209] Zhu S, Qi Q, Zhuang Z, Wang J, Sun H, Liao J. FedNKD: A Dependable Federated Learning Using Fine-tuned Random Noise and Knowledge Distillation. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. Newark NJ USA: ACM; 2022. p. 185-93. Available from: <https://dl.acm.org/doi/10.1145/3512527.3531372>.
- [210] Wang N, Xiao Y, Chen Y, Hu Y, Lou W, Hou YT. FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations. In: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. Nagasaki Japan: ACM; 2022. p. 946-58. Available from: <https://dl.acm.org/doi/10.1145/3488932.3517395>.
- [211] Wang Z, Yan B, Dong A. Blockchain Empowered Federated Learning for Data Sharing Incentive Mechanism. Procedia Computer Science. 2022;202:348-53. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1877050922005816>.
- [212] Giuseppi A, Manfredi S, Menegatti D, Pietrabissa A, Poli C. Decentralized Federated Learning for Nonintrusive Load Monitoring in Smart Energy Communities. In: 2022 30th Mediterranean Conference on Control and Automation (MED). Vouliagmeni, Greece: IEEE; 2022. p. 312-7. Available from: <https://ieeexplore.ieee.org/document/9837291/>.
- [213] Zhao J, Zhu H, Wang F, Lu R, Liu Z, Li H. PVD-FL: A Privacy-Preserving and Verifiable Decentralized Federated Learning Framework. IEEE Transactions on Information Forensics and Security. 2022;17:2059-73. Available from: <https://ieeexplore.ieee.org/document/9777682/>.
- [214] Lo SK, Liu Y, Lu Q, Wang C, Xu X, Paik HY, et al. Towards Trustworthy AI: Blockchain-based Architecture Design for Accountability and Fairness of Federated Learning Systems. IEEE Internet of Things Journal. 2022;1-1. Available from: <https://ieeexplore.ieee.org/document/9686048/>.
- [215] Gholami A, Torkzaban N, Baras JS. Trusted Decentralized Federated Learning. IEEE. 2022;6.
- [216] Yang Z, Shi Y, Zhou Y, Wang Z, Yang K. Trustworthy Federated Learning via Blockchain. IEEE Internet of Things Journal. 2022;1-1. Available from: <https://ieeexplore.ieee.org/document/9866512/>.
- [217] Abou El Houda Z, Hafid AS, Khoukhi L, Brik B. When Collaborative Federated Learning Meets Blockchain to Preserve Privacy in Healthcare. IEEE Transactions on Network Science and Engineering. 2022;1-11. Available from: <https://ieeexplore.ieee.org/document/9906419/>.
- [218] Li J, Yan T, Ren P. VFL-R: a novel framework for multi-party in vertical federated learning. Applied Intelligence. 2022 Sep. Available from: <https://link.springer.com/10.1007/s10489-022-04111-0>.
- [219] Sav S, Bossuat JP, Troncoso-Pastoriza JR, Claassen M, Hubaux JP. Privacy-preserving federated neural network learning for disease-associated cell classification. Patterns. 2022 May;3(5):100487. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2666389922000721>.
- [220] Divya Jatain, Vikram Singh, Naveen Dahiya. A contemplative perspective on federated machine learning: Taxonomy, threats & vulnerability assessment and challenges. Journal of King Saud University - Computer and Information Sciences. 2021. Available from: <https://www.sciencedirect.com/science/article/pii/S1319157821001312>.
- [221] Chuanxin Z, Yi S, Degang W. Federated Learning with Gaussian Differential Privacy. In: Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence. RICAI 2020. New York, NY, USA: Association for Computing Machinery; 2020. p. 296-301.

- [222] Jarin I, Eshete B. PRICURE: Privacy-Preserving Collaborative Inference in a Multi-Party Setting. In: Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics. IWSPA '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 25-35.
- [223] Grivet Sébert A, Pinot R, Zuber M, Gouy-Pailler C, Sirdey R. SPEED: secure, PrivatE, and efficient deep learning. *Machine Learning*. 2021;110(4):675-94.
- [224] Owusu-Agyemeng K, Qin Z, Xiong H, Liu Y, Zhuang T, Qin Z. MSDP: multi-scheme privacy-preserving deep learning via differential privacy. *Personal and Ubiquitous Computing*. 2021.
- [225] Nuria Rodríguez-Barroso, Goran Stipcich, Daniel Jiménez-López, José Antonio Ruiz-Millán, Eugenio Martínez-Cámarra, Gerardo González-Seco, et al. Federated Learning and Differential Privacy: Software tools analysis, the Sherpa.ai FL framework and methodological guidelines for preserving data privacy. *Information Fusion*. 2020;64:270-92. Available from: <https://www.sciencedirect.com/science/article/pii/S1566253520303213>.
- [226] Wibawa F, Catak FO, Kuzlu M, Sarp S, Cali U. Homomorphic Encryption and Federated Learning based Privacy-Preserving CNN Training: COVID-19 Detection Use-Case. In: EICC 2022: Procedings of the European Interdisciplinary Cybersecurity Conference. Barcelona Spain: ACM; 2022. p. 85-90. Available from: <https://dl.acm.org/doi/10.1145/3528580.3532845>.
- [227] Feng X, Chen L. Data Privacy Protection Sharing Strategy Based on Consortium Blockchain and Federated Learning. In: 2022 International Conference on Artificial Intelligence and Computer Information Technology (AICT). Yichang, China: IEEE; 2022. p. 1-4. Available from: <https://ieeexplore.ieee.org/document/9930188/>.
- [228] Abuadbba S, Kim K, Kim M, Thapa C, Camtepe SA, Gao Y, et al. Can We Use Split Learning on 1D CNN Models for Privacy Preserving Training? In: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. ASIA CCS '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 305-18.
- [229] Shayan M, Fung C, Yoon CJM, Beschastnikh I. Biscotti: A Blockchain System for Private and Secure Federated Learning. *IEEE Transactions on Parallel and Distributed Systems*. 2021;32(7):1513-25.
- [230] Ghamry ME, Halim ITA, Bahaa-Eldin AM. Secular: A Decentralized Blockchain-based Data Privacy-preserving Model Training Platform. In: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC); 2021. p. 357-63.
- [231] Agarwal A, Chattopadhyay P, Wang L. Privacy preservation through facial de-identification with simultaneous emotion preservation. *Signal, Image and Video Processing*. 2020.
- [232] Zhou T, Shen J, He D, Vijayakumar P, Kumar N. Human-in-the-Loop-Aided Privacy-Preserving Scheme for Smart Healthcare. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2020;1-10.
- [233] Zhou T, Shen J, He D, Vijayakumar P, Kumar N. Human-in-the-Loop-Aided Privacy-Preserving Scheme for Smart Healthcare. *IEEE Transactions on Emerging Topics in Computational Intelligence*, title=Human-in-the-Loop-Aided Privacy-Preserving Scheme for Smart Healthcare. 2020 Jan:1-10.
- [234] Bai Y, Fan M, Li Y, Xie C. Privacy Risk Assessment of Training Data in Machine Learning. In: ICC 2022 - IEEE International Conference on Communications. Seoul, Korea, Republic of: IEEE; 2022. p. 1015-5. Available from: <https://ieeexplore.ieee.org/document/9839062/>.
- [235] Abbasi W, Mori P, Saracino A, Frascola V. Privacy vs Accuracy Trade-Off in Privacy Aware Face Recognition in Smart Systems. In: 2022 IEEE Symposium on Computers and Communications (ISCC). Rhodes, Greece: IEEE; 2022. p. 1-8. Available from: <https://ieeexplore.ieee.org/document/9912465/>.
- [236] Montenegro H, Silva W, Gaudio A, Fredrikson M, Smailagic A, Cardoso JS. Privacy-Preserving Case-Based Explanations: Enabling Visual Interpretability by Protecting Privacy. *IEEE Access*. 2022;10:28333-47. Available from: <https://ieeexplore.ieee.org/document/9729808/>.
- [237] Mao Q, Chen Y, Duan P, Zhang B, Hong Z, Wang B. Privacy-Preserving Classification Scheme Based on Support Vector Machine. *IEEE Systems Journal*. 2022;1-11. Available from: <https://ieeexplore.ieee.org/document/9732431/>.
- [238] Harichandana BSS, Agarwal V, Ghosh S, Ramena G, Kumar S, Raja BRK. PrivPAS: A real time Privacy-Preserving AI System and applied ethics. In: 2022 IEEE 16th International Conference on Semantic Computing (ICSC). Laguna Hills, CA, USA: IEEE; 2022. p. 9-16. Available from: <https://ieeexplore.ieee.org/document/9736272/>.
- [239] Tian H, Zeng C, Ren Z, Chai D, Zhang J, Chen K, et al. Sphinx: Enabling Privacy-Preserving Online Learning over the Cloud. In: 2022 IEEE Symposium on Security and Privacy (SP). San Francisco,

- CA, USA: IEEE; 2022. p. 2487-501. Available from: <https://ieeexplore.ieee.org/document/9833648/>.
- [240] Tan AZ, Yu H, Cui L, Yang Q. Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*. 2022;1-17. Available from: <https://ieeexplore.ieee.org/document/9743558/>.
- [241] Gambelin O. Brave: what it means to be an AI Ethicist. *AI and Ethics*. 2021;1(1):87-91.
- [242] Gill KS. Ethical dilemmas // Ethical dilemmas: Ned Ludd and the ethical machine. *AI & SOCIETY*. 2021;36(3):669-76.
- [243] Stahl BC. Ethical Issues of AI. In: Stahl BC, editor. *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Cham: Springer International Publishing; 2021. p. 35-53.
- [244] Charles D Raab. Information privacy, impact assessment, and the place of ethics*. *Computer Law & Security Review*. 2020;37:105404. Available from: <https://www.sciencedirect.com/science/article/pii/S0267364920300091>.
- [245] Prunkl C, Whittlestone J. Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 138-43.
- [246] Ehsan U, Wintersberger P, Liao QV, Mara M, Streit M, Wachter S, et al. Operationalizing Human-Centered Perspectives in Explainable AI. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21. New York, NY, USA: Association for Computing Machinery; 2021. .
- [247] Alexandre Heuillet, Fabien Couthouis, Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*. 2021;214:106685. Available from: <https://www.sciencedirect.com/science/article/pii/S0950705120308145>.
- [248] Ztyek A, Liu D, Vaithianathan R, Veeramachaneni K. Sibyl: Explaining Machine Learning Models for High-Stakes Decision Making. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21. New York, NY, USA: Association for Computing Machinery; 2021. .
- [249] Sokol K, Flach P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 56-67.
- [250] Hailemariam Y, Yazdinejad A, Parizi RM, Srivastava G, Dehghantanha A. An Empirical Evaluation of AI Deep Explainable Tools. In: 2020 IEEE Globecom Workshops (GC Wkshps); 2020. p. 1-6.
- [251] Colaner N. Is explainable artificial intelligence intrinsically valuable? *AI & SOCIETY*. 2021.
- [252] Mercier D, Lucieri A, Munir M, Dengel A, Sheraz A. Evaluating Privacy-Preserving Machine Learning in Critical Infrastructures: A Case Study on Time-Series Classification. *IEEE Transactions on Industrial Informatics*. 2021;1-1. Conference Name: IEEE Transactions on Industrial Informatics.
- [253] Biswas S, Khare N, Agrawal P, Jain P. Machine learning concepts for correlated Big Data privacy. *Journal of Big Data*. 2021 Dec;8(1):157. Available from: <https://doi.org/10.1186/s40537-021-00530-x>.
- [254] Chang H, Shokri R. On the Privacy Risks of Algorithmic Fairness. In: 2021 IEEE European Symposium on Security and Privacy (EuroS P); 2021. p. 292-303.
- [255] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*. 2021;115:619-40. Available from: <https://www.sciencedirect.com/science/article/pii/S0167739X20329848>.
- [256] Zhang K, Yiu SM, Hui LCK. A Light-Weight Crowdsourcing Aggregation in Privacy-Preserving Federated Learning System. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020. p. 1-8. ISSN: 2161-4407.
- [257] Aminifar A, Rabbi F, Pun KI, Lamo Y. Privacy Preserving Distributed Extremely Randomized Trees. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing. SAC '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 1102-5.
- [258] Anastasiia Girka, Vagan Terziyan, Maria Gavrilushenko, Andrii Gontarenko. Anonymization as homeomorphic data space transformation for privacy-preserving deep learning. *Procedia Computer Science*. 2021;180:867-76. Available from: <https://www.sciencedirect.com/science/article/pii/S1877050921003914>.

- [259] He Q, Yang W, Chen B, Geng Y, Huang L. TransNet: Training Privacy-Preserving Neural Network over Transformed Layer. *Proc VLDB Endow.* 2020;13(12):1849-62.
- [260] Goldsteen A, Ezov G, Shmelkin R, Moffie M, Farkash A. Data minimization for GDPR compliance in machine learning models. *AI and Ethics.* 2021.
- [261] Boenisch F, Battis V, Buchmann N, Poikela M. “I Never Thought About Securing My Machine Learning Systems”: A Study of Security and Privacy Awareness of Machine Learning Practitioners. In: *Mensch Und Computer 2021. MuC ’21.* New York, NY, USA: Association for Computing Machinery; 2021. p. 520-46.
- [262] Xianxian Li, Jing Liu, Songfeng Liu, Jinyan Wang. Differentially private ensemble learning for classification. *Neurocomputing.* 2021;430:34-46. Available from: <https://www.sciencedirect.com/science/article/pii/S0925231220319512>.
- [263] Yuan H, Tang J, Hu X, Ji S. XGNN: Towards Model-Level Explanations of Graph Neural Networks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD ’20.* New York, NY, USA: Association for Computing Machinery; 2020. p. 430-8.