TOWARDS A DEFINITION OF A RESPONSIBLE ARTIFICIAL INTELLIGENCE

SABRINA GÖLLNER,[1] MARINA TROPMANN-FRICK,[1]
BOŠTJAN BRUMEN[2]

[1] Hamburg University of Applied Sciences, Department of Computer Science, Hamburg, Germany.
sabrina.goellner@haw-hamburg.de, marina.tropmann-frick@haw-hamburg.de
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia.
bostjan.brumen@um.si

Our research aims to contribute to the concept of responsible artificial intelligence (AI), a topic under significant discussion in EU politics, further emphasized by recent EU publications. Primarily, AI, while beneficial, can be a potential weapon, necessitating responsible use and prevention against misuse or misalignment. In recognizing the critical role of AI research in aiding legislators and machine learning practitioners, our work aims to help prepare for future AI advancements. To the best of our knowledge, we establish the first unified definition of responsible AI. As part of a structured literature review, we clarify the current state of the art in the context of responsible AI. Based on the knowledge of the analysis part we also have discussed an approach for developing a future framework for responsible AI. The results demonstrate that responsible AI should be a human-centered approach, encompassing ethical considerations, explainability of models, privacy, security, and trust.

Keywords:
structured literature review,
artificial intelligence,
responsible AI,
privacy-preserving AI,
explainable AI,
ethical AI,
trustworthy AI

University of Maribor Press

DOI https://doi.org/10.18690/um.feri.5.2023.2
ISBN 978-961-286-745-4

1 Introduction

Over the years, significant research has been conducted to enhance Artificial Intelligence (AI), which is already widely used in various life and industry sectors. In 2020 and 2021, the European Commission published a series of papers [1,2,3] outlining their strategy for AI. The white paper "A European Approach to Excellence and Trust" from 2020 outlines political strategies to encourage the use of AI while reducing the potential risks associated with certain applications of this technology. 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 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 "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 "responsible AI", which reflect 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 at 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 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 the first detailed and structured review regarding responsible AI.

The paper is organized as follows: First, we explain our research methodology, including our research aims and objectives, and the databases and research queries we used for searching. Next, we analyze the existing definitions for responsible AI in the literature, along with related expressions and their definitions. We compare these definitions to determine the essence of responsible AI. We then summarize our key findings within the previously defined scopes of responsible AI, conducting both qualitative and quantitative analyses. In the discussion section, we outline the key points and pillars for developing responsible AI. Finally, we conclude by mentioning the limitations of our work and discussing future research.

2 Research Methodology

In order to address the research questions, we conducted a systematic literature review (SLR) using the guidelines outlined in [4]. The process of performing the structured literature review for our study is explained in detail in the following subsections.

2.1 Research Aims and Objectives

Our research focuses on exploring the different aspects of "Responsible AI" including privacy, explainability, trust, and ethics. Our objectives are to define the term "responsible AI", examine the current state of research in this field, and identify areas that require further investigation. Ultimately, we aim to uncover any challenges, opportunities, and open problems that exist in this area.

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.

To ensure that all state-of-the-art papers were included, the search was limited to a three-year period from 2020 to 2022, with the search conducted in December 2022. The search results were sorted by relevance to eliminate non-relevant papers, as some search engines lack advanced options. During the screening stage, the authors followed specific guidelines to exclude irrelevant papers. 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).

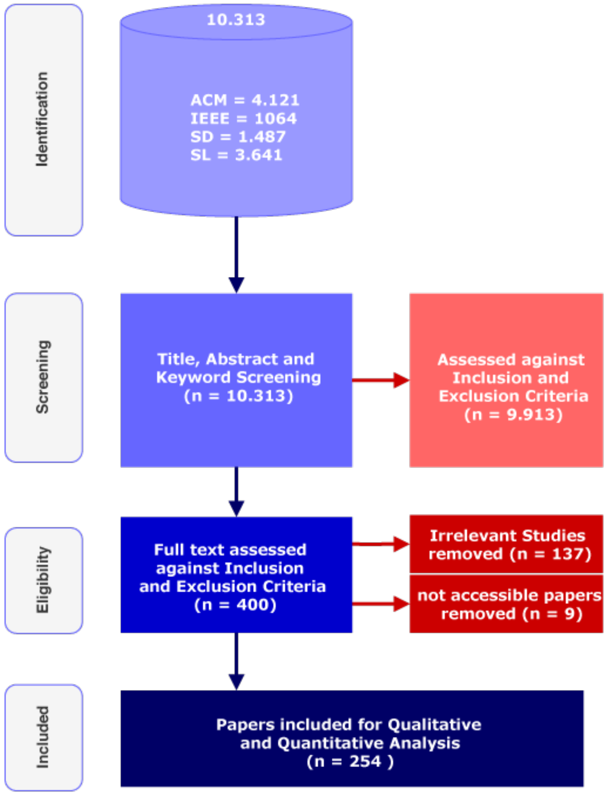


Figure 1: Structured review flow chart: the Preferred Reporting Items for Systematic Reviews and Meta– Analyses (PRISMA) flow chart detailing the records identified and screened, the number of full-text articles retrieved and assessed for eligibility, and the number of studies included in the review.

Source: own.

3 Analysis

In this section, we analyze existing definitions of "responsible AI" in literature. We also examine content-wise-similar expressions and their definitions, comparing and searching for any 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

Upon careful examination of 254 papers, it was found that a mere 5 of them specifically address the definition of "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

Our analysis uncovered that the terms "Responsible AI," "Ethical AI," and "Trustworthy AI" are frequently utilized interchangeably. Furthermore, we determined that "Human-Centered AI" holds a similar significance.

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.

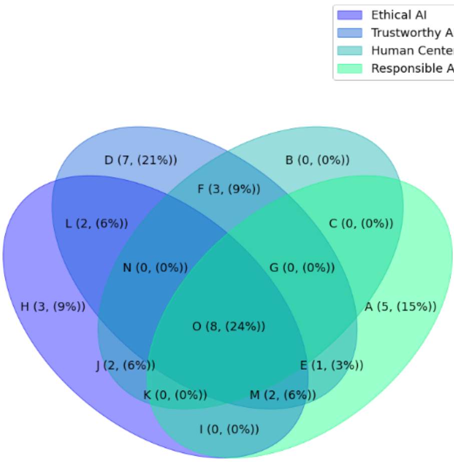
3.2 Collection of definitions

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:

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.
- 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 since 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.



Set	Terms
A	Solidarity, Performance, Sustainability, Soundness, Inclusiveness
B	-
C	-
D	Equality, Usability, Accuracy under Uncertainty, Assessment, Reliability, Data Control, Data Minimization Reproducibility, Generalization User Acceptance
E	Social Good
F	Human-Centered, Human Control, Human Agency
G	-
H	Autonomy, Non-Maleficience, Trust
I	-
J	Human Values, Non-Discrimination
K	-
L	Compliant with Rules and Laws, Social Robustness
M	Human Autonomy, Dignity
N	-
O	Explainability, Safety, Fairness, Accountability, Ethics, Security Privacy, Transparency

Figure 2: Venn diagram

Source: own.

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:

DEFINITION OF RESPONSIBLE AI

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.

Figure 3: Definition of responsible AI

Source: own.

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

After analyzing the literature, we have identified six categories related to responsible AI in section 3. These categories are Human-centered, Trustworthy, Ethical, Explainable, Privacy-preserving, and Secure AI. Adhering to these categories will ensure the responsible development and use of AI.

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 254 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 topics. The detailed analysis of these papers is beyond the scope of the present work and will be presented in our future work. Nevertheless, we highlight their most important features 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]

In summary, Trustworthy AI aims to provide the benefits of AI while addressing scenarios that have significant implications for people and society. To be accepted in society, it is crucial for AI applications to prioritize trust as a key goal and make every effort to maintain and measure it throughout all stages of development. Despite this importance, achieving trustworthy AI remains a significant challenge as it has not yet been comprehensively addressed.

3.3.2 Ethical AI

In this section, we will outline the discoveries made in the realm of ethical AI. The most fitting explanation of ethics in relation to AI is the one provided in source [28]:

”*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* [28].”

During our analysis, we noticed that Ethical AI deals often with fairness. 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 [29]. In the state-of-the-art overview of [30] 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 subsection.

3.3.3 Explainable AI

The choices made by AI systems or humans utilizing AI can greatly affect the welfare, liberties, and prospects of those influenced by those choices. That's why the issue of AI explainability is a crucial ethical concern. This subsection deals with the analysis of the literature in the field of 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]

Numerous XAI techniques have been extensively discussed in literature. The authors of [6] as well as [31] 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-dependent and data-independent mechanisms for gaining interpretability as well as global and local interpretability methods.

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 [32] the effectiveness and interpretability of two existing tools were investigated; the 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) [33,34]. Therefore, we need to protect data through secure and privacy-preserving AI-methods, which are analyzed in the following section.

3.3.4 Privacy-preserving and Secure AI

As previously mentioned, trust in AI is dependent on privacy and security. However, the success of ML models relies heavily on data, including sensitive information. This has resulted in increasing worries about privacy violations, such as the unlawful use and exposure of private data [35,36]. To ensure complete privacy protection, we require holistic methods that consider the usage of data and user activities and transactions.[37].

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.

Although the AI field is undergoing extensive research into privacy and security, achieving flawless privacy preservation and security in AI is currently not possible. Nonetheless, several challenges require addressing to further advance in this area.

Table 1: Quantitative Analysis

Feature of a study	Representation	Percentage	Sources
Trustworthy AI (28/254, 11%) *			
Reviews and Surveys	9/28	32%	[11,17,38,13,39,14,40,41,42]
Perceptions of trust	4/28	14%	[43,44,45,27]
Frameworks	9/28	32%	[26,46,47,48,49,15,50,51,52]
Miscellaneous	6/28	28%	[53,54,55,56,16,57]
Ethical AI (85/254,34%) *			
Frameworks	19/85	22%	[35,58,59,7,20,60,29,24,61,62]
			[63,64,65,66,67,68,69,70,71]
Ethical issues	22/85	26%	[72,20,73,74,75,76,77,78]
			[79,80,81,28,82,36,83,84]
			[85,86,87,88,89,90]
Miscellaneous	33/85	39%	[91,19,92,93,94,95,96,22,21,97,98]
			[99,100,101,102,9,103,104]
			[105,106,107,108,109,110,111]
			[112,113,114,115,116,117,118,8]
Reviews and Surveys	10/85	12%	[119,120,121,122,123,124,125,126,127,30]
Tools	1/85	1%	[128]
Explainable AI (46/254 , 18%) *			
Reviews and Surveys	10/46	22%	[6,31,33,12,129,34]
			[130,131,132,133]
Stakeholders	7/46	15%	[134,135,136,137]
			[32,138,139]
XAI Approaches	14/46	30%	[140,5,141,142,143,144]
			[145,146,147,148,149,150,151,152]
Frameworks	4/46	9%	[153,154,155,156]
Miscellaneous	11/46	24%	[157,158,159,160,161]
			[162,163,164,165,166,167]
Privacy-preserving and Secure AI (95/254 , 38%) *			
Reviews and Surveys	10/95	10%	[168,169,170,171,172,37]
			[173,174,175,176]
Differential Privacy	12/95	13%	[177,178,179,180,181,182]
			[183,184,185,186,187,188]
Secure Multi-Party Computation	2/95	2%	[189,190]
Homomorphic Encryption	4/95	4%	[142,191,192,193]
Federated learning	35/95	37%	[194,195,196,197,198,199,200,201]
			[202,203,204,205,206]
			[207,208,209,210,211,212,213,214,215]
			[216,217,218,219,220,221,222]
			[223,224,225,226,227,228,229]
Hybrid Approaches	8/95	xx%	[230,231,232,233,234,235,236,237]
Security Threats	7/95	8%	[238,239,240,241,242,243,244]
Miscellaneous	16/95	17%	[245,246,247,248,249,250,251,252,253,254]
			[255,256,257,258,259,260]

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

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 the 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 the explainability of AI models. Few of them presented miscellaneous topics that could not be assigned to any specific category or framework 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 that presented frameworks especially for trustworthiness or papers that reported on how Trust is perceived and described by different users.

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 ”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.

4.1 Pillars of Responsible AI

Here we highlight the most important criteria that a responsible AI should fulfil. These are also the points that a developer should consider if she 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,
- 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),
- 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).

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 functional 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.

Therefore, 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 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.

Each pillar of ethics includes "sub-modules" such as accountability, fairness, sustainability, and compliance. These are essential to ensure that AI meets ethical standards.

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.

Every "responsible system" requires humans to care for it. These individuals must handle the system responsibly, conducting maintenance work and regularly checking metrics to ensure that their responsibilities are fulfilled. To achieve this, special metrics are used as a continuous check. This makes responsible AI a joint effort between the system-side and the developer-side.

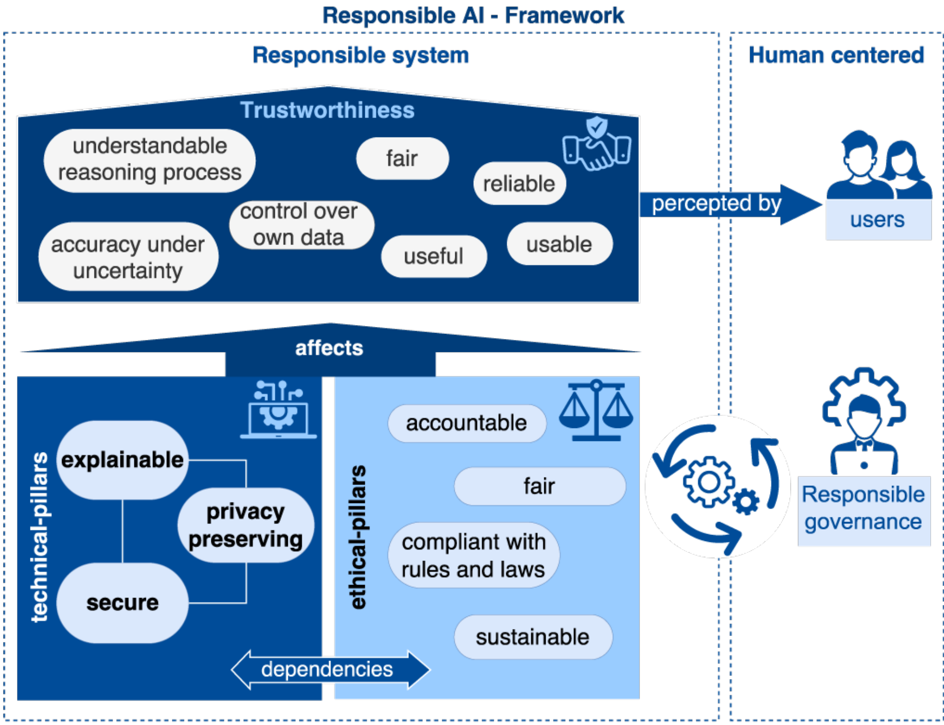


Figure 4: Pillars of the Responsible AI framework
Source: own.

In section 3.3, the concept of Human-Centered AI is highlighted as a crucial aspect of responsible AI. It is closely linked to the "Human-in-the-loop" approach, which emphasizes the importance of human involvement in the development and use of AI. This approach allows for the detection and correction of errors and retraining of the system throughout its lifespan, ensuring that AI is designed and utilized for the benefit of humans.

Therefore, responsible AI is interdisciplinary, and it is not 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 the cost of data privacy. For example, the methods that make the model more robust against attacks or methods that try to explain a model's behavior and could leak some information. Managing AI systems that are accurate, fair, private, robust, and explainable simultaneously is a challenging task. To begin, we suggest creating a benchmark for each requirement, which will determine the extent to which each requirement is met.

5 Research Limitations

Our study aims to provide a thorough and detailed analysis of the available literature on responsible AI from various journals. However, we encountered limitations in accessing some journals that were not freely available despite extensive access provided by our institutions. Despite our best efforts, accessibility remained an issue. It is also possible that some relevant research publications were not included in the databases we used for our search. Furthermore, our study only included the most recent state-of-the-art research, which may have caused us to miss out on some older but still relevant developments.

Another limitation of the presented work is the missing in-depth analysis of the papers reviewed. Due to paper length constraints, we have omitted a detailed overview of each of the reviewed papers' contributions in each of the subsections of section 3.3.

6 Conclusion

The field of AI is rapidly evolving and a legal framework is necessary to ensure responsible practices. However, the terms "trustworthy AI" and "responsible AI" lack clear definitions, making it difficult to establish efficient regulations. Instead of focusing solely on trust, regulations for responsible AI must be defined. As a leading authority in setting standards, such as the GDPR, the EU should be informed and prepared for upcoming research and legal regulations. This research provides an important contribution to the concept of responsible AI, being the first to address it comprehensively through a structured literature review and presenting an overarching definition. The review analyzed 254 recent high-quality works on the topic, and included a qualitative analysis of the papers covered.

We have defined the concept of "responsible AI" and conducted a thorough analysis of its key components. These components include human-centered design, trustworthy development, ethical considerations, explainability, privacy preservation, and security. By prioritizing these aspects, we can ensure the responsible development and use of AI products, and establish legal frameworks to regulate their use. In the discussion section, we propose a framework for responsible AI based on the insights gained from our analysis. In future research, we plan to analyze individual papers to determine their contributions to responsible AI, and explore topics such as human-centered AI and "human-in-the-loop" approaches. We also aim to develop benchmarking methods for responsible AI and establish a holistic framework to guide responsible AI development.

References

A complete list of 260 references is available at https://drive.google.com/file/ d/1Fm-9hKkrY_YAzS02TWec2L3lIqgPSmqm/view?usp=sharing, or by scanning the QR code below.

Figure 4: QR Code with the list of references
Source: own.

[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] Commission E. 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 D´ıaz-Rodr´ıguez, 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, Buc¸inca 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] level expert group on artificial intelligence H. 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] Ibánez 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, Achten N, Hilligoss H, Nagy A, Srikumar 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 GPDR 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] Hanna R, Kazim E. Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach. AI and Ethics. 2021.
[29] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, Pekka Abrahamsson. ECCOLA — 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.
[30] Zhou J, Chen F, Holzinger A. Towards Explainability for AI Fairness. In: Holzinger A, Goebel R, Fong R, Moon T, Mu¨ller KR, Samek W, editors. xx AI - 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.
[31] Burkart N, Huber MF. A Survey on the Explainability of Supervised Machine Learning. J Artif Int Res. 2021;70:245-317.
[32] 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.
[33] Choraś 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, Za´vodszky 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.
[34] 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.
[35] Cheng L, Varshney KR, Liu H. Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. J Artif Int Res. 2021;71:1137-81.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] 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.

[41] Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multicentre 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.
[42] 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/.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] 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.
[51] Thuraisingham B. Trustworthy Machine Learning. IEEE Intelligent Systems. 2022 Jan;37(1):21-4. Available from: https://ieeexplore.ieee.org/document/9756264/.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] 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.

[57] 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/.
[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] 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.
[62] 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/.
[63] 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/s43681022-00162-8.
[64] 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.
[65] 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.
[66] 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.
[67] 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.
[68] 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.
[69] 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.
[70] 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.
[71] 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.
[72] Ayling J, Chapman A. Putting AI ethics to work: are the tools fit for purpose? AI and Ethics. 2021.
[73] Maclure J. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. Minds and Machines. 2021;31(3):421-38.
[74] Gambelin O. Brave: what it means to be an AI Ethicist. AI and Ethics. 2021;1(1):87-91.
[75] 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.

[76] Gill KS. Ethical dilemmas // Ethical dilemmas: Ned Ludd and the ethical machine. AI & SOCIETY. 2021;36(3):669-76.
[77] 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.
[78] 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.
[79] Rochel J, Evéquoz F. Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics. AI & SOCIETY. 2020.
[80] 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.
[81] Stahl BC, Antoniou J, Ryan M, Macnish K, Jiya T. Organisational responses to the ethical issues of artificial intelligence. AI & SOCIETY. 2021.
[82] 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.
[83] Petrozzino C. Who pays for ethical debt in AI? AI and Ethics. 2021.
[84] 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.
[85] 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.
[86] 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.
[87] 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.
[88] 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.
[89] 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.
[90] Munn L. The uselessness of AI ethics. AI and Ethics. 2022 Aug. Available from: https://link.springer.com/10.1007/s43681-022-00209-w.
[91] Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines. 2020;30(1):99-120.
[92] Kiemde SMA, Kora AD. Towards an ethics of AI in Africa: rule of education. AI and Ethics. 2021.
[93] 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.
[94] 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.

[95] 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.
[96] Forbes K. Opening the path to ethics in artificial intelligence. AI and Ethics. 2021.
[97] 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.
[98] 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.
[99] 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.
[100] 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.
[101] 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.
[102] 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.
[103] 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.
[104] 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.
[105] 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.
[106] 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.
[107] 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.
[108] 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.
[109] Bélisle-Pipon JC. Artificial intelligence ethics has a black box problem. AI and Society. 2022:16.
[110] Haüßermann 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.

[111] 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.
[112] 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.
[113] 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.
[114] 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.
[115] 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.
[116] Ramanayake R. Immune moral models? Pro-social rule breaking as a moral enhancement approach for ethical AI. AI & SOCIETY. 2022:13.
[117] Hunkenschroer 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.
[118] 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.
[119] 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.
[120] 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.
[121] 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/.
[122] 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/.
[123] 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/.
[124] 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.
[125] 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.
[126] Attard-Frost B. The ethics of AI business practices: a review of 47 AI ethics guidelines. AI and Ethics. 2022:18.
[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] 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.

[132] 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/.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

[141] Sokol K, Flach P. One Explanation Does Not Fit All. KI - Künstliche Intelligenz. 2020;34(2):235-50.

[142] 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.

[143] Zytek 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. .
[144] 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.
[145] 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.
[146] 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.
[147] 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.
[148] 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.
[149] 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/.
[150] 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/.
[151] 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.
[152] Rožanec JM, Fortuna B, Mladenić 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.
[153] 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).
[154] 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.
[155] 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.
[156] 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/10.1145/3506860.3506866.
[157] 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.
[158] Colaner N. Is explainable artificial intelligence intrinsically valuable? AI & SOCIETY. 2021.

[159] 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.

[160] Tsiakas K, Murray-Rust D. Using human-in-the-loop and explainable AI to envisage new future work practices. In: The15th 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.

[161] 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.

[162] 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/.

[163] 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/.

[164] 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/.

[165] 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.

[166] 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.

[167] 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.

[168] 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.

[169] 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.

[170] 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.

[171] 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.

[172] 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.

[173] Sergey Zapechnikov. 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.

[174] 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).

[175] 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.
[176] 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/s10462-021-10123-y.
[177] 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.
[178] 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.
[179] 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.
[180] Guevara M, Desfontaines D, Waldo J, Coatta T. Differential Privacy: The Pursuit of Protections by Default. Commun ACM. 2021;64(2):36-43.
[181] 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.
[182] Alishahi M, Moghtadaiee 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.
[183] 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/.
[184] 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/.
[185] 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.
[186] 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.
[187] 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.
[188] 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.
[189] 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. Neurocomputing. 2021;422:245-62. Available from: https://www.sciencedirect.com/science/article/pii/S0925231220315095.
[190] 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.
[191] 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.

[192] 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/.

[193] 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.

[194] Can YS, Ersoy C. Privacy-Preserving Federated Deep Learning for Wearable IoT-Based Biomedical Monitoring. ACM Trans Internet Technol. 2021;21(1).

[195] 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.

[196] 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.

[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] 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.

[199] 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.

[200] 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.

[201] 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.

[202] 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.

[203] 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.

[204] 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/.

[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] 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.

[207] 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.
[208] 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.
[209] 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.
[210] 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.
[211] 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.
[212] 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.
[213] 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.
[214] 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.
[215] 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.
[216] Antunes RS, Andre´ 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.
[217] 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.
[218] 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.
[219] 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.
[220] 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.

[221] 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/.

[222] 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/.

[223] Gholami A, Torkzaban N, Baras JS. Trusted Decentralized Federated Learning. IEEE. 2022:6.

[224] 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/.

[225] 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/.

[226] Chowdhury A, Kassem H, Padoy 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.

[227] 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.

[228] 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/.

[229] 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.

[230] 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.

[231] 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.

[232] 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.

[233] 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.

[234] Nuria Rodríguez-Barroso, Goran Stipcich, Daniel Jiménez-López, José Antonio Ruiz-Millán, Eugenio Martínez-Cámara, 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.

[235] 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: Proccedings of the European Interdisciplinary Cybersecurity Conference. Barcelona Spain: ACM; 2022. p. 85-90. Available from: https://dl.acm.org/doi/10.1145/3528580.3532845.

[236] 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 (AICIT). Yichang, China: IEEE; 2022. p. 1-4. Available from: https://ieeexplore.ieee.org/document/9930188/.
[237] 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/.
[238] 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.
[239] 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.
[240] 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.
[241] 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.
[242] Brown H, Lee K, Mireshghallah F, Shokri R, Tramér 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.
[243] 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/.
[244] Giordano M, Maddalena L, Manzo M, Guarracino 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.
[245] Agarwal A, Chattopadhyay P, Wang L. Privacy preservation through facial de-identification with simultaneous emotion preservation. Signal, Image and Video Processing. 2020.
[246] 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.
[247] Anastasiia Girka, Vagan Terziyan, Mariia Gavriushenko, 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.
[248] 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.
[249] 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.
[250] Goldsteen A, Ezov G, Shmelkin R, Moffie M, Farkash A. Data minimization for GDPR compliance in machine learning models. AI and Ethics. 2021.
[251] 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.
[252] 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.
[253] 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.
[254] 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.
[255] 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/.
[256] Abbasi W, Mori P, Saracino A, Frascolla 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/.
[257] 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/.
[258] 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/.
[259] 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/.
[260] 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/.