Software Engineering's Key Role in AI Content Trustworthiness

Wumi AJAYI[1], Adekoya Damola Felix[2], Ojarikre Oghenenerowho Princewill[3], Fajuyigbe Gbenga Joseph[4]

[1]Software Engineering Department, Babcock University, Ilisan Remo, Ogun State Nigeria.

[2,3,4]Computer Science Department, Lead City University, Ibadan. Oyo State Nigeria.

DOI: https://doi.org/10.51244/IJRSI.2024.1104014

Received: 15 March 2024; Accepted: 22 March 2024; Published: 27 April 2024

ABSTRACT

Artificial intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems. It can also be defined as the science and engineering of making intelligent machines, especially intelligent computer programs. In recent decades, there has been a discernible surge in the focus of the scientific and government sectors on reliable AI. The International Organization for Standardization, which focuses on technical, industrial, and commercial standardization, has devised several strategies to promote trust in AI systems, with an emphasis on fairness, transparency, accountability, and controllability. Therefore, this paper aims to examine the role of Software Engineering in AI Content trustworthiness.

A secondary data analysis methodology was used in this work to investigate the crucial role that software engineering plays in ensuring the accuracy of AI content. The dataset was guaranteed to contain reliable and comprehensive material relevant to our inquiry because it was derived from peer-reviewed publications. To decrease potential biases and increase data consistency, a rigorous validation process was employed.

The findings of the paper showed that lawful, ethical, and robust are the fundamental components of reliable Artificial Intelligence. The criteria for Reliable Artificial Intelligence include Transparency, Human agency and oversight, technical robustness and safety, privacy and data governance, diversity, non-discrimination, fairness, etc. The functions of software engineering in the credibility of AI content are Algorithm Design and Implementation, Data Quality and Preprocessing, Explainability and Interpretability, Ethical Considerations and Governance, User feedback, and Iterative Improvements among others.

It is therefore essential for Software engineering to ensure the dependability of material generated by AI systems at every stage of the development lifecycle. To build and maintain reliable AI systems, engineers must address problems with data quality, model interpretability, ethical difficulties, security, and user input.

Keywords:- Artificial Intelligence, Software engineering, AI safety, Algorithm, **Explainable AI**

INTRODUCTION

A key component of artificial intelligence (AI) is the creation of machines with autonomous mental processes. AI, according to Legg and Hutter [1], is the process of teaching robots to do intelligent tasks by modeling human behavior and decision-making processes. Explainable AI is defined by Arrieta et al. [2] as using algorithmic techniques to produce transparent models that are easily understood and reliable by people. Explainability and interpretability are concepts that are frequently used synonymously [3]. Artificial intelligence (AI) and algorithmic decision-making have significantly changed daily life by offering guidance

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

and performing various tasks. Although algorithmic decision-making is not new, modern systems, powered by massive amounts of data, sophisticated algorithms, and powerful processors, are difficult to understand. These systems may be fragile and unjust due to their inherent limits, biases, and ethical considerations. As demonstrated by the CalGang database, which shows skewed and error-laden data leading to bias and injustice, relying solely on data for decision-making can add errors and biases [4].

There has been a noticeable increase in the government's and scientific communities' attention to trustworthy AI in recent times. With a focus on fairness, transparency, accountability, and controllability, the International Organization for Standardization (ISO), an organization that specializes in technical, industrial, and commercial standardization, has established a variety of techniques to foster confidence in AI systems [5]. The European Union (EU) has established moral standards for reliable AI to control and expedite the creation and functioning of AI systems [6]. A "right to explanations" for choices made using AI has also been granted to citizens by the EU through the General Data Protection Regulation [7]. A methodology for evaluating and improving user confidence in AI systems has been put out by the National Institute of Standards and Technology [8]. A framework for encouraging accountability and responsible AI use has been released by the U.S. Government Accountability Office (GAO) [9]. To further achieve the goal of making AI systems trustworthy and comprehensible, the Defence Advanced Research Project Agency (DARPA) initiated the Explainable Artificial Intelligence program [10]. Credibility is vital to the success of AI systems and the security of users and society, as demonstrated by these powerful institutions' active involvement. Gartner estimates that by 2025, thirty percent of all AI-based digital goods will need to follow a reliable AI framework [11]. Additionally, eighty-six percent of people will trust and stick with businesses that utilize ethical AI standards [12]. These incidents highlight how vital it is to create AI systems inside a framework that values reliability.

In response to the growing need for trustworthy AI, numerous new techniques and frameworks have emerged. Various approaches focus on different stages of the AI lifecycle to establish the dependability and credibility of AI systems. Some strategies focus on the design stage, creating reliable specifications and standards. Others improve data security, diversity, and justice by addressing data collection, protection, and pre-processing. Certain approaches concentrate on the modeling stage and provide interpretability and explainability. Simultaneously, some strategies focus on the supervision and execution stage, utilizing extensive testing and auditing for dependability and accountability. To ensure that AI is trustworthy, the European Union (EU) emphasized the significance of human involvement [6]. Furthermore, academics have put forth the idea of collaborative intelligence, which integrates human and computer decision-making [13]. These techniques aim to ensure that AI systems function as planned and do no harm, which will increase public confidence in these systems. This essay aims to investigate how important software engineering is to guarantee the reliability of AI content.

Problem statement

Artificial intelligence has played a more significant part in content generation in recent years. Concerns over the reliability of content produced by AI are, nevertheless, becoming more prevalent. Although artificial intelligence (AI) has the potential to completely transform several industries, including media and journalism, there are substantial obstacles in assuring the reliability of the data produced by these systems. Because AI systems are prone to biases, errors, and manipulation, inaccurate or misleading information may be disseminated. These problems seriously jeopardize people's access to impartial, trustworthy information and erode their faith in content produced by artificial intelligence. Software engineering is essential to solving the issue of trustworthy AI content. Creating and deploying reliable algorithms and systems that can identify and reduce biases, errors, and manipulations in AI-generated material falls under the purview of software engineering. Software engineering can help create AI systems that are more dependable and trustworthy by considering elements like fairness, explainability, and transparency [14,15]. Therefore, this

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

paper aims to examine the role of Software Engineering in AI Content trustworthiness.

LITERATURE REVIEW

The Collaborative Potential of Software Engineering and Artificial Intelligence

Since both software engineering and artificial intelligence have their roots in computer science, they are inextricably linked. Although modern artificial intelligence (AI) dates to the early 1950s or 1940s, software development, and AI first came together in 1959 when the IBM 704 computer's Mark 1 perceptron was designed for picture identification. Interestingly, there was little connection between the perceptron—which was originally designed to be a machine—and software engineering as it is known today. To address the difficulties in developing ever-more-complex software systems, the term "software crisis" was coined in the late 1960s, which also marked the formalization of software engineering as a profession and field of study [16]. Artificial intelligence (AI) gained prominence again in the 1980s, especially with the introduction of Expert Systems [17]. A new set of Software Engineering difficulties, such as validation and verification, inevitably arose with the integration of expert systems into production systems. One survey that examined the complex connections between AI and software engineering was conducted by Partridge [18]. Even though they address different issues, artificial intelligence, and software engineering have been combined since their birth. Both Software Engineering methods and AI approaches have been used to support Software Engineering tasks (AI4SE) and to produce AI software (SE4AI).

Perkusich et al. [19] recently defined AI for Software Engineering (AI4SE) as intelligent SE, describing it as a collection of SE techniques that explore data taken from digital artifacts or domain experts for knowledge discovery, reasoning, learning, planning, natural language processing, perception, or decision support. The increasing complexity and size of software systems drive the development of AI4SE and correspondingly increase the volume of SE activities. Research started concentrating on automatable techniques when software engineers faced cognitive constraints. The SE community investigated the incorporation of machine learning techniques to offer support after realizing that many SE jobs might be framed as data analysis (learning) tasks. AI algorithms were first implemented as stand-alone programs using SE, as demonstrated by the Mark 1 perceptron. However, when AI-powered software systems became more complex and gained more real-world and business uses, there was a need for more sophisticated SE techniques. The turning point came when AI elements were integrated into well-known software systems, including driving control or expert systems. It became clear that standard SE methods were no longer appropriate because of the special qualities of AI, which resulted in problems like technological debt [20]. This required the creation of new techniques as well as a reevaluation of traditional SE concepts.

Trustworthy Artificial Intelligence

Technically speaking, trustworthiness is the assurance that a system or model will function as intended. Confidence arises when faced with a particular issue [21], which instills confidence in the model user (the audience). There are several ways to strengthen this trust, including giving thorough justifications for system choices [22]. Lipton asserts that when a user is aware of a model's operation and decision-making process, their trust in using it grows [23]. In a similar vein, confidence in the model's dependability in a variety of scenarios, its dedication to privacy, and its resilience to biases in the training data can all contribute to increased trust. Thus, for people and communities to create, deploy, and use AI systems, trustworthiness becomes a multifaceted requirement. This is a necessary step to realize the enormous potential social and economic benefits that AI may provide [24]. Furthermore, reliability encompasses not just the system but also various actors and procedures across the AI life cycle. This calls for a thorough analysis of the components and prerequisites that support the development of user trust in AI-based systems.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

The fundamental components of reliable Artificial Intelligence

Broadly speaking, a pillar is a fundamental truth that forms the basis of a specific notion or concept and provides essential conditions for making that idea a reality. In a similar vein to structural engineering, pillars are integral to the notion of trustworthy AI; each is necessary, but none by itself is adequate to attain trustworthy AI. Similar to how formwork, concrete, and cantilevers support pillars in supporting a building's structure during construction, key requirements may support one or more pillars. Throughout the AI system's life cycle, these standards must be continually upheld, utilizing approaches that go beyond technical details to include human involvement. The three essential characteristics that trustworthy AI systems should possess are outlined in the EU Ethical Guidelines for Trustworthy AI [24].

(i). Lawful: Reputable AI systems ought to abide by all relevant laws and guidelines, both vertically—such as the domain-specific requirements enforced in some high-risk application fields, like banking or medicine—and horizontally—such as the European General Data Protection Regulation.

(ii). Ethical: AI systems that are considered reliable need to adhere to ethical standards and values in addition to legal requirements. Current AI-based systems are advancing at a rapid pace, raising ethical questions that are not always addressed by governmental activities. Current examples that highlight the importance of ethics as a cornerstone component of a reliable AI system are the pervasive use of big language models and the propagation of false information via deepfakes.

(iii). Robust: Reputable AI systems must make sure they don't do unintentional harm and operate safely and dependably from a technical (performance, confidence) and social (use, context) perspective.

These three pillars form the basis of Trustworthy AI. In an ideal world, these pillars would be harmonious and supportive of one another, guiding the development of reliable AI. However, conflicts could arise between them since, for example, morality and the law do not always coincide. On the other hand, moral considerations may need changes to the law, which would go against current regulations. For trustworthy AI to achieve its intended effects on the socioeconomic context in which it is deployed, it must guarantee adherence to moral standards and values, obey legal requirements, and function reliably. Changes that might be in opposition to the rules as they currently stand.

Criteria for Reliable Artificial Intelligence

1. Human agency and oversight

Artificial intelligence technologies ought to provide people with more authority so they can defend their fundamental rights and make educated decisions. Establishing efficient supervision methods is also critical, and this may be done by utilizing strategies like human-in-the-loop, human-on-the-loop, and human-in-command. To achieve autonomy and control, artificial intelligence systems essentially need to improve human autonomy and decision-making capacities. AI systems' unfair manipulation, deceit, herding, and conditioning endanger people's independence, autonomy, and rights. Therefore, reliable AI systems should enable user oversight, evaluation, and the freedom to accept or reject judgements made by the system, avoiding automated decisions that are made in the absence of human input [25].

a. The agency of humans

The application area and potential dangers will determine the different solutions for human oversight. Computing paradigms that support human computation or interactive machine learning techniques [25], human-centric methodologies [26], human-compatible approaches [27], AI for social good initiatives [28], and AI for human computation and human rights uphold these requirements. More organized toolkits, like

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

the ones described in the C-Suite [29] or the World Economic Forum [30] must be created, though, to seamlessly take this requirement into account that is specific to certain domains. Language appears as the common language used by machines and people when it comes to technical instruments intended for a variety of audiences. To help humans monitor and make the best judgments possible based on AI system outputs, artificial intelligence models that use natural language processing and/or provide counterfactual and natural language explanations are therefore essential [31].

b. Human oversight

The European Commission [30] has identified several tiers of human involvement in managing AI-powered systems:

1. Human-in-the-loop refers to the supervisor's ability to modify every cycle of the system under observation [25].
2. Human-on-the-loop refers to the participation of humans in the AI-based system's design and monitoring cycles.
3. Human-in-command, which denotes the supervisor's capacity to supervise the AI system's overall operations, including its wider moral, legal, and economic ramifications. This entails making certain that human intervention is available for decisions made by the AI system.

Mechanisms can be designed to accommodate any of the levels of human monitoring, depending on the application in question. Since the supervisor's background, skills, and the AI-based solution's design can all have an impact on the user algorithm interface, the approaches that have been proposed so far have primarily been domain-specific.

2. Technical robustness and safety

This secondary prerequisite comprises an array of features that are all focused on mitigating the effects of deliberate damage and averting accidental injury. The ability of AI-based systems to withstand security threats and attacks, the creation of backup plans in case of mistakes, maintaining correctness, dependability, and reproducibility, and guaranteeing overall safety are some of these functionalities. The importance of robustness and safety highlights the need for AI systems to be safe, dependable, and able to withstand mistakes or inconsistencies at every step of their existence [32]. When AI systems operate in real-world scenarios, their surroundings may change, which might cause changes in the inputs they receive (e.g., idea drift). Adversarial interactions between malevolent users and the AI system may lead to such modifications. The AI system's credibility depends on how well it can reduce the negative effects of these changes on its predictions, regardless of whether these modifications are deliberate or not. Reputable artificial intelligence systems must evaluate pertinent safety precautions in high-risk scenarios and have contingency plans in place in case the system deviates from its intended course. Furthermore, trustworthiness is intimately linked to repeatability and reliability, guaranteeing the validation of AI systems' anticipated functionality and functionality. These elements are vital for guaranteeing that the system adapts robustly to variations and idiosyncrasies, reliably producing the desired results, especially in scenarios where AI systems are employed in various surroundings and integrated into various systems. Three aspects can be used to evaluate methods that meet this requirement: reproducibility, safety, and technological robustness.

a. Technical robustness

Robustness and reliability in the context of AI-based systems refer to the system's capacity to continue operating as usual even in the presence of unusual data in contrast to standard operating scenarios [32]. Robustness becomes important when one expects the model to degrade, be perturbed, or have an impact on it in the future. A good model should be resilient against adversarial assaults or models, robust against data

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

disturbances, and strong in terms of its generalization or generative capacities. AI maintenance frameworks ought to assist in achieving robustness and tracking the state of AI throughout its life cycle [33]. Monitoring can take place in two ways: either passively, by measuring different quantitative metrics related to robustness over the data, model, or both regularly, or actively, by simulating scenarios in which the robustness of the model might be compromised (e.g., by simulating adversarial attacks or perturbations of known samples). In both scenarios, systematic deviations in the metrics inside data and models can be detected by AI maintenance frameworks to identify model degradation over time [34].

Fascinatingly, current AI research aims to provide AI-based systems with the capacity to continually learn from an infinite variety of data streams, measure and convey confidence in their outputs, and recognize and synthesize new patterns over time. To wrap off the talk about strengthening technological robustness in AI-based systems, it's important to highlight the methods that can be used to fulfill other needs. Explainability approaches, for example, might improve the robustness of a model by identifying which characteristics are more resistant to adversarial attacks or changes in input that occur outside of the model's distribution. Like this, in a counterfactual generation, the size of modifications required to reach a desired adversarial confidence score is a good predictor of how much a particular data instance might depart from the distribution. All things considered; these examples show how extra features added to an AI-based system can concurrently support several trustworthiness criteria [35].

b. Safety

The idea of safety in AI is developing, originating from a broad framework of information technologies, and emphasizing congruence with human values. It is crucial to maintain safety in AI, even though defining precise standards and procedures might be difficult. The field of AI safety comprises several open research questions [36], such as:

-To build systems less vulnerable to adversarial threats, such as adversarial perturbations that result in high-confidence errors, and resilient against long-tail difficulties, robustness must be achieved.

– Creating instruments to examine AI-powered systems, spot risks, and irregularities, adjust them, recognize real results, and spot new abilities. The possibility of backdoors is one risk associated with AI systems that highlights the need for safety tools: backdoored models function correctly in nearly all scenarios, apart from a few special cases where they are purposefully trained to behave incorrectly through training on tainted data, injecting backdoors. This is especially troublesome for foundational models, which are the structural cornerstone of downstream models and are all derived from initially contaminated data in large training datasets.

– Setting up safety goals to direct models, both outside (figuring out how to safely pursue this safety compliance) and internally (figuring out how models should learn to assure compliance with safety measurements). Among the difficulties in this regard are:

– Learning human values: this refers to the difficulty AI systems face in learning human values that are subjective, such as safety, happiness, sustainability, or meaningful experiences. Although exposing models to a variety of real-world inputs can aid in the distinction between happy and unpleasant states, the utility values assigned to these states are merely reflections of the model's learned utility function and not absolute facts.

– Gaming via a proxy: This is the result of adversaries and optimizers manipulating objective proxies. According to Goodhart's law, when a measure is the only goal, it becomes less reliable as an indicator. Proxy gaming, for example, happens when hacking rewards are used in reinforcement learning. Similarly, objective countable measures may unintentionally replace human values when opaque AI models are forced

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

to optimize a single quantitative measure. Therefore, it is not enough to just obtain a proxy for human values; models also need to be resistant to manipulation attempts [35].

c. Reproducibility

Reproducibility is a critical component in the essential criterion for reliable AI, following robustness and safety. In this sense, reproducibility refers to an AI experiment's capacity to exhibit consistent behavior when carried out in the same settings. It is strongly related to replicability, which is the process of independently coming to findings that are not the same but are at least comparable despite differences in data analysis, research methods, and sampling [26]. An experiment should be seen as an assessment of the safety, robustness, and performance of a particular AI-based system in the context of reliable AI systems. Different organizations, such as research groups or certification labs in the case of commercial software-based solutions, may conduct this review. Reproducibility in reliable AI systems is dependent on various elements, including system secrecy and the particularity of the experimental configuration used to construct the AI system. Public distribution of the source code for the proposed AI-based system facilitates third-party repeatability of experiments in less constrained situations, including research.

3. Privacy and data governance

This requirement makes sure that data protection and privacy are maintained across the AI system's whole life cycle, which includes design, training, testing, deployment, and operation. Robust data governance processes are necessary to ensure authorized access to data and adherence to processing rules, while also taking data quality, integrity, and domain relevance into account. Artificial intelligence systems, especially those that rely on digital records of human behavior, can reveal sensitive personal information such as age, gender, sexual orientation, and political and religious beliefs, as well as infer individual preferences. Because AI-based systems are data-driven, it is critical to guarantee the confidentiality of such personal data as it is being processed, stored, and retrieved throughout the AI life cycle. This entails putting in place systems to track data usage (governance) and ensuring privacy awareness by ensuring that sensitive data is inaccessible at all stages of the life cycle. In addition to undermining user confidence in AI systems, not providing these guarantees results in non-compliance with current laws, such as the European GDPR. With the guarantee that their data won't be exploited unfairly or illegally to cause them harm or to discriminate against them, citizens should have total control over it [32]. Enforcing human rights such as the right to privacy, intimacy, dignity, and the right to be forgotten requires compliance with these criteria. To prevent digital information from being used to categorize people into profiles that may not accurately reflect reality, it emphasizes the significance of limiting, protecting, and informing people about the use and scope of their data [37]. This can help to avoid maintaining historical and cultural biases, reinforcing stereotypes, or perpetuating historical differences among minorities.

a. Privacy of data

The importance of Federated Learning (FL), homomorphic computing, and differential privacy (DP) is emphasized as examples of privacy-aware technologies in the present AI landscape to transform the data privacy requirement into concrete technology. Federated Learning involves training a model on several dispersed devices without moving the data to a single hub. Devices locally train models using their data, transmitting just numerical model updates to the central server, as opposed to sending all the data to a central server. To generate a new model, the central server compiles the updated model parameters from all servers or devices. This allows for the learning of a global model while maintaining the confidentiality of data. FL expedites model training, lowers communication costs, and protects the privacy of local data. Encrypted data can be processed using homomorphic computing without requiring decryption. By conducting operations directly on encrypted data while maintaining the underlying data structure, this guarantees data security and privacy. Only those with permission and the decryption key can view the

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

encrypted computer result. In AI-based systems, homomorphic computing works well for enabling privacy-aware preprocessing, training, and inference. Finally, DP reduces the possibility of identifying specific persons in the dataset while enabling data processing and learning. Before processing the data, DP adds random noise to it. This noise is adjusted to preserve statistical accuracy while hiding any information that can be used to identify specific people or violate their privacy. The amount of noise added strikes a compromise between the AI-based system's performance degrading in comparison to circumstances without noise and privacy protection. By putting any of these technologies—or combinations of them—into practice, the risk of harm is reduced and individual privacy in datasets is protected [35].

b. Data governance: Quality and integrity of data and access to data

Data protocols ought to supervise the accuracy and availability of data for every person, including those who do not utilize the AI system directly. Access to personal data should be restricted to trained people who have demonstrated a clear necessity and level of competence. Data governance necessitates a more comprehensive regulatory framework than that of a single nation or continent as part of AI governance. Due to the significance of guaranteeing data quality, integrity, and access, rules and suggestions for AI governance have been developed because of this viewpoint. As an illustration, consider the 2018 publication of the Universal Guidelines for AI [38], which received support from 300 experts from 40 different nations as well as 70 organizations. The Data Quality Obligation is one of these rules; it was created as a tenet to be included in moral norms, accepted in laws and international accords, and deeply embedded in the architecture of artificial intelligence (AI) systems. The Council of Europe Convention on AI, the OSTP AI Bill of Rights (2022), the UNESCO Recommendation on AI Ethics (2021), the OECD AI Principles (2019), and the EU AI Act have all been impacted by these proposals. The Information Commissioner's Officer (ICO) has provided guidelines for establishing data governance and suggestions for the appropriate and legal use of AI and personal data [39].

These include developing and implementing AI using a risk-based approach, addressing bias and discrimination risks early on, guaranteeing meaningful human reviews of AI decisions, gathering only the data that is required, and working with outside vendors to ensure appropriate AI use. The 2020 European Strategy for Data seeks to establish the European Union as a data-powered society at the continental level. The European Data Governance Act [40] was developed because of this policy and makes it easier for Member States and sectors to share data. For example, the EU Data Governance Act aims to facilitate the exercise of rights under the General Data Protection Regulation (GDPR), allow data use for altruistic purposes, enable the use of personal data through a "personal data-sharing intermediary," encourage data sharing among businesses, and make public sector data reusable [40]. The Data Act [41], a law that harmonizes regulations on equitable access to and use of data, was suggested in 2022 by the European Union's data strategy. By outlining who can and when they can derive value from data, this legislation enhances the Data Governance Act.

4. Transparency

According to R. Mariani's definition of transparency [42], it guarantees that pertinent parties are informed. Transparency in AI-based systems can be divided into three categories: algorithmic transparency (knowing the workings of the model and how it will behave for any given output), decomposability (explaining the behavior and components of the model), and simulatability (the model's human-understandability) [43]. Another classification emphasizes the role of the targeted stakeholder audience, which includes developers, designers, owners, users, regulators, or society, and takes transparency into account at the algorithmic, interaction, and social levels [44]. Transparency in data, the system itself, and AI business models is essential for reliable AI systems. People should be aware that they are dealing with AI systems and should be educated on the capabilities and limitations of systems [24]. Consequently, explanations must be given to the relevant stakeholder audience—whether it is a layperson, regulator, researcher, or another

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

stakeholder—promptly and with appropriate adaptation. It's also crucial to guarantee the traceability of AI systems. This criterion includes aspects like communication, explainability, and traceability, all of which are necessary to implement transparent AI-based systems.

a. Traceability

According to Estévez in [26], traceability is the collection of practices and techniques intended to keep an eye on the data, development, and deployment activities of the system, usually through recorded identity that is documented. AI-based systems benefit from early traceability and logging deployment in the design phases, which facilitates auditing and attains the required degree of transparency catered to the requirements of the pertinent audience. According to Pérez [45], provenance technologies are essential for supporting the traceability or lineage of data and model decisions, which helps meet the transparency criterion. Blockchain techniques, which focus on the provenance of databases and their related quality, bias, and fairness in this context, offer promise in guaranteeing the integrity of data used for training and elucidating machine learning models.

b. Explainability

Explainability approaches are becoming indispensable for algorithmic auditing; they are a critical first step toward understanding and verifying the information gathered by black-box models, which are models that only allow for the observation of inputs and outputs without providing insight into their underlying workings. Elucidating the input parameters that influence complex black-box algorithms' decisions can provide a useful general understanding of the model's operation. This adds to a thorough understanding of the model's functioning, together with traceability techniques and efficient information dissemination that is catered to the intended audience. Many taxonomies of eXplainable AI (XAI) methodologies have been presented over time, given that the quality of explanations depends on the audience and is linked to the reason for which they are generated [43]. Model-agnostic and model-specific ways to explain machine learning models are classified depending on whether the XAI technique can be applied to any machine learning model, regardless of its structure and learning algorithm. This is a basic classification. Ex-ante and post-hoc XAI approaches are divided into several categories based on whether explainability is taken into consideration before or following the model's creation and training. Ex-ante techniques, also known as the explainable-by-design paradigm, seek to have AI models automatically offer explanations without adding extra models or needless complexity. This way, explanations are as close as possible to the actual reasoning that the model is carrying out.

On the other hand, post-hoc XAI approaches generally add more components to the original AI model or create a stand-in version, like a local approximation or a streamlined iteration of the original model, to enable a simpler explanation of the original one (e.g., LIME). Furthermore, certain XAI approaches can make use of outside information, gleaned from websites, Wikipedia, or discussion boards, to clarify language models or conversation models that react on the fly to questions concerning a model's particular choice. Several factors can be used to categorize XAI technologies, such as how the explanations are presented (attribution techniques, counterfactual analyses, streamlined model surrogates, or the combination of explanations presented in many modalities, such as text and visual). Making explanations understandable to non-expert audiences requires the use of natural language explanations, quantitative assessments of explanation quality, and models that support their learning processes with formal symbolic foundations like language, rules, compositional relationships, or knowledge graphs (neural-symbolic learning and reasoning). According to Díaz-Rodríguez [35], these interfaces enable users to evaluate the model's functionality in a more comprehensible way, supporting the need for human agency and oversight in reliable AI systems.

c. Communication

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

The third aspect of transparency is how the audience is informed about the AI-based system—that is, how the user is given explanations or specifics about how the system works. People must be conscious of when they are interacting with AI systems, getting signals about their limitations, performance, and advice on what they can do.This openness includes providing the user with an explanation of the model's operation and output. The explanations must be modified to consider the unique features of the AI system being evaluated as well as the audience's cognitive capacities (knowledge, prior experience). As such, good communication is essential to making sure that the audience is presented with all aspects of transparency in a way and format that suits their experience and understanding. This flexibility is essential to gaining the audience's trust in the relevant AI-based system.

5. Diversity, non-discrimination, and fairness

This criterion covers several areas, including the avoidance of unjust bias, the encouragement of diversity, accessibility for all people with disabilities, and stakeholder participation throughout the whole life cycle of an AI system. The aim of this comprehensive requirement, which is various, is to guarantee that AI-based systems do not deceive people or unnecessarily limit their freedom of choice. As such, it is a prerequisite that is strongly related to the moral and equitable values that support AI's reliability. If the data used to train models contains hidden biases, this mandate is crucial for extending the good impact of AI across all social strata and reducing any potential negative consequences that automated judgments might have in practice. A model trained on such data may make judgments that have a variety of unfavorable effects, such as marginalizing vulnerable groups and fostering prejudice or discrimination [24]. For this reason, it is imperative to avoid unjust bias in data. Strategies for fulfilling these criteria can be divided into groups according to the aspect they are meant to promote. Examples of these groups include techniques for guaranteeing diversity, nondiscrimination, accessibility, universal design, and involvement of stakeholders.

a. Diversity, non-discrimination, accessibility, universal design, and stakeholder participation

To provide accessibility for everyone, AI systems must take into consideration the entire range of human talents, skills, and requirements. It is essential to create procedures based on the ideas of bias reduction and non-discrimination to match requirements with moral standards. Equal treatment is promoted by characteristics including non-discrimination, fairness, and bias mitigation, which aim to remove systemic disparities in the treatment of groups. For example, suggestions advocate hiring people from a variety of backgrounds to guarantee a diversity of viewpoints. Ensuring that persons who may be marginalized have fair access to the advantages of AI is ensured by this criterion, which emphasizes the incorporation of varied data and individuals. To provide a diverse representation that includes minorities, implementations include addressing the consequences of missing data, pushing for diversity in data-gathering methods, and evaluating the impact of demographic imbalances. Biased models and non-discriminating automated processes are actively opposed. Diversity can be promoted through strategies like unbalanced learning or data augmentation. Diversity is employed during model training by penalizing the lack of diverse prototypes. The emphasis also includes the solutions generated by AI systems, highlighting the significance of varied results to guard against prejudice. Unifying quality and diversity optimization frameworks provide a variety of solutions, which is essential for robotics to learn a wide range of behavioral regulations. Using techniques like compositional fairness, discrimination-aware data mining, and interpretation through sanity checks and ablation studies, the discrimination-conscious by-design paradigm, applied globally, integrates discrimination detection and prevention from the outset of AI system design [35].

b. Fairness

The second aspect of this need is fairness, which focuses on methods designed to reduce the likelihood that AI-based systems may produce biased results. If an algorithm produces results that benefit a particular

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

group, it is deemed unfair. Such biased decisions can result from a variety of sources, including algorithmic biases affecting users (e.g., popularity, evaluation, or measurement biases); data-related biases like measurement, omitted variable sampling, or representation biases; and user-induced biases affecting the data (e.g., temporal, historical, social, or content production biases). It is well known that attaining fairness could necessitate making certain compromises with accuracy. Nonetheless, there are methods for removing false features from machine learning models, which could improve their effectiveness. Another factor to consider is the trade-off between privacy and fairness, where adversarial learning can be used to simultaneously train an adversary to model a protected variable and a predictor. Reducing the adversary's capacity to forecast this variable can lead to precise forecasts with less stereotyping, getting closer to the fairness principle of chance equality. Fair washing, or the creation of a false impression that a machine learning model complies with moral principles through outcome explanations and fairness measures, is a key idea in the discussion of fairness in AI-based systems. To audit unfair opaque models successfully, it is imperative to grasp how manipulable fair washing is and use strategies such as Laundry ML [46,47].

c. Avoidance of unfair bias

During the entire AI life cycle, which includes the design, development, and deployment stages of AI-based systems, it is imperative to ensure variety, representation, and completeness in both data and models. Since potential bias-inducing factors can affect data and models, a thorough approach to this problem is necessary. A major challenge is proxy discrimination, in which models may incorporate proxy variables that unintentionally cause prejudice. For instance, utilizing zip codes as proxy variables in predictive policing may result in biased results [35]. Beyond algorithms, prejudice exists in the form of a vicious cycle that begins with bias in human behavior, moves on to bias in data, and includes bias in online sampling, algorithmic bias, self-selection bias, and interaction bias. According to Baeza-Yates [48], the latter may feed back into the algorithm, producing a second-order bias.

Using a variety of strategies, from general guidelines and toolkits to taxonomies of bias at distinct phases of the AI life cycle, is necessary to mitigate prejudice. Various concepts of justice, including counterfactual justice and causal justice, offer frameworks for addressing bias [49]. Counterfactual fairness takes other possibilities into account, whereas causal fairness depends on causal linkages and creates causal graphs. By measuring fairness associated with various explanatory factors, causal mediation analysis, which leverages causality, can help debug algorithmic bias mitigation or explain models [35].

6. Societal and environmental wellbeing

AI systems should be beneficial to humanity in the now and the future, requiring environmental friendliness and sustainability. Artificial intelligence technology should not deplete natural resources or upset ecological balance. An extensive assessment of AI's social and societal effects, sustainability, and environmental friendliness are important factors that support this need. The intention is for AI systems to support ecological responsibility and sustainability while positively impacting society. Problems arise, nevertheless, from the damaging greenhouse gases released during the computationally demanding training of sophisticated AI models. For example, throughout their training, a single AI model can produce as much CO2 as five cars. As the number of parameters in the model increases, so does the computational and environmental cost of the model [35]. A study [50], which concentrated on big language models, found an annual cost of about 8.4 tons. To put things in perspective, each person's annual carbon footprint is approximately 4 tons. Even though the emissions were dispersed during the model's lifetime, the ChatGPT model was discovered to have a 1287 MWh consumption and 522 tCO2e in cost. As a result, organizations and businesses using AI must take energy and regulatory concerns into account. Currently, there are two approaches used to satisfy this requirement: societal impact and sustainability and environmental well-being.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

a. Sustainability and environmental wellbeing

According to [51], sustainable AI adopts a holistic approach that addresses hardware, data algorithms, and models. It also emphasizes the potential of software-hardware co-design to reduce the carbon footprints associated with the full life cycle of AI models, including the phases of design, training, and deployment. The field of sustainable AI seeks to disseminate essential knowledge, best practices, metrics, and standards to direct the sustainable development of AI systems considering the significant energy consumption of large AI models. Technical advancements in the field of Green AI [52] concentrate on developing eco-friendly and effective designs for AI-based assets, systems, and algorithms. Many approaches have been put forth to satisfy this requirement, especially for AI models like deep neural networks that have many parameters and require a long training period.

b. Societal wellbeing

AI can greatly improve social welfare on a societal level. AI-based solutions can increase productivity and enhance human well-being by safely and effectively handling everyday chores on their own. Artificial Intelligence (AI) has the potential to accelerate workflows, simplify administrative procedures, and minimize paperwork in public administration. It can also help with legislation and support city planners in several ways, such as detecting urban heat islands, forecasting future floods, and illustrating the effects of climate change. Advances in AI have created more opportunities for society to benefit from them as more sectors are becoming digitally connected. The entire potential of AI may be leveraged by sectors such as infrastructure planning, health, justice, equality, inclusion, education, economic empowerment, security, and hunger reduction to address societal concerns.To fully realize these advantages, though, enormous volumes of data must be used in AI-based systems that handle learning tasks meant to address important social issues. Fairness, privacy, openness, and human oversight are critical aspects of trustworthy AI since the judgments made by these systems affect people and are scrutinized by society. Most significantly, since judgments in fields like justice, education, and security must be made following current legal constraints and fundamental human rights, AI ethics and regulation are critical to the health of society.

7. **Accountability**

The last prerequisite for reliable AI systems is the creation of procedures that guarantee accountability and responsibility during the creation, implementation, upkeep, and utilization of AI systems and their results.An essential component of accountability is auditability, which makes it possible to evaluate data, algorithms, and design procedures and link outcomes to actions that were performed in response to the AI-based system's outputs. Minimizing harm, disclosing adverse effects, informing users of design trade-offs, and putting in place easily accessible redress mechanisms for AI-based systems are all parts of accountability. As a result, auditability and accountability are fundamental to responsible AI systems and are closely related. The development of useful tools to confirm desired neural network properties, such as stability, sensitivity, relevance, or reachability, as well as metrics beyond explainability, like traceability, data quality, and integrity, is required because auditability is essential for trustworthy artificial intelligence systems. Auditability, which includes recommendations from organizations like as IEEE, ISO/IEC, and CEN/CENELEC for the implementation of reliable AI needs in industrial setups, is becoming more and more important as standards are being developed. However, accountability is necessary for redress in cases where an AI model produces erroneous conclusions, offering justifications and suggestions for situations in which such decisions hurt cases. A key element of accountability is adherence to moral and legal requirements. Other elements include answerability, reporting, supervision, attribution, and the imposition of sanctions. Accountability is essential for allocating expenses, risks, responsibilities, and liabilities among the different parties participating in the AI life cycle, and it is framed within ethical and regulatory guidelines

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

for AI [35].

a. Accountability

Accountability mechanisms are especially important in high-risk situations since they assign blame for decisions made during the AI system's design, development, and deployment stages. Algorithmic accountability policy toolkits, which involve post-hoc model output analysis (e.g., via local relevance attribution methods) or algorithms for causal inference and reasoning, are among the tools to meet this criterion. Since fairness principles and accountability are closely related, risk management and accountability are closely related because unjust consequences may occur. Accountability requires the open identification and mitigation of risks, enabling third-party verification. For AI-based systems to make judgments that are accountable, auditing techniques and tools for data, algorithms, and design processes are required. Many risk assessment frameworks exist (53, 54] describe derisking procedures that are incorporated into the design and development stages. These procedures, which include bias mitigation and explainability, operationalize risk management in machine learning pipelines. A discussion of further resources for tackling fairness and bias can be found in [54]. Any emerging trade-offs between needs should be spelled out and evaluated for their potential to jeopardize moral standards or violate fundamental rights. When a risk-free trade-off for these factors cannot be identified, no AI system should be employed [24]. As a result, multi-criteria decision-making and MLOps-level pipelines that outline and explain these trade-offs to the user are frequently included in AI models created for accountability.

b. Auditability

According to Mökander[55], the AI Act serves as the foundation for AI audits in Europe. More broadly, even within worldwide ISO standards for AI robustness, the need to certify systems with AI-based features is becoming more and more prevalent. By extending formal methods for requirement verification or satisfaction—which are widely used in software engineering—these standards enable the verification of desired attributes of AI models. Stability, sensitivity, relevance, or reachability are desired qualities for neural network certification [56]. Grading schemes specific to the use case are required for auditing procedures, especially when AI systems engage with users[57]. This is because models need to be validated. Examples are the Muir Trust Scale [58] and the System Causability Scale [57], which are extensively used in robotics and human-robot interaction. These scales are based on four factors: competence (the user's level of confidence in the system's ability to handle similar situations in the future), trust (the user's overall level of trust in the system), predictability (the degree to which the robot behavior or the output of the AI-based system can be predicted from moment to moment), and reliability (how much the user can count on the system to do its job).

c. Minimizing and revealing negative consequences and trade-offs.

Given the state of advanced generative AI, reliable and verifiable auditing procedures for AI-based systems must be established. It is becoming difficult to discern between machine-generated and human-created multimodal material due to the maturity of generative AI. Inaccurately classifying such content could cause misunderstandings and misinformation, which could have negative effects on society by influencing public opinion or disseminating false information. Verifiable statements, which are defined as falsifiable claims supported by arguments and evidence that might affect the likelihood of their truth, are a viable solution to these problems [59]. This project is the result of the cooperative efforts of developers, regulators, and other stakeholders in artificial intelligence. It is motivated by the necessity to identify the attributes of AI systems that can be effectively demonstrated, the techniques used to accomplish so, and the measurable costs or benefits associated with them. While the achievable degree of assurance may vary depending on specific

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

claims and situations, the general idea is to demonstrate that more substantial evidence can be provided for claims regarding AI development than is typically done, which will make the auditing process easier.

d. Redress

Finally, to keep the user's faith when unfavorable or unfair effects happen, it is critical to let them know about their right to seek redress in the case that a confirmed incident results from the recognized risk [24]. Redress is a technique to correct or reverse an AI system's conclusion that is deemed erroneous. It is strongly linked to the idea of algorithmic recourse [60]. Trustworthy AI must guarantee effective redress against decisions made by both people operating the systems and the AI itself. When these mechanisms fail, users, particularly vulnerable persons, or groups, must have access to these procedures. Setting up and utilizing accountability frameworks and disclaimers in addition to redress procedures is crucial, as certification is domain-specific and cannot take the place of comprehensive responsibility. Supportive techniques for redress in AI-based systems can include methods like machine unlearning [61], counterfactual explanations [62], or the investigation of uneven impacts [63].

The Imperative for Trustworthy AI

Artificial intelligence (AI) systems are rapidly transforming several aspects of life, from movie recommendations to illness diagnosis, customer support, and more [13]. Even though AI has many uses, its quick development has raised a lot of questions. The potentially catastrophic effects of unchecked AI growth were highlighted by the late Stephen Hawking [64]. AI systems can be harmful and dangerous if strict design and oversight guidelines are not followed [65]. Nowadays, a lot of different industries use AI systems to make decisions, however, these systems don't always produce the best results. AI systems are useful, but they have a big obligation to make sure they don't hurt people. Sadly, there have been times when these AI systems malfunctioned and caused dangerous situations for people. For instance, it was discovered that the COMPAS algorithm, which is often used across the country to estimate the likelihood of criminal recidivism, has a bias against Black people [66]. Furthermore, because the training sample data for the facial recognition software was of low quality, black people were mistakenly classified as white [67]. Additionally, a large tech company's resume screening process showed prejudice against women [68]. These examples show how prejudice can mislead opaque systems, resulting in injustice or injury. Because of their erratic behavior, some of these systems have even caused harm. For example, when a pedestrian was detected by a self-driving car's sensors, the algorithm malfunctioned and the car failed to react, resulting in the person's death [69]. Furthermore, the complexity of these systems makes decision reasoning more difficult to understand, which restricts the full potential of these systems' applications. Fan et al. [70] showed that, despite their potential usefulness in routine clinical procedures, healthcare professionals' adoption of AI-based medical diagnosis support systems is very low. The low adoption rate among doctors is ascribed to the incomprehensible character of these systems, which erodes their trust and acceptability. These instances highlight how important it is to guarantee the security and reliability of AI systems, which are already capable enough to be used extensively in society.

People's lives are already being significantly impacted by these technologies [71]. But it is important to understand that just because something is useful, it does not mean that it is trustworthy or dependable. It is improper to use these methods informally, particularly in high-stakes situations when a single wrong choice could have dire repercussions. Biases and fragility are common in AI systems. Marcus and Davis [72] use an example involving facial recognition software to demonstrate this. While less trustworthy software could be fine for automatically identifying users in social network photos, it is improper for law enforcement to utilize the same technique to identify suspects in surveillance images. This demonstrates how people often accept AI systems only in situations where they do not pose a threat to their lives. An ethical framework for AI system control and governance is crucial to guaranteeing the reliable deployment of these systems in

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

important applications.

The function of software engineering in the credibility of AI content

Software engineering plays a critical role in guaranteeing the reliability of content provided by artificial intelligence. Let's examine the role that software engineering plays in this area:

1. Algorithm Design and Implementation: Software engineers create dependable and strong algorithms that serve as the foundation for artificial intelligence systems. They must be conscious of data biases that could compromise the reliability of content produced by artificial intelligence.
2. Software engineers should create procedures for training artificial intelligence models, confirming their accuracy, and reducing errors. Trustworthiness may be impacted by difficulties in ensuring that models generalize well to a variety of data sources.
3. Data Quality and Preprocessing: To improve the quality and applicability of input data, software engineers are essential to the process of data preprocessing. Content generated by artificial intelligence (AI) may lose credibility due to biases or inaccuracies in training data.
4. Explainability and Interpretability: To increase trust, software engineers should focus on improving the explainability and interpretability of AI models. It can be difficult to achieve great interpretability, particularly in intricate models like deep neural networks.
5. Ongoing Monitoring and Maintenance: To identify anomalies and guarantee the continued reliability of AI systems, software developers should set up monitoring systems. Constant work is needed to stay on top of changing data patterns and possible adversarial attacks.
6. Ethical Considerations and Governance: By embracing the values of justice, accountability, and transparency, software engineers may help advance ethical AI. As societal norms change, it could be necessary to revise ethical concerns frequently.
7. Security Measures: To guard AI systems against malevolent actions that can undermine confidence, software engineers should put security procedures into place. Constant updates are necessary to preserve the integrity of AI-generated material since cybersecurity threats are always changing.
8. User Feedback and Iterative Improvements: To update AI content iteratively, software engineers should create systems to gather user feedback. It can be difficult to accommodate different user viewpoints and preferences, and this calls for a sophisticated strategy.

GAPS IN RESEARCH

A review of software engineering's contribution to ensuring AI systems' reliability reveals several unaddressed research gaps. Research is first and foremost needed to understand how ethical considerations can be easily incorporated into the AI development process using software engineering approaches. Furthermore, studies ought to concentrate on how software engineering contributes to adhering to new AI rules and guidelines. Inquiries into the gaps in software engineers' education and training about the reliability of AI should also be conducted. By filling in these research gaps, we may build strong software engineering techniques that prioritize and improve the reliability of AI systems, promoting the development of ethical and responsible AI.

METHODOLOGY

For this study, we employed a secondary data analysis methodology to examine the critical function of software engineering in guaranteeing the reliability of AI content. Since the dataset was taken from peer-reviewed publications, it was guaranteed to contain accurate and thorough information relevant to our investigation. We used a strict validation procedure to improve data consistency and reduce any possible

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

biases. Utilizing secondary data offers a productive and economical way to accomplish our research goals, adding significantly to the corpus of computer science knowledge already in existence.

CONCLUSION

The concepts, foundations, and prerequisites that must be met for AI systems to be considered reliable have been clarified in this book. We have given clear definitions for all relevant terms, drawing on well-established regulatory and supervisory frameworks like the AI Act, and we have emphasized the importance of each trustworthiness requirement in promoting user trust in AI-based systems. The conversation moves on to practical issues that are critical to the design, development, and deployment of reliable AI systems, highlighting the significance of assessing their regulatory compliance (auditability) and clarifying the thought processes that go into their decision-making (accountability). Suitability for these two pragmatic requirements is essential for responsible AI systems. We hope that this paper will be an invaluable resource for scholars, professionals, and those just starting in the field of artificial intelligence, especially for those looking for a thorough grasp of reliable AI. For the responsible design and development of AI systems throughout their life cycle, a detailed analysis of the meaning of trust in AI-based systems and the needs that go along with it, as this manuscript presents, is essential. To bridge the gap between technology and regulation, improve scientific research, foster economic prosperity, and eventually benefit humanity—all while adhering to ethical and legal guidelines—reliable AI and responsible AI systems are crucial, particularly in high-risk circumstances.

REFERENCES

1. Shane Legg and Marcus Hutter. 2007. A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications* 157 (2007), 17.
2. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, SihamTabik, Alberto Barbado, Salvador García, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
3. PangWei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In Proceedings of the International Conference on Machine Learning. 1885–1894.
4. Kate Crawford. 2021. *The Atlas of AI*. Yale University Press.
5. ISO 24028:2020. 2020. *Information Technology–Artificial Intelligence–Overview of Trustworthiness in Artificial Intelligence.* International Organization for Standardization.
6. European Commission. 2018. Ethics Guidelines for Trustworthy AI. Retrieved November 2, 2021 from https://ec. europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
7. Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology* 31 (2017), 841.
8. National Institute of Standards and Technology. 2021. NIST Proposes Method for Evaluating User Trust in Artificial Intelligence Systems. Retrieved November 2, 2021 from https://www.nist.gov/news-events/news/2021/05/nistproposes-method-evaluating-user-trust-artificial-intelligence-systems.
9. U. S. Government Accountability Office. 2021. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. Retrieved November 2, 2021 from https://www.gao.gov/products/gao-21-519sp.
10. David 2017. *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency.
11. B. Burke, D. Cearley, N. Jones, D. Smith, A. Chandrasekaran, C. K. Lu, and K. Panetta. 2019. Gartner Top 10 Strategic Technology Trends for 2020-Smarter with Gartner. Retrieved November 2, 2021 from https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020/.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

12. Edelman Trust Barometer. 2019. Edelman Trust Barometer Global Report. Retrieved November 2, 2021 from https://edelman.com/sites/g/files/aatuss191/files/2019-02/2019_Edelman_Trust_Barometer_Global_Report.pdf.
13. H. James Wilson and Paul R. Daugherty. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review* 96, 4 (2018), 114–123.
14. Peter. Naur, Brian Randell, Friedrich Ludwig Bauer, and NATO Science Committee. (Eds.). 1969. Software engineering: report on a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7th to 11th October 1968. Scientific Affairs Division, NATO.
15. Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan. Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 2017. DOI 10.1093/qje/qjx032. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5947971/ .
16. Lepri, Oliver, Pentland. Ethical machines: The human-centric use of artificial intelligence. iScience, 2021. Volume 24, Issue 3, Pages 102249-102249.
17. Stuart J. Russell and Peter Norvig. 2021. Artificial intelligence: A modern approach (fourth edition ed.). Pearson, Hoboken.
18. D. Partridge and Y. Wilks. 1987. Does AI have a methodology which is different from software engineering? Artificial Intelligence Review 1, 2 (1987), 111–120. https://doi.org/10.1007/BF00130012.
19. Mirko Perkusich, Lenardo Chaves e Silva, Alexandre Costa, Felipe Ramos, Renata Saraiva, Arthur Freire, EdnaldoDilorenzo, Emanuel Dantas, Danilo Santos, KyllerGorgônio, Hyggo Almeida, and Angelo Perkusich. 2020. Intelligent software engineering in the context of agile software development: A systematic literature review. Information and Software Technology 119 (2020), 106241. https://doi.org/10.1016/j.infsof.2019.106241.
20. D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean François Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems 2015-Janua (2015),2503–2511.
21. E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2020) 4793–4813.
22. D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, 2017, arXiv preprint arXiv:1710.00794.
23. Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57.
24. European Commission High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019.
25. A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Inf. 3 (2) (2016) 119–131.
26. M. EstévezAlmenzar, D. Fernández Llorca, E. Gómez, F. Martinez Plumed, Glossary of Human-Centric Artificial Intelligence, Tech. Rep. JRC129614, Joint Research Centre, 2022.
27. C. Widmer, M.K. Sarker, S. Nadella, J. Fiechter, I. Juvina, B. Minnery, P. Hitzler, J. Schwartz, M. Raymer, Towards human-compatible XAI: Explaining data differentials with concept induction over background knowledge, 2022, arXiv preprint arXiv:2209.13710.
28. N. Tomašev, J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D.C. Belgrave, D. Ezer, F.C.v.d. Haert, F. Mugisha, et al., AI for social good: Unlocking the opportunity for positive impact, Nature Commun. 11 (1) (2020) 2468.
29. World Economic Forum, Empowering AI Leadership: AI C-Suite Toolkit , Tech. Rep., 2022.
30. World Economic Forum, Empowering AI Leadership An Oversight Toolkit for Boards of Directors, Tech. Rep., 2019.
31. E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, N. Nobani, A survey on XAI and natural language explanations, Inf. Process. Manage. 60 (1) (2023) 103111.
32. L. Floridi, Establishing the rules for building trustworthy AI, Nat. Mach. Intell.1 (6) (2019) 261–262.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

33. A. Ruospo, E. Sanchez, L.M. Luza, L. Dilillo, M. Traiola, A. Bosio, A survey on deep learning resilience assessment methodologies, Computer 56 (2) (2023) 57–66.
34. S. Speakman, G.A. Tadesse, C. Cintas, W. Ogallo, T. Akumu, A. Oshingbesan, Detecting systematic deviations in data and models, Computer 56 (2) (2023) 82–92.
35. N. Díaz-Rodríguez, J. Del Ser, Mark Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera. "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation." *Information Fusion* (2023): 101896.
36. D. Hendrycks, N. Carlini, J. Schulman, J. Steinhardt, Unsolved problems in ml safety, 2021, arXiv preprint arXiv:2109.13916.
37. C. O'neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2017.
38. Public Voice coalition, Universal Guidelines for Artificial Intelligence, 2018, online https://thepublicvoice.org/ai-universal-guidelines/. (Accessed 20 April 2023).
39. Information Commissioner's Office (ICO), How to use AI and personal data appropriately and lawfully, 2022, online https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf.(Accessed 20 April 2023).
40. European Union, Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), 2022.
41. European Union, Proposal for a regulation of the European parliament and of the council on harmonised rules on fair access to and use of data (Data Act), 2022.
42. R. Mariani, F. Rossi, R. Cucchiara, M. Pavone, B. Simkin, A. Koene, J. Papenbrock, Trustworthy AI – Part 1, Computer 56 (2) (2023) 14–18.
43. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
44. K. Haresamudram, S. Larsson, F. Heintz, Three levels of AI transparency, Computer 56 (2) (2023) 93–100.
45. B. Pérez, J. Rubio, C. Sáenz-Adán, A systematic review of provenance systems, Knowl. Inf. Syst. 57 (2018) 495–543.
46. U. Aïvodji, H. Arai, S. Gambs, S. Hara, Characterizing the risk of fairwashing, Adv. Neural Inf. Process. Syst. 34 (2021) 14822–14834.
47. B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.
48. R. Baeza-Yates, Bias on the web, Commun. ACM 61 (6) (2018) 54–61.
49. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.
50. E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650, http://dx.doi.org/10.18653/v1/P19-1355, URL https://aclanthology.org/P19-1355.
51. C. -J. Wu, R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, G. Chang, F. Aga, J. Huang, C. Bai, et al., Sustainable AI: Environmental implications, challenges and opportunities, Proc. Mach. Learn. Syst. 4 (2022) 795–813.
52. R. Schwartz, J. Dodge, N.A. Smith, O. Etzioni, Green AI, Commun. ACM 63 (12) (2020) 54–63.
53. B. Xia, Q. Lu, H. Perera, L. Zhu, Z. Xing, Y. Liu, J. Whittle, Towards concrete and connected AI risk assessment (C2 AIRA): A systematic mapping study, 2023, arXiv:2301.11616.
54. J. Silberg, J. Manyika, Notes from the AI frontier: Tackling bias in AI (and in humans), McKinsey Global Inst. 1 (6) (2019).
55. J. Mökander, M. Axente, F. Casolari, L. Floridi, Conformity assessments and post-market monitoring:

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue IV April 2024

A guide to the role of auditing in the proposed European AI regulation, Minds Mach. 32 (2) (2022) 241–268.
56. ISO/IEC, ISO/IEC TR 24029-1, information technology — Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview, 2021, https://www.iso.org/standard/77609.html.
57. A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (SCS) comparing human and machine explanations, KI-KünstlicheIntell. 34 (2) (2020) 193–198.
58. Z. Han, H. Yanco, Communicating missing causal information to explain a robot's past behavior, ACM Trans. Hum.-Robot Interact. 12 (1) (2023) 1–45.
59. M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, et al., Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020, arXiv preprint arXiv:2004. 07213.
60. A. -H. Karimi, J. von Kügelgen, B. Schölkopf, I. Valera, Towards causal algorithmic recourse, in: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Springer, 2022, pp. 139–166.
61. L. Bourtoule, V. Chandrasekaran, C.A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: IEEE Symposium on Security and Privacy, SP, IEEE, 2021, pp. 141–159.
62. S. Verma, V. Boonsanong, M. Hoang, K.E. Hines, J.P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, in: NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses, ML-RSA, 2020.
63. S. Barocas, A.D. Selbst, Big data's disparate impact, California Law Rev. (2016) 671–732.
64. Mike Thomas. 2019. 6 Dangerous Risks of Artificial Intelligence. Retrieved November 2, 2021 from https://builtin.com/artificial-intelligence/risks-of-artificial-intelligence.
65. Bernard Marr. 2018. Is artificial intelligence dangerous? 6 AI risks everyone should know about. *Forbes* (2018).
66. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica, May 23, 2016.
67. Jethro Mullen. 2015. Google Rushes to Fix Software That Served Up Racial Slur. Retrieved November 2, 2021 from https://www.cnn.com/2015/07/02/tech/google-image-recognition-gorillas-tag/.
68. Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Available at https://www.reuters.com.
69. Puneet Kohli and Anjali Chadha. 2019. Enabling pedestrian safety using computer vision techniques: A case study of the 2018 Uber Inc. self-driving car crash. In *Proceedings of the Future of Information and Communication Conference*. 261–279.
70. W. Fan, J. Liu, S. Zhu, and P. M Pardalos. 2020. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annalsof Operations Research* 294, 1 (2020), 567–592.
71. Anthony 2019. *The Culture of AI: Everyday Life and the Digital Revolution*. Routledge.
72. Gary Marcus and Ernest Davis. 2019. Rebooting AI: Building Artificial Intelligence We Can Trust. Vintage.