Information Modelling and Knowledge Bases XXXV
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA231146

Towards a Definition of a Responsible Artificial Intelligence

Sabrina GÖLLNER [a], Marina TROPMANN-FRICK [a] and Boštjan BRUMEN [b]
[a] *Department of Computer Science, Hamburg University of Applied Sciences, Germany.*
sabrina.goellner@haw-hamburg.de, marina.tropmann-frick@haw-hamburg.de
[b] *Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia.*
bostjan.brumen@um.si
ORCiD ID: Sabrina Göllner https://orcid.org/0000-0002-1817-7440, Marina Tropmann-Frick https://orcid.org/0000-0003-1623-5309, Boštjan Brumen https://orcid.org/0000-0002-0560-1230

Abstract. Our investigation seeks to enhance the understanding of responsible artificial intelligence. The EU is deeply engaged in discussions concerning AI trustworthiness and has released several relevant documents. It's crucial to remember that while AI offers immense benefits, it also poses risks, necessitating global oversight. Moreover, there's a need for a framework that helps enterprises align their AI development with these international standards. This research will aid both policymakers and AI developers in anticipating future challenges and prioritizing their efforts. In our study, we delve into the essence of responsible AI and, to our understanding, introduce a comprehensive definition of the term. Through a thorough literature review, we pinpoint the prevailing trends surrounding responsible AI. Using insights from our analysis, we've also deliberated on a prospective framework for responsible AI. Our findings emphasize that human-centeredness should prioritized. This entails adopting AI techniques that prioritize ethical considerations, explainability of models, and aspects like privacy, security, and trustworthiness.

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

1. Introduction

Over the past few years, intensified research has been directed towards advancing Artificial Intelligence (AI), reflecting its increasing integration into various sectors. The European Commission, in 2020 and 2021, unveiled several documents, notably [1,2,3], outlining their strategic vision for AI. Their 2020 white paper "A European Approach to Excellence and Trust" proposes political strategies to amplify AI's potential benefits and mitigate its inherent risks. Their objective is to create a legal infrastructure in Europe, promoting trustworthy AI, anchored firmly in the core values and rights of EU citizens. Emphasizing a human-centric AI approach, the paramount importance of European values is highlighted. These documents tackle intricate challenges like ethics, privacy, and sustainability, underscoring the security's pivotal role within AI domains and introducing

a stratified risk framework for AI systems.

An observation from the documents is the current absence of a unified European AI framework, shedding light on its political significance. Another crucial document, the "Communication on Fostering a European Approach to AI," sets out the EU Commission's roadmap. It underscores a *"human-centric, sustainable, secure, inclusive and trustworthy artificial intelligence (AI) [which] depends on the ability of the European Union"*.

The Commission's ambition revolves around fostering AI excellence, emphasizing collaboration, research augmentation, and funding prospects. The discourse on trust spirals towards innovation, with the EU's approach characterized as *"human-centered, risk-based, proportionate, and dynamic"*. The vision encapsulated is for an innovative, ethical, and human-focused AI. The document wraps up by highlighting the opportunity to support the EU's innovative prowess, competitiveness, and responsible AI deployment.

Moreover, the European Commission's "Proposal for a Regulation" details potential prohibitions in AI practices and stipulations for high-risk AI systems, with an emphasis on transparency. Notably, there's a discernible inconsistency in the terminologies used across political texts related to trustworthy AI, often leading to ambiguity. While the documents accentuate both the promise and perils of AI, they avoid a clear definition of trustworthy AI. The discourse touches upon aspects like ethical considerations, transparency, and safety, but without offering an unequivocal definition.

Our contention is that merely targeting trust, as vaguely outlined in these documents, doesn't suffice for AI integration. A broader "responsible AI" approach, resonating with European values, is imperative, where trust forms just an element of its broader responsibility. Consequently, in this paper, our quest is to discern the current academic consensus on *"trustworthy AI"* and probe if there's a definition for *"responsible AI"*. This understanding is crucial for steering towards *"excellence"* in AI.

In our endeavor to decipher responsible AI, we embark on a structured literature review, aiming to unveil its true essence. Our initial probe reveals a plethora of inconsistencies in terminologies not just in political texts but across various sources. Definitional overlaps for responsible AI, coupled with semantically similar terms, further complicate the landscape. Although paradigms exist in arenas like ethics and security, myriad challenges loom ahead. Best to our knowledge this is the first detailed and structured review about responsible AI. The structure of our paper is as follows: Initially, our research methodology, encompassing aims, objectives, as well as specifying the databases and research queries we used for searching, is detailed. We then go through extant literature to cull out definitions of responsible AI, compare them against analogous terms. From this analysis, we derive a definitive understanding of responsible AI. The ensuing sections encapsulate our core findings, underpinned by a meticulous analysis of every single paper regarding the terms "Trustworthy, Ethics, Explainability, Privacy, and Security" in a structured table and quantitative analysis of the study features. Culminating the paper, our discussion highlights the foundational pillars for nurturing responsible AI, and we conclude with our research constraints, inferences, and avenues for future research.

2. Research Methodology

In addressing our research queries, we conducted a systematic literature review (SLR) adhering to the guidelines delineated in [4]. The methodology and steps involved in our comprehensive literature review are expounded upon in the ensuing subsections, with a concise outline presented in the Systematic Review Protocol.

2.1. Research Aims and Objectives

In our current study, our objective is to delve into the multifaceted role of ”Responsible AI” encompassing diverse facets like privacy, explainability, trust, and ethics. Our primary goal is to decipher the composite components that make up ”responsible AI”. Subsequently, we intend to survey the prevailing advancements in this domain. Concludingly, our focus will shift to pinpointing unresolved issues, potential challenges, and arenas that demand deeper investigative efforts.

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 scour the various databases for relevant literature, we utilized the following search string: `("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Neural Network" OR "AI" OR "ML") AND (Ethic* OR Explain* OR Trust*) AND (Privacy*).`
Acknowledging the varied terminology often associated with ”Artificial Intelligence”,

we incorporated terms like "Machine Learning", "Deep Learning", and "Neural Network", viewing them as synonymous. Given the prevalent use of the acronyms AI and ML in many existing papers, these too were integrated into our synonym set. We used the wildcard asterisk (*) with terms like "Ethic", "Trust", "Explain", and "Privacy" to ensure all potential variations stemming from these root words were captured (for instance, explain would match "explainability"). Boolean operators, namely OR and AND, facilitated our search strategy. The OR operator allowed for inclusiveness of any terms, while the AND operator ensured all our specified categories intersected. Parentheses were used to demarcate these sets.

We focused our search on literature from 2020 and 2021, offering a snapshot of the most recent advancements. This search was executed in December 2021. Upon retrieval, results were ranked by relevance. This prioritization was essential, especially since certain databases, lacking refined search capabilities, yielded a multitude of unrelated documents. To exclude irrelevant papers, the authors followed a set of guidelines during the screening stage. Papers did not pass the screening if:

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

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

The initial search produced 10.313 papers of which 4.121 were retrieved from ACM, 1064 from IEEE, 1.487 from Elsevier Science Direct, and 3.641 from Springer Link. The screening using the title, abstract, and keywords removed 6.507 papers. During the check of the remaining papers for eligibility, we excluded 77 irrelevant studies and 9 inaccessible papers. We ended up with 254 papers that we included for the qualitative and quantitative analysis (see Figure 1).

3. Analysis

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

3.1. Responsible AI

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

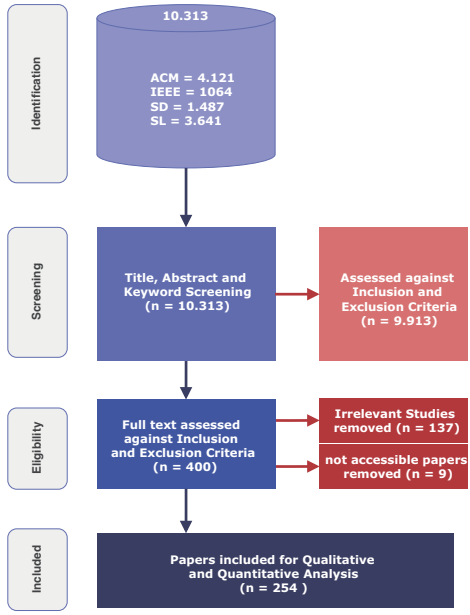


Figure 1. Structured review flow chart: the Preferred Reporting Items for Systematic Reviews and Meta–Analyses (PRISMA) flow chart detailing the records identified and screened, the number of full-text articles retrieved and assessed for eligibility, and the number of studies included in the review.

3.1.1. Terms defining Responsible AI

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

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

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

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

Therefore we summarize these terms of the above definitions to: ”Ethics, Trustworthiness, Security, Privacy, and Explainability”. However, only the terms alone are not

enough to get a picture of responsible AI. Therefore, we will analyze and discuss what the *meaning* of the five terms "Ethics, Trustworthiness, Security, Privacy, and Explainability" in the context of AI is, and how they *depend* on each other. During the analysis, we found also content-wise similar expressions to the concept of "responsible AI" which we want to include in the findings. This topic will be dealt with in the next section.

3.1.2. Content-wise similar expressions for Responsible AI

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

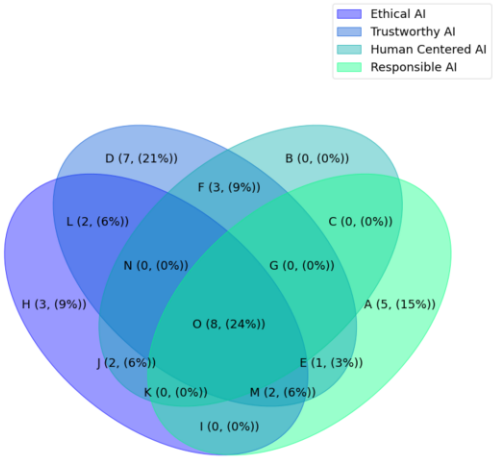
Therefore, we treat the terms:

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

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

3.2. Collection of definitions

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



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

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

- From all four sets there is an overlap of 24% of the terms: Explainability, Safety, Fairness, Accountability, Ethics, Security Privacy, Transparency.
- The terms occurring in the set of the definition for 'trust' only occurred in these, which is why this makes up the second largest set in the diagram. This is due to the fact that most of the terms actually come from definitions for trustworthy AI.
- There are also 6 null sets.

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

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

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

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.

As mentioned in the sections before, the terms defining "responsible AI" result from the analysis of the terms in sections 3.1.1 and 3.1.2. We presented a figure depicting the overlapping of the terms of content-wise similar expressions of Responsible AI, namely "Ethical AI, Trustworthy AI, and Human-Centered AI", and extracted the main terms of it. Also by summarizing the terms Fairness and Accountability into Ethics, and clarifying the synonyms (e.g., explainability instead of transparency), we finally redefined the terms defining "responsible AI" as **"Human-centered, Trustworthy, Ethical, Explainable, Privacy(-preserving) and Secure AI"**.

3.3. Aspects of Responsible AI

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

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

Trustworthy AI is about delivering the promise of AI's benefits while addressing the scenarios that have vital consequences for people and society.
In this subsection, we summarize which are the aspects covered by the papers in the category "Trustworthy AI" and what are the issues to engender users' trust in AI.

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

3.3.2. Ethical AI

In this subsection, we list the findings in the field of ethical AI. In our opinion, the definition found in [28] best describes ethics in conjunction with AI:

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

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

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

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

There are many different XAI techniques discussed in the literature. [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.

We will analyze this in the next section.

3.3.4. Privacy-preserving and Secure AI

As it was noted before, privacy and security are seen as central aspects of building trust in AI. However, the fuel for the good performance of ML models is data, especially sen-

sitive data. This has led to growing privacy concerns, such as unlawful use of private data and disclosure of sensitive data[35,36]. We, therefore, need comprehensive privacy protection through holistic approaches to privacy protection that can also take into account the specific use of data and the transactions and activities of users [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.

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

3.4. Quantitative analysis

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

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

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

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

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

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

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

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

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]

Table 1. Quantitative Analysis

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

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 fulfill. 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.

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

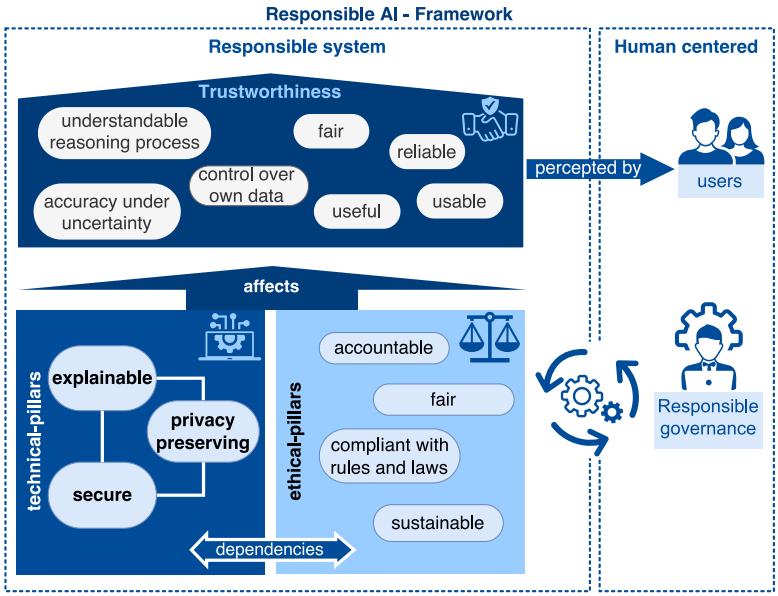


Figure 3. Pillars of the Responsible AI framework

Figure 4 shows that on the one hand there are social/ethical requirements/pillars and on the other hand the technical requirements/pillars. All of them are dependent on each other. If the technical and ethical side is satisfied the user trust is maintained. Trust can be seen as the perception of the users of AI.
There are also ”sub-modules” present in each of the pillars, like accountability, fairness, sustainability and compliance in the field of ethics. They are crucial that we can say the AI meets ethical requirements.
Furthermore, the explainability methods must value privacy, meaning they must not have that much access to a model so that it results in a privacy breach. Privacy is dependent on security, because security is a prerequisite for it.
With each ”responsible system” there are the humans that care for the system. The people who take care of the system must also handle it responsibly and constantly carry out

maintenance work and check by metrics whether the responsibility is fulfilled. This can be ensured by special metrics which are considered as a kind of continuous check as standard. This means responsible AI encompasses the system-side and the developer-side.
Human-Centered AI (mentioned in 3.3) needs to be considered as a very important part of responsible AI and it is closely connected to the approach "Human-in-the-loop". The human in the loop here is very important because this is the person who checks and improves the system during the life cycle. so the whole responsible AI system needs to be Human-Centered, too. This topic will not be dealt with in detail in this study, but is a part of the future work.
Therefore, responsible AI is interdisciplinary, and it is not a static but it is a dynamic process that needs to be taken care of in the whole system lifecycle.

4.2. Trade-offs

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

5. Research Limitations

In the current study, we have included the literature available through various journals and provided a comprehensive and detailed survey on the literature in the field of responsible AI. In conducting the study, we unfortunately had the limitation that some journals were not freely accessible despite a comprehensive access provided by our institutions. Although we made a good effort to obtain the information needed for the study on responsible AI from various international journals, accessibility was still a problem. It is also possible that some of the relevant research publications are not listed in the databases we used for searching. Additional limitation is the time frame of searched articles; this was carefully addressed to include only the state-of-the-art in the field. However, some older yet still current developement might have been missed out.
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' contribution in each of the subsections of the section 3.3.

6. Conclusion

The field of AI is such a fast changing area and a legal framework for responsible AI is strongly necessary. From the series of EU-Papers on Artificial Intelligence of the last 2 years we noticed that "trustworthy AI" and "responsible AI" are not clearly defined, and as such a legal framework could not be efficiently established. Hence, the trust as a goal

to define a framework/regulation for AI is not sufficient. Regulations for 'responsible AI' need to be defined instead. As the EU is a leading authority when it comes to setting standards (like the GDPR) we find it is absolutely necessary to help the politicians to really know what they are talking about. On the other hand, helping practitioners to prepare for what is coming next in both research and legal regulations is also of great importance.
The present research made important contributions to the concept of responsible AI. It is the first contribution to wholly address the "responsible AI" by conducting a structured literature research, and an overarching definition is presented as a result. The structured literature review covered 254 most recent high quality works on the topic. We have included a qualitative analysis of the papers covered.
By defining "responsible AI" and further analyzing the state of the art of its components (i.e., Human-centered, Trustworthy, Ethical, Explainable, Privacy(-preserving) and Secure AI), we have shown which are the most important parts to consider when developing AI products and setting up legal frameworks to regulate their development and use.
In the discussion section we have outlined an idea for developing a future framework in the context of Responsible AI based on the knowledge and insights gained in the analysis part.
In future work and research we will include a detailed analysis of the contribution of each of the analyzed papers to the defined aspects of responsible AI. Furthermore, the topic of Human-Centered AI and "Human-in-the-loop" should be developed in the context responsible AI. Other important topics to be worked upon are the benchmarking approaches for responsible AI and a holistic framework for Responsible AI as the overarching goal.

7. 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

References

[1] European Commission. White Paper on Artificial Intelligence A European approach to excellence and trust. European Commission,.; 2020. Available from: https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence.

[2] European Commission. Coordinated Plan on Artificial Intelligence 2021 Review. European Commission.; 2021. Available from: https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

[3] 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 Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115. Available from: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[7] Eitel-Porter R. Beyond the promise: implementing ethical AI. AI and Ethics. 2021;1(1):73-80.

[8] Werder K, Ramesh B, Zhang RS. Establishing Data Provenance for Responsible Artificial Intelligence Systems. ACM Transactions on Management Information Systems. 2022 Jun;13(2):1-23. Available from: https://dl.acm.org/doi/10.1145/3503488.

[9] Jakesch M, Buçinca Z, Amershi S, Olteanu A. How Different Groups Prioritize Ethical Values for Responsible AI. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM; 2022. p. 310-23. Available from: https://dl.acm.org/doi/10.1145/3531146.3533097.

[10] 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] Ibáñez JC, Olmeda MV. Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI & SOCIETY. 2021.
[23] Fjeld J, 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, Müller KR, Samek W, editors. xxAI - Beyond Explainable AI. vol. 13200. Cham: Springer International Publishing; 2022. p. 375-86. Series Title: Lecture Notes in Computer Science. Available from: https://link.springer.com/10.1007/978-3-031-04083-2_18.
[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, Závodszky G, Lees MH, Dongarra JJ, Sloot PMA, Brissos S, et al., editors. Computational Science – ICCS 2020. vol. 12140. Cham: Springer International Publishing; 2020. p. 615-28.
[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 Comput-ing 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 multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion. 2022 Jan;77:29-52. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1566253521001597.