Received: 1 March 2023 | Revised: 15 April 2023 | Accepted: 28 April 2023

DOI: 10.1002/spe.3216

SPECIAL ISSUE PAPER

WILEY

Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges

Muhammad Azeem Akbar[1] | **Arif Ali Khan**[2] | **Sajjad Mahmood**[3] | **Saima Rafi**[4] | **Selina Demi**[5]

[1]Software Engineering Department, Lappeenranta-Lahti University of Technology, Lappeenranta, Finland

[2]M3S Empirical Software Engineering Research Unit, University of Oulu, Oulu, Finland

[3]Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

[4]Department of Informatics and Systems, University of Murcia, Murcia, Spain

[5]Faculty of Computer Sciences, Østfold University College, Halden, Norway

Correspondence
Muhammad Azeem Akbar, Software Engineering Department, Lappeenranta-Lahti University of Technology, 53851 Lappeenranta, Finland.
Email: azeem.akbar@lut.fi

Abstract

The significance of artificial intelligence (AI) trustworthiness lies in its potential impacts on society. AI revolutionizes various industries and improves social life, but it also brings ethical harm. However, the challenging factors of AI trustworthiness are still being debated. This research explores the challenging factors and their priorities to be considered in the software process improvement (SPI) manifesto for developing a trustworthy AI system. The multivocal literature review (MLR) and questionnaire-based survey approaches are used to identify the challenging factors from state-of-the-art literature and industry. Prioritization based taxonomy of the challenges is developed, which reveals that lack of responsible and accountable ethical AI leaders, lack of ethics audits, moral deskilling & debility, lack of inclusivity in AI multistakeholder governance, and lack of scale training programs to sensitize the workforce on ethical issues are the top-ranked challenging factors to be considered in SPI manifesto. This study's findings suggest revising AI-based development techniques and strategies, particularly focusing on trustworthiness. In addition, the results of this study encourage further research to support the development and quality assessment of ethics-aware AI systems.

KEYWORDS

challenges, multi-vocal literature review, questionnaire, SPI manifesto, trustworthy AI software

1 | INTRODUCTION

Trustworthy AI refers to ethical considerations when designing, developing, and using artificial intelligence (AI) systems. Since AI could have a huge impact on society and individuals, it is important to be created and used ethically and responsibly. AI automates tasks, makes decisions, and performs various functions.[1] This means that the ethical considerations of AI have the potential to affect a large number of individuals and businesses.

Abbreviations: AI, artificial intelligence; ART, accountability, responsibility, and transparency; AIED, artificial intelligence in education; EAD, ethically aligned design; ECCOLA, ethical considerations for commercializing AI; IEC, international electrotechnical commission; ISO, international organization for standardization; MLR, multivocal literature review; SPI, software process improvement.

In the industry, there is an ever-growing push to develop policies to counter potential ethical damage caused by AI applications.[2] However, there is still disagreement over what constitutes "ethical AI" and "which ethical prerequisites, guidelines, and technical standards are required for its fulfillment. According to Greene et al.,[3] the ethics of AI systems are "up for grab," or open to initiatives. These initiatives provide objectives and definitions for what ethical AI systems should be capable of according to Vakkuri et al.,[4] incidents of misuse and failure of AI systems in the actual world show the need for and conversation on AI ethics. Furthermore, the ethical analyses of AI technologies noted that autonomous systems and AI should not merely be seen as technological endeavors. There is a widespread argument that cultural and ethical norms impact the creation and use of AI-based systems.[5] In addition to technical work, developing AI-based systems needs to consider political, intellectual, economic, legal, and sociological concerns.[5] These systems greatly impact people's values and cultural norms.[5]

Within an organization, implementing and enforcing policies can pose significant challenges. While numerous stakeholders advocate for ethical AI, the exact consequences and benefits of embracing or neglecting ethical efforts and commitments are still being determined. In the AI industry, єethically driven self-commitmentsε made by businesses and research organizations impede the establishment of a legally binding framework, leaving AI ethics regulations largely vague and superficial.[6] Many prominent businesses, organizations, and communities have declared their commitment to ethical principles, as suggested by Greene et al.[3] However, the resulting value statements often raise more questions than they provide answers. A possible reason for this could be that businesses do not fully recognize the severe consequences of neglecting to uphold and implement ethical principles in AI development.

AI practitioners and industries should have a firm grip on this field's ethics. Press and public views on AI ethics have recently supported substantial relevant research.[3] However, the subject must be thoroughly researched academically and in real-world environments.[5] Although there have been few academic studies on this subject, most AI practitioners are still unaware. It was noted in the IEEE's ethically aligned design (EAD)[7] principles that industrial-scale ethics in AI is still in its infancy.[8] The fact that the AI business knows little to nothing about ethics suggests that further theoretical and applied study in required.

Moreover, current guidelines and principles tend to be broad and require more detailed information on specific areas to foster the development of ethically robust AI systems. This research study aims to investigate factors that could negatively impact the ethical AI-based software development process, which practitioners should consider when developing ethically sound AI systems. We expect that this comprehensive study will contribute to a knowledge base that both research communities and practitioners can utilize to develop innovative and practically reliable methodologies and procedures for creating morally sound AI-based systems. Consequently, this study addressed the following research questions:

RQ1. What are the trustworthy AI challenges against the SPI manifesto reported in multivocal literature? ■

RQ2. What are the most important challenges to be considered in SPI manifesto categories for developing ethically trustworthy AI systems? ■

RQ3. What would be the ranked-based taxonomy of trustworthy AI system challenges? ■

The remainder of this article is organized as follows: Section 2 delves into the study background; Section 3 outlines the research design; Section 4 presents the results and discussion; Section 5 offers a summary of the study findings. The limitations of this research are addressed in Section 6, while Section 7 discusses the study's implications. Finally, Section 8 provides the conclusion and recommendations for future research.

2 | BACKGROUND

AI applications have spurred a technological revolution, transforming both science and society. This shift in human-to-machine power dynamics has prompted significant societal discussions surrounding guiding principles and regulations for the use and implementation of AI systems.[9] Numerous organizations have formed ad hoc committees to establish AI ethical policy guidelines, resulting in the creation of various AI guidelines and regulations.[9] In 2018, technology companies such as SAP and Google released public rules and standards for AI-based systems.[9]

Similarly, Access Now, Amnesty International, and the Association of Computing Machinery (ACM) have created guidelines and suggestions for AI technologies. The trusted AI European Commission's standards were created

to encourage legal, moral, and reliable AI systems.[10] The Obama administration's "preparing for the Futures of Artificial Intelligence" report thoroughly examined current AI research, its uses, and societal effects.[11] The report also included suggestions for the next AI-related initiatives. The "Beijing AI Principles" guidelines[12] established several recommendations for the governance use and development of AI. These guidelines offered a framework with an emphasis on AI ethics.

The EAD guidelines,[7] released by IEEE, the largest technical professional organization in the world, presented a framework to address the moral and technical values of AI systems based on rules and principles. The EAD framework comprised eight broad concepts: well-being, human rights, efficacy, accountability, data agency, awareness of misuse, and integrity. These principles served as a foundation for the design and implementation of AI-based systems. A standard for AI is also being developed by groups like ISO and IEC.[11] A combined ISO/IEC international standard group called JTC 1/SC 42 focuses on the complete artificial intelligence ecosystem, including standardization, AI governance, trustworthiness, and computational approach.[13] Guidelines, methods, and strategies are required for handling ethics in a way that satisfies organizational priorities, and this is determined by the effort made by various organizations to create AI ethics. However, according to recently published studies, current AI ethics standards could be more useful and widely used than before.[14] This is clear from empirical studies by McNamara et al.[15] examining how the ACM code of ethics affected the decision-making of software development. The study's findings showed that the ACM code of ethics had no bearing on moral judgments. It is difficult to implement the guidelines successfully due to the lack of efficient methods.[14]

To develop the ethics standards for reliable AI systems, the European Commission organized a high-level expert group on artificial intelligence (AI-HLEG).[15] They emphasized the seven essential core aspects (technical robustness and safety, human agency and oversight, transparency, privacy, and data governance, non-discrimination, accountability, and societal and environmental well-being). For the purpose of creating an ethically sound AI system, Jobin et al.[9] conducted a mapping study and clarified the significance of embracing the ethics principle.

Vakkuri et al.[14] explored ethical issues in the context of AI using the accountability, responsibility, and transparency (ART) framework and conceptual model. They conducted case studies are conducted to empirically validate the conceptual model. The empirical findings showed that AI ethics principles are still not being used, but some well-known ideas, such as documentation, are being considered. Moreover, the results showed that professionals take the societal impact of AI systems into account.[14] Another model, ECCOLA was developed by Vakkuri et al.[16] using a cyclical action design research approach. This model is developed to guide the developers and managers to implement the core AI ethics practically during software development. Thus, the method is developed based on the RESOLVED study, the essence theory of software engineering, the existing theoretical and conceptual research, and the ethical guideline developed by higher organizations. Therefore, ECCOLA is divided into eight themes (taken from the AI ethical guidelines), 21 cards, and 1–6 cards are included in each theme, following the theory of the essence of software engineering. Each card is categorized into three sections: inspiration (i.e., why this is significant), suggestions for action (to address the issue), and an actual instance of the subject (to make the issues more tangible). The cards further contain a section for note-making purposes. Holmes et al.[17] developed an artificial intelligence in education (AIED) framework. First, they introduced concerns about the morality of using AI in education. Second, they addressed the crucial points presented by the 17 responders and summarized their contributions. Based on the expert's opinions, Holmes et al.[17] created a framework for discussing AIED ethics based on the perspectives of the experts, combining a multidisciplinary approach with a set of strict standards that seem essential in this situation.

Several other AI-specific designs have been released, for instance, a Microsoft nine-step pipeline,[18] a five-step "stairway to heaven" AI model,[19] and a maturity framework for the AI process.[20] However, they are not primarily concerned with creating high-quality or morally upright AI systems. Although these models portray processes in certain organizational contexts, there aren't any models that SMEs and beginning businesses may use.[21] There are no frameworks, methods, or tools to guide the organizations conceding to the influencing factors (negatively and positively) in AI system development. Furthermore, Khan et al.[22] explored ethical AI systems' principles and influencing factors in a systematic literature review study. Khan et al.[22] study focused on understanding the impact of ethics challenges on AI principles. In another empirical study, Khan et al.[23] discovered that the most prevalent AI ethics difficulties are a lack of ethical expertise, oversight bodies, and legal frameworks.

To leverage the benefits of ethical AI in system development, there is a need to design tools or frameworks that will guide organizations toward developing trustworthy, sound AI systems. This study provides insights into AI ethics challenges through a multivocal literature review and empirical investigations. The study uncovered nineteen critical

challenges that are the main hurdles to ethical AI systems. The gap between AI ethical systems and their adoption in practice is covered by finding the challenges hindered in performing AI software development tasks. Furthermore, these challenges were explained briefly with the SPI concept. The SPI manifesto is used to emphasize critical factors to improve in terms of software processing during the development of AI systems. The SPI manifesto prioritized the values of people, business focus, and a belief that organizational change is at the core of software process improvement. Therefore, this study uncovered the ongoing challenges, and mapping with the SPI manifesto helped reveal the importance of software process improvement in AI systems while managing ethical challenges.

3 | RESEARCH METHODOLOGIES

To achieve the objective of this study, we employed a hybrid research methodology. Firstly, we conducted a multivocal literature review to survey existing literature, following the guidelines provided by Garousi et al.[24] In the second step, we carried out a questionnaire survey study aimed at validating the findings of the multivocal literature review with industry practitioners.[25] Lastly, we utilized the fuzzy AHP approach to rank the identified challenges based on their significance for adopting ethics in software process improvement initiatives. This combination of methods has been used in various software engineering research studies.[26–29]

3.1 | Multivocal literature review

MLR[24] gives that it is a kind of literature review that offers perspectives from both the state-of-the-practice and state-of-the-art to address the research objectives in this study.[24,30,31] We reviewed the formal (conferences, peer-reviewed journals, etc.) and informal (white papers, blogs, industry standards, videos, etc.) literature as part of MLR. An overview of our research process is shown in Figure 1.

3.1.1 | Planning the review

Before beginning the review process, a thorough MLR protocol was created following the MLR recommendations.[24] All study team members completed the MLR phases. Data is gathered using the MLR approach from grey literature and formally published literature. Both data extraction methods are covered in detail in the following sections in Figure 1.

3.1.2 | Executing the review

To extract and consider the literature for data extraction, the following protocols were used:

***Search string*:** By combining the keyword and its alternatives, we developed the search strings according to the criteria given by Zhang and Babar.[32] The search strategy is based on the following steps:

1. From population, intervention, and result, derive the key concepts.
2. Look for alternate spellings and synonyms for the derived phrases.
3. Search several academic resources and search engines to validate these terms.
4. The Boolean operators were utilized (and, if permitted, the AND operator was used to connect the key components of the search phrase; OR operator was considered to combine synonyms and similar spellings if permitted).

Based on the above search strategy, the search string elements are provided in Table 1.

***Used digital platforms*:** To extract the relevant and potential literature, academic digital repositories were search using the developed search string (formal literature\pre-reviewed primary studies) and search engines (grey literature). The explored vanes for both types of data are presented in Figure 2.

Additionally, to find the needed data, we also used backward and forwarded snowballing tactics.[33,34] Studies that quoted the paper are referred to as εforward snowballing,ε whereas studies that are mentioned in the article are referred to as εbackward snowballingε (reference list of the paper).[35,36]

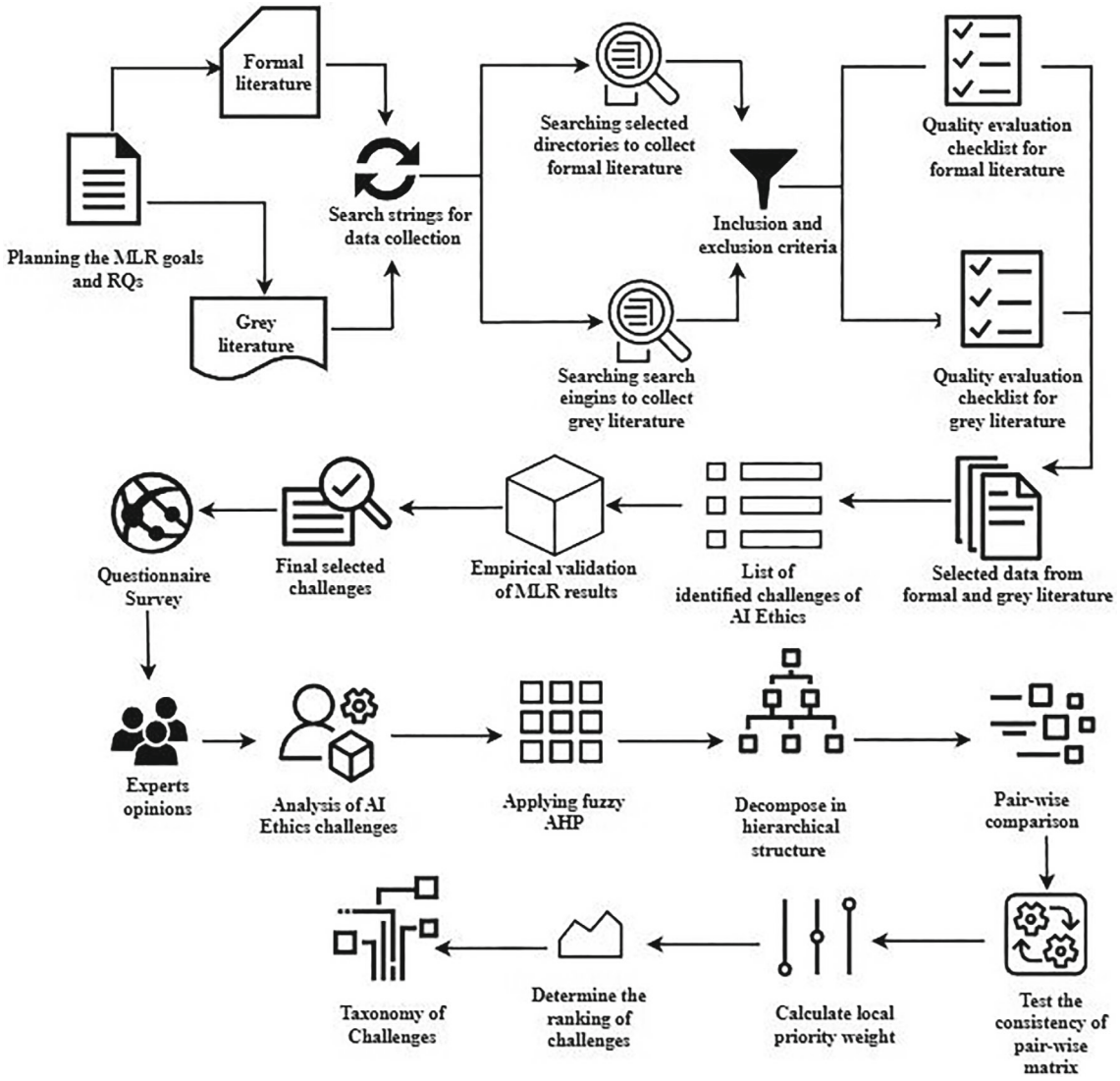

FIGURE 1 Overview of research methodology.

TABLE 1 Search terms.

Particular	Search terms
SI1 (Factors)	*Challenges*: ("barriers" OR "obstacles" OR "hurdles" OR "difficulties" OR "impediments" OR "hindrance" OR "challenges" OR "limitations")
SI2 (Intervention)	*Artificial intelligence*: ("artificial intelligence" OR "AI" OR "pattern recognition" OR algorithms)
	Machine learning: ("machine learning" OR "predictive analytics" OR "pattern recognition" OR "deep learning."
SI3 (Intervention)	Ethics OR "human rights" OR "human values" OR "responsibility" OR "human control" OR "fairness" OR discrimination OR nondiscrimination OR "transparency" OR "explainability" OR "safety and security" OR "accountability" OR "privacy".
SI4 (Population)	"Software development community," "software operation community," "policymakers, lawyers," "observers"
SI5 (Experimental)	*Formal literature*: ("grounded theory," OR "interviews," OR "case studies," OR "questionnaire survey," OR "theoretical studies," OR "content analyses," OR "action research").
	Grey literature: ("Videos," OR "Blogs," OR "white papers," OR "expert reports," OR "industry standards," OR "tweets," OR "website Q&A").
"Final search string =	(SI1) AND (SI2) AND (SI3) AND (SI4) AND (SI5)"

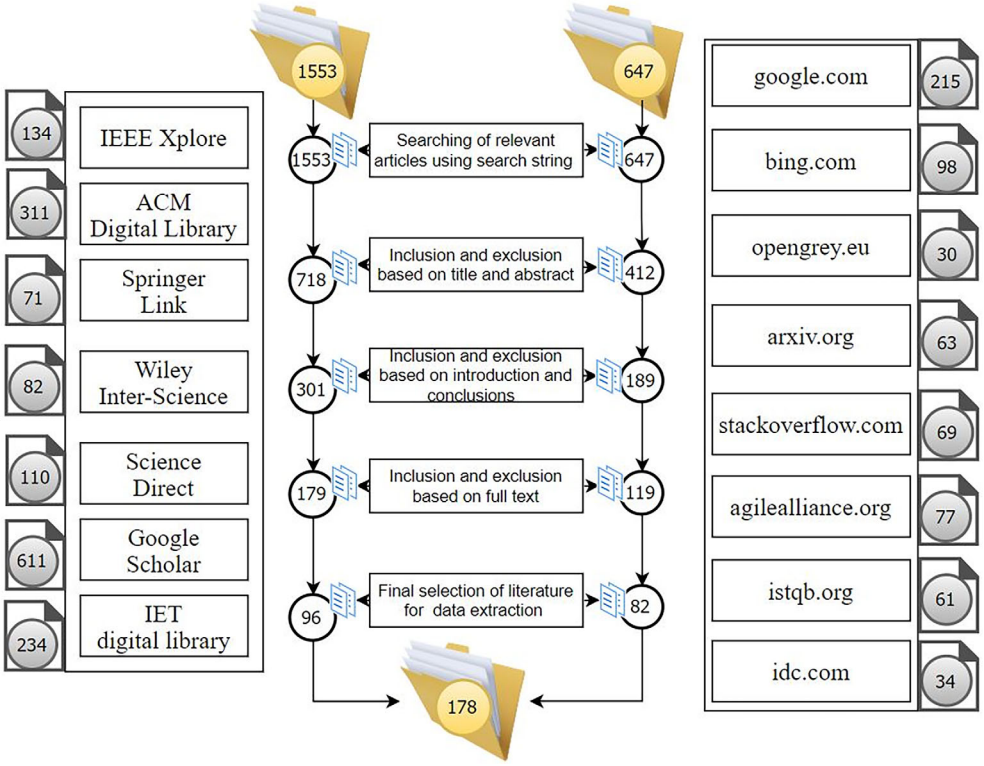

FIGURE 2 Final selected primary studies.

***Inclusion criteria*:** We utilized the inclusion criteria created in accordance with the recommendations provided in References 24,37–39. The main components of the inclusion criterion are: (i) literature should be about AI ethics; (ii) literature should discuss the influencing factors of ethical AI, the social impact of AI, and AI principles; (iii) literature of the study should be industry- oriented; (iv) literature provides contextual information about the subject under investigation; (v) The value of literature is clear to both academic and industrial researchers.

***Exclusion criteria*:** Based on the recommendations in References 36,39,40, we used the exclusion criteria to weed out irrelevant material. The main exclusion criterion points are: (i) studies that do not have implications for researchers and industry experts; (ii) literature not based on primary evidence; (iii) studies written in a language other than English were excluded.

***Determination of literature quality (LQ)*:** The chosen studies' quality was evaluated using standards derived from prior SLR software engineering studies.[24,37,38,41,42] The evaluation standards include a checklist that comprises LQ assessment standards for formal literature and grey literature in Table 2. Each selected piece of literature was assessed against the assessment criteria questions using the Likert scale "fully answered = 1, partially answered = 0.5, no answered = 0". The final LQ score of each source of formal and gray literature is provided in Appendix A.

***Literature selection*:** The most pertinent studies were chosen using three processes. The methods used to choose the studies are based on advice from Reference 36. Initially, seven research were manually selected while considering QGS recommendations.[43] The hand-chosen studies had a direct bearing on the study's topic. Moreover, we extracted the material from digital repositories using the search strings. Following the inclusion and exclusion criteria application, 1553 papers were pulled from academic repositories in response to the execution of a search string for formal data collection. Using inclusion and exclusion criteria, we gathered 647 items of grey literature while running the search string through search engines. To further refine the literature source set, we applied the five steps of the tollgate approach,[31] and 96 formal studies and 82 grey literature sources were considered for data extraction. The complete literature source (formal and grey literature) is provided in Appendix A.

***Data extraction*:** We used the coding approach[44] to extract the data from selected literature to address our study's research questions. The chosen data was labeled, sorted, and classed according to the specified contributions, concepts, ideas, or discoveries. With a thorough and persuasive study of qualitative data sets, the elements in the analytical data were discovered using the coding approach.

TABLE 2 LQ assessment criteria for formal literature.

S. No	Criteria for formal literature
Q1	Is the selected literature an empirical study?
Q2	Is the aims and objective of the study clear and rational?
Q3	Is the description of the research context adequate?
Q4	Is the research design appropriate to address the aim of the research?
Q5	Is the description of the sample used and the methods of identifying and recruiting the sample adequate?
Q6	Are the data collection methods appropriate and adequately described?
Q7	Are the data analysis methods adequately described and grounded in the data?
Q8	Is the relationship between the researcher and participants adequately considered?
Q9	Are the findings clearly stated with credible results?
Q10	Is the study valuable for research or practice?
S. No	**Criteria for grey literature**
Q1	Is the publishing organization reputable? For example, the Software Engineering Institute (SEI)
Q2	Is an individual author associated with a reputable organization?
Q3	Does the author have expertise in the area? (e.g., job title principal software engineer).
Q4	Does the source have a clearly stated aim?
Q5	Does the source have a stated methodology?
Q6	Are any limits clearly stated?
Q7	Does the work refer to a particular population or case?
Q8	Does the work seem to be balanced in the presentation?
Q9	Is the statement in the sources as objective as possible? Or is the statement a subjective opinion?
Q10	Does the item have a clearly stated date?

TABLE 3 Kendall's coefficient of concordance test.

"Data set"	"Kendall chi-squared"	"df"	"Subjects"	"Raters"	"*p*-value"	"W"
CGSD	35.434	14	10	3	0.001267	0.906761

We performed an inter-rater reliability test after completing the data extraction process to look at the interpersonal bias. To do this, we consulted three impartial specialists, and they chose ten data sources (5 published primary studies and five sources of grey literature). They carried out each step of data extraction and data collection. We determined Kendall's nonparametric coefficient of concordance (W) to measure the inter-rater agreement[36] between the data extraction team and independent experts. As indicated in Table 3, the outcome, $W = 0.91$ ($p = 0.002$), demonstrates agreement between the independent experts and the data extraction team.

***Data synthesis*:** To address this study's research objective, we identified the factors that could negatively and positively impact the development of ethically sound AI systems. We also identified the social influencing areas caused by unethical AI systems. Moreover, the fundamental process areas (principles) that software development firms should consider when developing AI-based software.

3.1.3 | Reporting the review

***LQ assessment score*:** The quality of the selected data sources (primary studies and grey literature) was assessed to ensure each data source's effectiveness in addressing our study's research questions. The LQ assessment for both formal and grey literature is given in Table 2. The QA results showed that 78% of the primary studies scored greater

than 70%, and 80% of grey literature scored at least 70%. The quality assessment score indicated that the selected literature set effectively addresses this study's research questions. The detailed LQ assessment results are provided in Appendix A.

3.2 | Empirical research

The empirical study described in this paper aimed to characterize and assess the relationship between the key categories of the SPI manifesto and the MLR findings (challenges) from the perspective of the AI-based software development community and policymakers. This investigation aims to benefit the software development community to develop ethically sound AI systems and advance the body of evidence in this area. As indicated by the research objective and the RQ2, the nature of our study is εexploratory".[45] Hence, to gather data, we created and implemented an online questionnaire-based opinion survey. Below, we go over the survey's planning and execution.

3.2.1 | Survey instrument development

Based on the information gathered from the MLR, we created an online survey (Section 4.1). We formulated a draft set of questions to cover the challenges concerning AI ethics adoption in the SPI manifesto. The first draught of the survey questionnaire was created using our research team's expertise[26,28,46] and by taking into account reports on questionnaire development guidelines, various survey guidelines, and experience reports, for example, References 47–49. The developed questionnaire was sent to five industrial experts as a pilot assessment. The purpose of this stage in our survey design was to ensure the participants were familiar with the terms utilized in our questionnaire. This is because scholars and industrial practitioners frequently employ slightly different terminology, as other studies have revealed, such as References 26,28,46. Discussing and agreeing on consistent terminology is important when conducting joint studies or projects. Although their responses to the survey questions were not included in the survey data, the industrial practitioners' opinions helped finalize the survey questions' collection. Initial expert feedback was beneficial for assuring the survey's appropriate context and industrial respondents' familiarity with its vocabulary.

We had 34 questions after refining them in response to the pilot study's industry practitioners' feedback. Appendix B contains a list of all the questions asked during the survey. The first 12 questions were used to collect information about the projects, participants, and companies' profiles and demographics. The majority of these inquiries had quantifiable, already created single or multiple responses. For instance, participants could give several answers to the second question about their present job because they might have more than one role within their employer organizations. However, the first question regarding participants' gender needed a single response answer (male or female).

The remaining queries focused on information about AI ethics adoption in the SPI manifesto in the context of the challenges. To get the opinions of survey participants, we used the five scales Likert scale from (strongly disagree to strongly agree). In the five-scale Likert scale, we used neutral as an important option due to its vital role in collecting unbiased data.

3.2.2 | Recruitment of subjects and survey execution

The Google Forms tool (docs.google.com/forms) was utilized to create the survey instrument, while Google Drive (drive.google.com) hosted it online. The survey received approval from LUT University's Research Ethics Board before initiating the data collection process. The survey was made available to participants from February to April 2022 following ethics approval. Participants were asked to complete the survey questionnaire voluntarily and anonymously, with the option to withdraw at any time.

To ensure data collection from as many real-world practitioners as possible, we designed and implemented a PR and execution plan. We sent email invitations to 63 software industry partners and connections within our network (and their departments). Utilizing the snowballing principle, we asked survey respondents to share the invitation email containing the survey URL with their contacts[50] to increase the survey's reach. Consequently, we could track the progress weekly but needed to determine the response rate of practitioners who received the email and completed the survey. Furthermore, we extended public invitations to the software engineering community by emailing

the management offices of over 16 Technology Development Zones, Research Finland, and posting messages on social media platforms such as Facebook, Twitter, and LinkedIn (where we found 22 software engineering-related LinkedIn groups).

As mentioned earlier, we created and adhered to a publicity and execution strategy for our survey. To avoid burdening practitioners with multiple duplicate invitations, we shared our invitation with each group through a single point of contact. Our plan included each organization's target industries, contact information, publicity schedule, and status. Since participation in our survey was anonymous, we did not track response rates. However, an iterative publicity schedule (with three key iterations) allowed us to observe changes in the survey population and estimate response rates after each iteration.

It is important to note that, similar to previous online surveys on software engineering, such as those conducted in References 26,28,46, several participants opted not to answer some questions. This is a common occurrence in virtually all online surveys for software engineering. We found that partial responses to survey questions—answers that did not fully address all the questionnaire's items—still provided valuable insights. During the data collection process, we received 156 complete responses, which were used for data analysis. To ensure data replicability and allow other researchers to conduct further studies on our dataset, we will make the data available upon request.

3.2.3 | Data analysis

At this step, we only analyzed the complete 156 responses. We have effectively applied the frequency analysis method to analyze the descriptive data types.[51] Comparatively analyzing the survey variables, it calculates the degree of agreement among survey respondents using the chosen Likert scale. Numerous other researchers have used similar data analysis techniques.[37,52–55]

4 | RESULTS

4.1 | RQ1 (MLR results)

The aspects under investigation were also divided into the εPeople,ε εBusiness,ε and εChangeε subcategories of the software process improvement (SPI) manifesto (Figures 3 and 4). The software engineering experts developed the SPI manifesto concept to execute and formally improve software development activities. Due to the technological revolution and the implication of AI in each field of life (healthcare, banking, etc.), ethics is an important subject to be addressed. The SPI covers the core domain of the software development process. Thus, we mapped the identified challenges, motivators, social impact areas, and core process areas against the core categories of SPI.[56] The mapping team comprises three individuals, which include the prior two authors of this work and two outside industry professionals from the software development companies εVirtual-Forceε and εFin-Tec.ε The experts have experience in AI-based software development and requirements engineering for AI-related software. We mapped the identified problems, social influencing factors, motivators, and process areas against the SPI manifesto's key categories, taking into account the concepts drawn from the literature (by the prior two study authors) and based on the practical knowledge of industry professionals. This mapping aims to elaborate on the key aspects that need to be considered by a particular software development expert toward the development of ethically sound AI systems. The identified challenges against each category of the SPI manifesto are presented in Table 4. In what follows, we explain the set of AI-ethics challenges identified from the multivocal literature review (See Table 4):

4.1.1 | Lack of responsible and accountable ethical AI leader (C1)

The importance of assigning leadership roles accountable for the ethical aspects of AI applications lies in ensuring that these aspects get the attention they deserve. However, a recent report on ethical [GL1] revealed the lack of such roles in 47% of organizations, indicating the negligence of ethical aspects. This negligence may lead to potential risks, which can be prevented with a proactive attitude towards ethics. Organizations should adopt a proactive approach by assigning a

FIGURE 3 SPI manifesto core categories.[56]

FIGURE 4 SPI manifesto fundamental principle.

Chief Ethics Officer [GL2, GL3] or an internal/external board [GL3] to monitor ethical issues related to AI systems and offer advice. In addition, organizations should adopt complementary approaches, such as establishing communication channels with industry and public oversight groups to report ethical issues and share best practices [GL3].

4.1.2 | Lack of human agency and oversight (C2)

AI systems should promote human autonomy, agency, and oversight in order to ultimately reach the objective of functioning as enablers in an egalitarian and democratic society [GL1]. Impact assessments should be carried out prior to the development of AI systems to ensure that they do not compromise users' rights and human autonomy [GL1]. Moreover, governance tools, such as human-on-the-loop, human-in-the-loop, and human-in-command techniques, should be used to offer human oversight of AI systems [GL1]. Given that governance is a complex, integrated discipline that considers different dimensions, such as business, organizational, and IT processes, organizations may use the expertise of digital ethicists to define suitable governance approaches that address the ethical ramifications of AI systems [GL4].

TABLE 4 Challenges for trustworthy AI.

SPI manifesto	Challenges
People	Lack of responsible and accountable ethical AI leader (C1)
	Lack of human agency and oversight (C2)
	Lack of scale training programs to sensitize the workforce on ethical issues (C3)
	Lack of cultural sensitivity determination (C4)
	Lack of ethical knowledge (C5)
	Moral deskilling & debility (C6)
Business	Lack of ethical responsibility parameters (C7)
	Difficult to scope the ethical issue (C8)
	Lack of privacy standards (C9)
	Lack of AI systems ethical impact determination (C10)
	Data governance issues (C11)
	Lack of diversity consideration (C12)
	Lack of quality data (C13)
	Lack of assessment framework (C14)
Change	Lack of transparency of AI tools (C15)
	Lack of clarity on AI decisions (C16)
	Lack of ethics audits (C17)
	Lack of inclusivity in AI multistakeholder governance (C18)
	Lack of legal frameworks (C19)

4.1.3 | Lack of scale training programs to sensitize the workforce on ethical issues (C3)

Previous survey studies [GL1] pointed out the lack of relevant training programs for AI developers as a significant factor that leads to ethical concerns in AI systems. Therefore, organizations should invest in training programs, such as cognitive bias, data bias, and human-centered design, that aim to raise awareness about ethical issues in AI. The scope of these programs should be extended beyond AI developers to users and management. The ultimate goal of education and training programs should be to boost the ethical mindset of all the stakeholders of AI systems, for example, designers, developers, users, and society at large [GL3]. The knowledge regarding the impact of AI systems should escalate across society by ensuring the appropriate training of AI ethicists.

4.1.4 | Lack of cultural sensitivity determination (C4)

Cultural factors influence how users respond to incorporating AI systems into their daily lives [GL4]. For instance, experience has shown that AI-enabled surveillance technologies have been more accepted in countries like China than in America [GL4]. Hence, digital ethicists must understand cultural influences on users' responses to AI systems beyond acquiring technical knowledge.

4.1.5 | Lack of ethical knowledge (C5)

Given the ethical complexity of AI systems, it is not surprising that the lack of ethical knowledge contributes to the immaturity of this emerging field in practice [GL5]. While governmental support for establishing ethics in AI is

essential, organizations perceive the inability of governments to provide ethical experts [GL6]. Moreover, there is a need for additional standards and frameworks to guide the ethical development of AI systems.

4.1.6 | Moral deskilling & debility (C6)

AI systems aim to complement or replace humans in making decisions. The increased decision-making capability of machines may lead to humans becoming deskilled and debilitated at making decisions [GL7]. To prevent moral deskilling and debility, emphasis should be put on ethical issues in building AI systems and ethical training.

4.1.7 | Lack of ethical responsibility parameters (C7)

It is unclear who holds responsibility in the case of unforeseen events that may occur, particularly in AI-enabled safety-critical systems. For instance, if an AI-enabled autonomous taxi causes a pedestrian's life loss, it is not trivial where the liability lies, for example, rideshare company, manufacturer, or programmer [GL2]. Companies should get ready for these situations before AI systems are developed.

4.1.8 | Difficult to scope the ethical issue (C8)

The complex nature of AI algorithms complicates understanding these systems' decision-making logic. In the case of hazardous decisions made by AI systems, humans should be able to explain the reasons behind these decisions. The explainability of such systems is not trivial, as often conclusions are drawn based on humans' interpretation rather than reality [GL2].

4.1.9 | Lack of privacy standards (C9)

AI systems pose a significant threat to privacy as they collect vast user data [GL2]. Therefore, preventive measures should be taken to protect humans' privacy rights. These measures encompass data security standards established by regulators [GL2] and governance strategies focused on privacy issues that should be formulated by digital ethicists [GL4]. On a larger security scope, AI systems may introduce new vulnerabilities due to their self-modifying nature [GL4]. In addition, anyone can potentially access and modify illegitimate AI algorithms, which are often open-source. Unethical hackers may also inject malicious code into organizational operations. These security and privacy issues should be considered and addressed by digital ethicists.

4.1.10 | Lack of AI systems ethical impact determination (C10)

Data protection impact assessments are necessary for the context of AI-enabled data-intensive applications, yet they do not cover ethical and societal issues of AI systems. Incorporating ethical and societal perspectives into human rights impact assessments has been proposed in the literature [GL8]. A self-awareness questionnaire and an ad hoc committee comprise the Human Rights, Ethics, and Social Impact Assessment (HRESIA), presented by Mantelero [GL8]. This broader perspective may raise awareness among data controllers regarding the importance of considering social and ethical aspects of data use [GL8]. While assessing the ethical impact before developing AI systems is essential, it is also necessary to monitor the system and identify issues at early stages [GL9].

4.1.11 | Data governance issues (C11)

Organizations should adopt adequate data governance mechanisms to ensure data quality, integrity, and legitimate accessibility [GL1]. Moreover, organizations should establish a governance body that enables employees with

the necessary means, for example, internal hotlines or channels, to outline emerging issues with AI systems. The governance body should also train employees and partners and communicate with customers regarding ethical issues [GL1].

4.1.12 | Lack of diversity consideration (C12)

The lack of diversity in the AI domain raises ethical concerns because AI algorithms are trained on specific, non-inclusive datasets that may lead to incorrect AI outputs [GL10]. For instance, male dominance in AI has been observed in academic settings [GL1, GL10], and this may have affected how AI systems have been built or how ethical issues have been approached [GL10]. To avoid unpredictable outcomes from AI systems, it is necessary to make inclusive teams with diverse workforces regarding gender, demography, education, and viewpoints [GL1]. This diverse workforce may contribute to the inclusive design of AI systems.

4.1.13 | Lack of quality data (C13)

Ethical issues range from data protection and accuracy to misuse, lack of trust, and negative influence on democracy [GL11]. Privacy and data protection are among the most discussed ethical issues in literature [GL11]. There are three core data protection issues related to AI systems, as follows: (i) training requires large datasets, the access to which may trigger data protection concerns, (ii) it is possible to detect patterns without the need to access personal data, and (iii) it is possible to re-identify anonymized personal data. Existing data protection regulations have yet to consider these emerging AI-related issues [GL11].

4.1.14 | Lack of assessment framework (C14)

Incorporating AI ethics into systems' development raises tensions among principles requiring making political decisions [GL12]. Conducting a risk-benefit analysis to support the decision-making process has been recommended. A risk assessment approach can contribute to predicting ethical issues and their consequences [GL9]. To guide the implementation of such an approach, it may be useful to adopt a methodology or framework.[12] In order to ensure the development of moral and reliable AI systems, the European Commission developed an assessment list in the report εEthics guidelines for trustworthy AIε [GL3]. The responsible department should monitor and update such a list to address technological and regulation-related changes. In addition, the department should update AI systems' standards and policies and ensure adherence to the current regulatory framework and organizational values [GL3].

4.1.15 | Lack of transparency of AI tools (C15)

According to the guidelines for trustworthy AI established by the European Commission [GL3], AI systems should ensure transparency over a set of components involved, such as data, systems, and business models. However, in practice, it has been reported that executives need to be more certain about the transparency of their AI systems [GL1]. Transparent AI systems are those systems that operate in a clear, consistent, and understandable manner, or in simplified terms, systems that can show how they work [GL1]. Recent reports indicate that transparency of AI systems reduced in 2020, with only 59% of organizations informing users regarding the impact of AI decisions, compared with 73% of organizations surveyed in 2019 [GL1]. Two main factors can explain this result: (i) there is an inevitable trade-off between transparency and guarding trade secrets and business practices. The latter needs to improve the full transparency of AI systems. Unsurprisingly, the report [GL1] found out that banks ranked the lowest in ensuring transparency of their AI systems. (ii) The increased complexity of AI models makes them difficult to intelligible to humans [GL1, GL13].

4.1.16 | Lack of clarity on AI decisions (C16)

The increased complexity of AI systems makes them less intelligible to users and reduces the transparency of how AI decisions are made. This reduced transparency has been perceived by customers, as indicated by a recent survey [GL1]. This survey [GL1] revealed that only 62% of customers in 2020 believed that organizations informed them about the decision-making process of AI systems, compared with 77% in 2019. This data suggests the need for increased clarity on AI decisions.

4.1.17 | Lack of ethics audits (C17)

The capability of AI systems to be evaluated in terms of their systems' algorithms, design processes, and data is known as appropriateness, according to the European Commission.[3] While the system's suitability does not imply complete transparency on business models and AI systems' Intellectual Property, it is also true that it requires traceability through logging mechanisms since the early design stages of AI systems. Organizations should conduct ethical audits on their AI systems, which refer to systems and business processes inspections to ensure adherence to ethical requirements [GL14]. The goal of such audits is to detect potential improprieties. For instance, ethical audits may unveil intra-organizational illegal activities, such as breaching workplace safety regulations or employees' excessive working hours [GL15]. In addition, ethical audits can detect behaviors that may not breach law requirements but are still considered unacceptable in a workplace environment [GL16]. Despite ensuring the lack of prohibited practices, ethical audits ensure the application of policies, procedures, and codes of conduct in the organizational environment. There should be consistency between business regulations and value statements, and employees' behaviors and ethical audits ensure this consistency [GL15].

4.1.18 | Lack of inclusivity in AI multistakeholder governance (C18)

The wide variety of AI applications requires inclusive governance. International organizations, governments, the commercial sector, and civil society are the typical four types of stakeholders involved in multistakeholder governance projects. [GL17]. These initiatives should include diverse stakeholders that may contribute with diverse perspectives. However, they fail to achieve inclusivity, as experience has shown that Western nations often dominate these initiatives through big technological organizations and international non-profit organizations with a limited understanding of what is happening locally [GL17].

4.1.19 | Lack of legal frameworks (C19)

Business leaders' growing use of AI systems increases their expectations for lawyers to adopt them [GL18]. The enhanced AI knowledge of lawyers can potentially make them more valuable to organizations and customers. Lawyers can use AI systems to improve their legal services' speed, quality, and efficiency to organizations. Given these benefits, the use of AI systems is expected to increase in the following years. Therefore, lawyers should be aware of AI and its use in improving the delivery of client services in terms of accuracy and efficiency, ethical requirements, and associated challenges. Ultimately, lawyers should perform independent judgment, monitor AI work, and communicate the impact that AI systems might have on customers [GL18].

4.2 | RQ2 (empirical investigations)

This section will present the survey participants' demographic analysis and participants' perceptions concerning the significance of identified challenges to developing ethically trustworthy AI systems. Moreover, we also apply the fuzzy AHP analysis to rank the significance of SPI manifesto core categories and their respected identified challenging factors.

4.2.1 | Survey participants demographics

Demographic data on survey respondents is crucial for the survey results to be reliable and useful and for researchers to better understand the factors influencing participants' responses. Moreover, the survey participants' demographics analysis will enable researchers to ensure that the sample of participants is representative of the studied population.[57] Therefore, we analyzed the data demographics of survey participants, and a summary of the results is provided below and is depicted in Figures 5, 6, and 7.

The data collection process was executed globally using snowballing; we collected feedback from 41 countries on five continents (Figure 5). The frequency analysis shows that 60 out of 156 total responses (i.e., 43.6%) were received from Europe, and the second most frequent respondents (50, 32.5%) were from Asia. The demographic analysis indicates that the survey participants are involved across the globe. Figure 6 presents the survey participants' professional designation, roles, and organizations. We noted that the survey participants hold nine different designations and belong to nine companies. In addition, the experiences of survey participants and their organization size were also analyzed. The frequency-based results presented in Figure 7A show the survey participants' generic and AI-focused experience. Out of 156 total respondents, 120 participants have specific AI-focused expertise, and the rest have a significant generic experience. It is also observed that the highest number of survey participants has 3 to 5 years of experience. Still, overall, there is a good mix of survey participants' experiences, as shown in Figure 7A. The organizational sizes of the survey participants were also analyzed, as illustrated in Figure 7B, and it

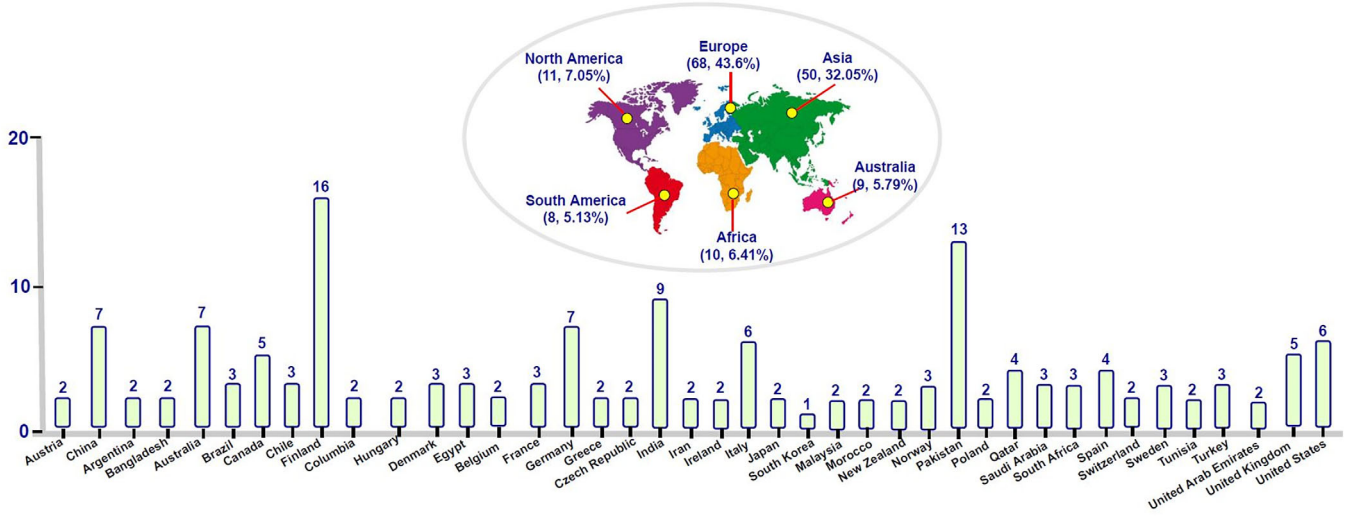

FIGURE 5 Geo distribution of survey participants.

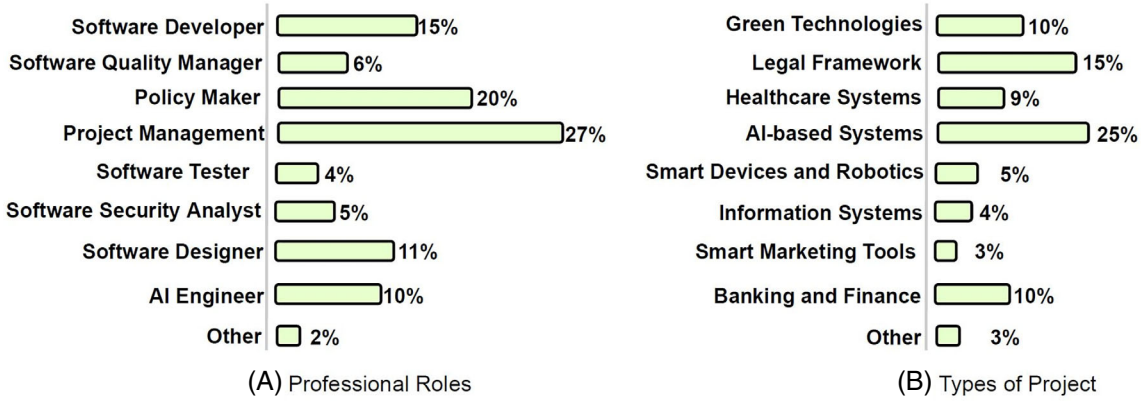

FIGURE 6 Participants' professional roles and projects.

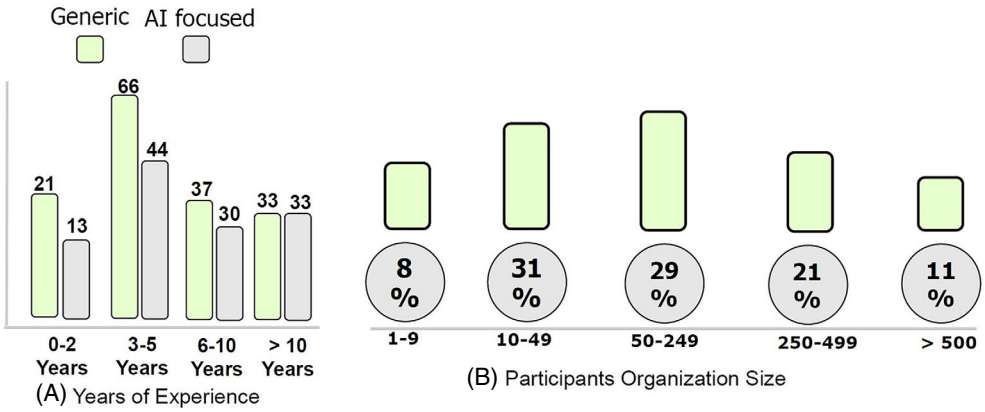

FIGURE 7 Participants' experiences and their organization size.

is noted that the survey participants belong to all sizes of organizations. However, the vast majority of responders work for medium-sized companies. Nevertheless, we found a good mix of participants' organizations' sizes from small to large.

4.2.2 | Frequency analysis of AI ethics challenges

The purpose of the survey study is to gain industry experts' perspectives on the challenges of AI ethics discovered through a multivocal literature review. The questionnaire survey consisted of the list of identified AI ethics challenges and requested survey participants to rate as per their understanding. The findings (Figure 8) showed that the majority of survey respondents believed that the issues listed could have a detrimental impact on the creation of an AI system that is morally sound. Based on the results, C10 (Lack of AI systems ethical impact determination) is agreed upon by 86% of the survey participants as the most important challenging factor that needs to be considered when developing ethically sound AI systems. Thilo[58] indicated that ethics is essential in developing and using AI systems. Because AI systems have a substantial social influence, it is crucial to make sure they are created and used in ways that are moral and consistent with the values of the people they will affect. One primary concern with AI systems is their potential to perpetuate or amplify existing biases and inequalities. For instance, an AI system trained on partial data may make biased decisions or recommendations. It is important for AI developers to carefully consider their systems' potential impacts and work to eliminate bias whenever possible. Furthermore, Desmond[59] indicated that another ethical concern with AI systems is their potential to be used for nefarious purposes, such as surveillance or manipulation. AI developers must consider their systems' potential risks and unintended consequences and design them with appropriate safeguards and controls. Thus, it is essential for AI developers to carefully consider their work's ethical implications and be transparent about their systems' potential impacts.

We further noted that C11 (Data governance issues, 82%) and C14 (Lack of assessment framework, 82%) are the second most important challenges for developing ethically sound AI systems. C11 (Data governance issues) refers to the policies, procedures, and practices an organization puts in place to ensure that the data it collects, processes, and uses is handled ethically and responsibly. In the context of AI ethics, there are several data governance issues that organizations need to consider, like data privacy, security, quality, transparency, and data accountability. Concerning C14 (Lack of assessment framework), there is currently a lack of a universally accepted assessment framework for evaluating the ethical implications of AI systems. This can make it challenging for organizations to ensure that their AI systems are ethically designed, implemented, and operated and for regulators and other stakeholders to evaluate the ethical implications of AI. It is also important for organizations to engage with stakeholders, including regulators, academics, and civil society organizations, to ensure that the ethical implications of their AI systems are thoroughly considered and addressed. This can help build trust in AI and ensure that it is used to benefit society.

To conclude, the questionnaire survey results indicated that all the identified challenges are essential to be addressed to develop an ethically sound AI system, as no challenge is marked by less than 65% of the respondents.

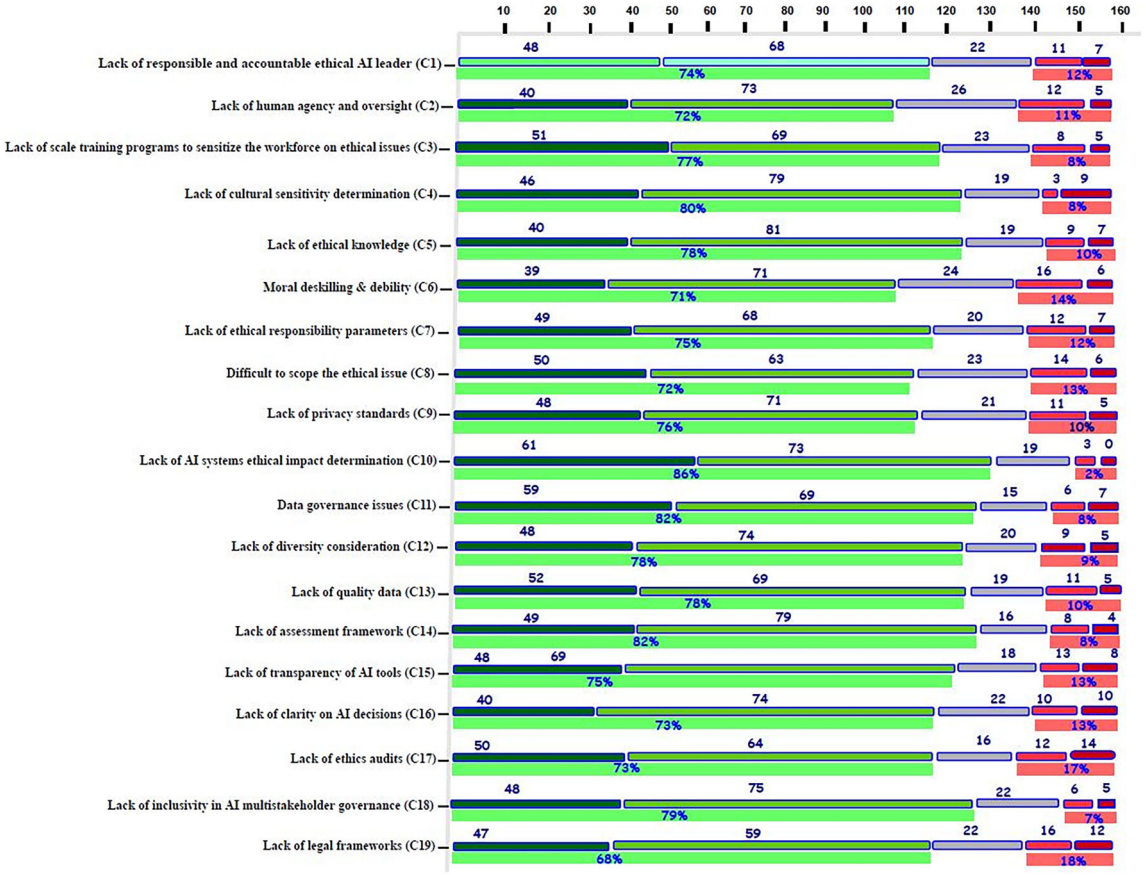

FIGURE 8 Survey participants' agreement and disagreement with identified AI ethics challenges.

4.3 | RQ 3 (application of fuzzy analytic hierarchy process)

We completed all the FAHP steps in this section to determine the values of enablers within and between their categories. The research was carried out on a desktop pc with an Intel Corei3 3.5 GHz CPU and 8 GB of RAM using the MATLAB R2016b programming environment created by math works, a US privately held company. The following sections carry out the FAHP's sequential phases.

4.3.1 | Step 1: Decompose a complicated decision problem into a hierarchical structure

According to Shameem[60] and Albayrak,[61] a complex decision-making problem is separated into connected decision factors at this level. The problem's hierarchical structure is split into at least three stages, as shown in References 62,63. The purpose of the problem is stated in stage 1 of this hierarchical structure. Nevertheless, level 2, and level 3 structures are used for the main categories and their respected enablers. Figure 9 shows the suggested hierarchy structure.

4.3.2 | Step 2: Pairwise comparison

Prioritizing the researched enablers and their basic categories in relation to the importance of AI ethics difficulties in the SPI manifesto is the main goal of FAHP study. The paired comparison is performed to assess each researched challenge and its main categories. To conduct a pairwise comparison in the second survey, experts were involved. We created a survey tool and communicated with the survey participants from the first study. A 43 individuals responded and consented to take part in the pairwise comparison analysis. The experts received the survey form created for pairwise comparison. In Appendix C, a sample of the questionnaire that was utilized for the paired comparison is included. After receiving the expert responses, we carefully evaluated them to look for any incomplete ones. All 43 responses were considered for the

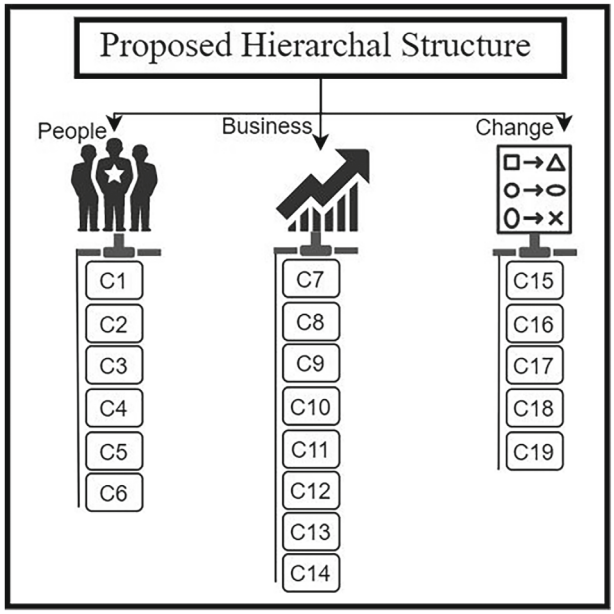

FIGURE 9 Proposed hierarchal structure of the problem.

TABLE 5 Triangular fuzzy conversion scale.[76]

Linguistic scale	Triangular fuzzy scale	Triangular fuzzy reciprocal scale
"Just equal (JE)"	(1, 1, 1)	(1, 1, 1)
"Equally important (EI)"	(0.5, 1, 1.5)	(0.6, 1, 2)
"Weakly important (WI)"	(1, 1.5, 2)	(0.5, 0.6, 1)
"Strongly more important (SMI)"	(1.5, 2, 2.5)	(0.4, 0.5, 0.6)
"Very strongly more important (VSMI)"	(2, 0.5, 3)	(0.3, 0.4, 0.5)
"Absolutely more important (AMI)"	(2.5, 3, 3.5)	(0.2, 0.3, 0.4)

FAHP analysis because no incomplete reactions were discovered during the manual inspection. The pairwise comparison survey's sample size (43 replies) is modest and might not be reliable enough to generalize the findings of the FAHP analysis. According to the literature, the FAHP seems to be a subjective methodology. However, given that numerous other research also takes into account the minimal sample size for AHP analysis, the tiny data set is equally appropriate for results generalization. For instance, Cheng and Li[64] gathered information from nine professionals to analyze and order the success variables for construction collaboration. Five replies were obtained, and Shameem et al.[60] used an AHP analysis to prioritize the success criteria of the distributed agile software development methodology.

Similarly, Wong and Li[60] used an AHP analysis to choose intelligent building systems while considering the feedback from nine experts. Thus, we performed FAHP research based on information collected from 43 specialists. We can demonstrate that the sample group of 43 experts is appropriate for generalizing the study conclusions by looking at the sample data of previous studies.

The second survey (FAHP survey) collected responses that were transformed using the geometric mean, a useful technique for transforming expert opinions into TFN numbers. The formula for calculating the geometric mean is given below (Equation 1), and the triangular fuzzy conversion scale from Table 5 was used to translate it into a fuzzy triangulation.

$$\text{Geometric mean} = \sqrt[n]{t1x\ t2 \times t3\ \ldots\ \ldots\ \ldots\ .tn} \tag{1}$$

$$t = \text{Weight of each response}$$

$$n = \text{Number of responses}$$

TABLE 6 Pairwise comparison of challenges categories.

Challenges categories			
	People	**Business**	**Change**
People	(1,1,1)	(1.5, 2.5, 3)	(1, 1.5, 2)
Business	(0.3, 0.4, 0.6)	(1,1,1)	(0.4, 0.5, 0.6)
Change	(0.5, 0.6, 1)	(1.5, 2, 2.5)	(1,1,1)

TABLE 7 Results of V values for criteria.

	People	**Business**	**Change**	**d (priority weight)**
V (People ≥ …)	—	1	1	1
V (Business≥ …)	0.030019	—	0.26502	0.030019
V (Change≥ …)	0.69836	1	—	0.69836

4.3.3 | Step 3: Test the consistency of the pair-wise matrix

In this part, we provided a step-by-step computation of the method used to determine the consistency of a given pairwise matrix. For this, we took into account the Table of categories (Table 6). Equation (2) is used to defuzzify a triangular fuzzy number from the pair-wise comparison matrix of the primary kinds to a crisp number, producing the matching Fuzzy Crisp Matrix (FCM), as illustrated in Table 7:

4.3.4 | Step 4: Calculating the local priority weight of each challenge and their respective categories

A numerical example

Table 6 lists the priority vector for each major category of issues. Equation (2) was used to determine the local priority weight (LPW) for each of the major challenge categories. Then, the synthetic extent values of three categories—people, business, and change—were established, and Equation (3) was used to determine the priority weight of each category. We have included the priority weight calculation for each category of enablers in the section below.

$$\sum_{i}^{n}\sum_{j}^{m} F_{gi}^{j} = (1,1,1) + (1.5{,}2,2.5) + (1{,}1.5{,}2) \ \ldots \ + (0.5{,}0.6{,}1) + (1,1,1) = (14.1{,}18.2{,}22.8) \tag{2}$$

$$\left[\sum_{i}^{n}\sum_{j}^{m} F_{gi}^{j}\right]^{-1} = \left(\frac{1}{22.8}, \frac{1}{18.2}, \frac{1}{14.1}\right) = (0.04386, 0.054945, 0.070922) \tag{3}$$

$$\sum_{j=1}^{m} F_{g1}^{j} = (1,1,1) + (1.5{,}2.5{,}3) + (1{,}1.5{,}2) = (5,7,8.5)$$

$$\sum_{j=1}^{m} F_{g2}^{j} = (0.3{,}0.4{,}0.6) + (1,1,1) + (0.4{,}0.5{,}0.6) = (2.2{,}2.5{,}3.2)$$

$$\sum_{j=1}^{m} F_{g3}^{j} = (0.5{,}0.6{,}1) + (1.5{,}2,2.5) + (1,1,1) = (4{,}5.1{,}6.5)$$

TABLE 8 Fuzzy crisp matrix (FCM) of challenges categories.

	People	Business	Change
People	1.9	2.5	1.5
Business	0.9	1.0	0.7
Change	0.8	2.0	1.0
Column sum	3.6	5.5	3.2

TABLE 9 Normalized matrix of challenges categories.

	People	Business	Change	Priority vector weight
People	0.37037	0.35714	0.40541	0.37938
Business	0.18519	0.14286	0.13514	0.14945
Change	0.25926	0.28571	0.27027	0.27593

The syntheses values of the three primary categories—εpeople,ε εbusiness,ε and εchangeε—were determined using Equation (4) and are as follows:

$$People = \sum_{j}^{m} F_{g1}^{j} \otimes \left[\sum_{i}^{n}\sum_{j}^{m} F_{gi}^{j}\right]^{-1} = (5, 7, 8.5) \otimes (0.04386, 0.054945) = (0.219298, 0.384615) \quad (4)$$

$$\text{Business} = (2.2,2.5,3.2) \otimes (0.04386, 0.054945) = (0.096491, 0.137363)$$

$$\text{Change} = (4,5.1,6.5) \otimes (0.04386, 0.054945) = (0.175439, 0.280220)$$

Using Equation (4), the degree of probability is calculated. The lowest degree of probability (priority weight) for every pair-wise comparison was computed and the outcomes are presented in Table 7.

The weight vector was thus calculated to be W' = (1, 0.030019, 0.69836) (Table 7). The relevance of the qualities was estimated using the formula W = (0.4789, 0.01435, 0.3337). The results show that people are the most important category, having the largest priority weight compared to the other enabling categories.

Test the consistency of the pair-wise matrix

In this part, we provided a step-by-step computation of the method used to determine the consistency of a given pairwise matrix. We have taken into account the Table of Categories (Table 6) and the calculated V-value from Table 7 for this. Equation (2) is used to convert a triangular fuzzy number from the pair-wise comparison matrix of the major categories to a crisp number, yielding the matching Fuzzy Crisp Matrix (FCM), as shown in Table 8:

The greatest Eigenvector (λmax) value of the FCM matrix is determined by computing the column total of every FCM matrix (Table 9) and splitting each element of the FCM matrix by the column sum. Furthermore, as seen in Table 9, the precedence weight is determined by averaging each row.

$$\lambda_{max} = \sum ([\sum \text{Cj}] \times \{\text{W}\}) \quad (5)$$

Where ∑Cj= sum of the columns of Matrix [C] (Table 9),
W= weight vector (Table 9), therefore

$$\lambda_{max} = 3.6^{*}0.37938 + 5.5^{*}0.14945 + 3.2^{*}0.27593 + = 3.0707$$

Based on the calculation, the largest Eigenvalue (λ_{max}) of the matrix FCM is 3.0707 (Equation 5). The dimension of FCM is 4. Therefore, n=4, and its RI is 0.9 (Table 10). Therefore, Equations (6) and (7) are used to calculate the consistency

TABLE 10 Random consistency index (RI).

Size of the matrix	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

TABLE 11 Pairwise comparison of the 'people' category.

	C1	C2	C3	C4	C5	C6
C1	(1,1,1)	(0.4, 0.5, 0.6)	(1.5, 2, 2.5)	(0.3, 0.4, 0.5)	(0.4, 0.5, 0.6)	(1, 1.5, 2)
C2	(1.5, 2, 2.5)	(1,1,1)	(2, 2.5, 3)	(0.5, 1, 1.5)	(1, 1.5, 2)	(1, 1.5, 2)
C3	(0.4, 0.5, 0.6)	(0.3, 0.4, 0.5)	(1,1,1)	(2, 2.5, 3)	(2.5, 3, 3.5)	(0.4, 0.5, 0.6)
C4	(2, 2.5, 3)	(0.6, 1, 2)	(0.3, 0.4, 0.5)	(1,1,1)	(0.5, 0.6, 1)	(1, 1.5, 2)
C5	(1.5, 2, 2.5)	(0.5, 0.6, 1)	(0.2, 0.3, 0.4)	(1, 1.5, 2)	(1,1,1)	(2, 2.5, 3)
C6	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(1,1,1)

Note: $I_{max} = 5.85$, CI = 0.21, CR = 0.19.

TABLE 12 Pairwise comparison of the 'business' category.

	C7	C8	C9	C10	C11	C12	C13	C14
C7	(1,1,1)	(1, 1.5, 2)	(2.5, 3, 3.5)	(0.6, 1, 2)	(1.5, 2, 2.5)	(1, 1.5, 2)	(0.5, 0.6, 1)	(0.3, 0.4, 0.5)
C8	(0.5, 0.6, 1)	(1,1,1)	(0.5, 0.6, 1)	(1, 1.5, 2)	(0.4, 0.5, 0.6)	(1, 1.5, 2)	(2, 2.5, 3)	(1, 1.5, 2)
C9	(0.2, 0.3, 0.4)	(1, 1.5, 2)	(1,1,1)	(0.5, 1, 1.5)	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(1, 1.5, 2)	(0.5, 0.6, 1)
C10	(0.5, 1, 1.5)	(0.5, 0.6, 1)	(0.6, 1, 2)	(1,1,1)	(0.2, 0.3, 0.4)	(2, 2.5, 3)	(0.5, 1, 1.5	(2, 2.5, 3)
C11	(0.4, 0.5, 0.6)	(1.5, 2, 2.5)	(1, 1.5, 2)	(2.5, 3, 3.5)	(1,1,1)	(0.4, 0.5, 0.6)	(0.2, 0.3, 0.4)	(1, 1.5, 2)
C12	(0.5, 0.6, 1)	(0.5, 0.6, 1)	(1, 1.5, 2)	(0.3, 0.4, 0.5)	(1.5, 2, 2.5)	(1,1,1)	(0.4, 0.5, 0.6)	(2, 2.5, 3)
C13	(1, 1.5, 2)	(0.3, 0.4, 0.5)	(0.5, 0.6, 1)	(0.6, 1, 2)	(2.5, 3, 3.5)	(1.5, 2, 2.5)	(1,1,1)	(0.4, 0.5, 0.6)
C14	(2, 2.5, 3)	(0.5, 0.6, 1)	(1, 1.5, 2)	(0.3, 0.4, 0.5)	(0.5, 0.6, 1	(0.3, 0.4, 0.5)	(1.5, 2, 2.5)	(1,1,1)

Note: $I_{max} = 10.4$, CI = 0.17, CR = 0.11.

index and consistency ratio as follows:

$$CI = \frac{\lambda_{max}}{n - 1\frac{3.0707-3}{3-1}0.035553} \tag{6}$$

$$CR = \frac{CI}{RI} = \frac{0.035553}{0.58} = 0.061 \tag{7}$$

The calculated value of CR is 0.061 < 0.10; therefore, the pairwise comparison matrix developed for the categories of enablers is consistent and acceptable. Similarly, the consistency ratio for all the classes are checked, and the results of the 'people,' 'business,' and 'change' category are given in Table 11, 12, and 13, respectively.

4.3.5 | Step 5: Determining the ranking of the challenges

The challenging factors were also ranked within their respective category and for overall software process improvements, considering their significance towards ethically sound AI system development. Therefore, to check the criticality of the identified challenges within their category, we calculated local ranks by considering the local weights, that is, the value of 'W.' We used the criteria to determine challenges rankings; a challenge is regarded as the highest priority challenge if its value of weight vector 'W' is higher. For instance, the value of 'W' of *C1 (Lack of responsible and accountable ethical AI*

TABLE 13 Pairwise comparison of the 'change' category.

	C15	C16	C17	C18	C19
C15	(1,1,1)	(0.3, 0.4, 0.5)	(0.4, 0.5, 0.6)	(1.5, 2, 2.5)	(0.4, 0.5, 0.6)
C16	(2, 2.5, 3)	(1,1,1)	(2, 2.5, 3)	(0.5, 1, 1.5)	(1, 1.5, 2)
C17	(1.5, 2, 2.5)	(0.3, 0.4, 0.5)	(1,1,1)	(2, 2.5, 3)	(2.5, 3, 3.5)
C18	(0.4, 0.5, 0.6)	(0.6, 1, 2)	(0.3, 0.4, 0.5)	(1,1,1)	(0.5, 0.6, 1)
C19	(1.5, 2, 2.5)	(0.5, 0.6, 1)	(0.2, 0.3, 0.4)	(1, 1.5, 2)	(1,1,1)

Note: $I_{max} = 5.50$, CI = 0.12, CR = 0.10.

TABLE 14 Local and global ranks of identified challenges.

Categories	Categories weight (CW)	Challenges	Local weights (LW)	Local ranking (LR)	Global weights (GW)	Global ranking (GR)
People	0.379	C1	0.420	1	0.159	1
		C2	0.160	6	0.061	8
		C3	0.241	3	0.091	5
		C4	0.180	5	0.068	7
		C5	0.201	4	0.076	6
		C6	0.342	2	0.130	3
Business	0.149	C7	0.176	5	0.026	15
		C8	0.378	1	0.056	10
		C9	0.104	8	0.015	19
		C10	0.170	6	0.025	16
		C11	0.320	3	0.048	12
		C12	0.340	2	0.051	11
		C13	0.191	4	0.028	14
		C14	0.161	7	0.024	17
Change	0.276	C15	0.210	3	0.058	9
		C16	0.120	4	0.033	13
		C17	0.487	1	0.134	2
		C18	0.450	2	0.124	4
		C19	0.063	5	0.017	18

leader) is 0.420. Thus, it is the highest priority challenge in the 'people' category (Table 14). We further noted that the value of 'W' of C6 (Moral deskilling & debility) is 0.342, and it is declared the second most critical challenging area for developing ethically trustworthy AI systems. Determination of local rankings aims to assist the practitioners in developing strategies by considering the most critical challenges for successfully addressing the challenges of a software process improvement area towards developing an ethically sound AI system.

Moreover, the identified challenges were globally ranked to determine them critically in all areas of software process improvements. The global weights were calculated by multiplying each challenge's category and local weight. For instance, the global rank of C1 (Lack of responsible and accountable ethical AI leader) is (category weight × challenge weight), that is, (0.379 × 0.420 = 0.159), and it is ranked as first highest priority challenge (Table 14). The calculated global ranks provide the criticality of the identified challenges across software process improvement areas for developing ethically trustworthy AI systems. Hence, as per the calculated global weights, C1 (Lack of responsible and accountable ethical AI leader), C17 (Lack of ethics audits), C6 (Moral deskilling & debility), C18 (Lack of inclusivity in AI multistakeholder

governance), and C3 (Lack of scale training programs to sensitize the workforce on ethical issues) are the top-ranked challenges for ethically trustworthy AI system (Table 14).

4.3.6 | Step 6: Taxonomy of challenges

The last stage of this study is constructing a taxonomy based on prioritizing the discovered AI ethical challenges. Prioritization helps to make informed decisions about allocating time and resources, which can improve the quality of your work and help you achieve desired goals. Youssef and Barry[65] underlined that prioritization is an essential skill that can help to manage time and resources effectively, increase productivity, reduce stress, and make better decisions. The objective of prioritization-based taxonomy development is to inform practitioners regarding the AI ethics challenging factors of specific SPI manifesto categories and the overall SPI manifesto process. This will help the practitioners develop or revise strategies for addressing the challenges of developing ethically sound AI systems.

The developed taxonomy presented in Figure 10 indicated that C1 (Lack of responsible and accountable ethical AI leader) stands out as the most important challenge to developing an ethically sound AI system within the 'people' category and for the overall SPI manifesto. This indicated that practitioners need to consider C1 a priority because, as an AI leader, it is important to be responsible and accountable for ethical AI practices. This entails ensuring that algorithms and AI systems are created, developed, and applied ethically, openly, and equitably. It's critical to comprehend the ethical ramifications of the AI systems and algorithms you are in charge of, according to Stahl.[66] This includes understanding the potential impacts on individuals and society and taking steps to mitigate any negative consequences. AI leaders can shape the culture of their teams and organization. This includes promoting an ethical culture that values transparency, fairness, and responsibility.[67,68] It is also important to engage with stakeholders, including those who may be affected by the AI systems and algorithms you are responsible for. This could include working with community groups, advocacy organizations, and other stakeholders to understand their concerns and incorporate feedback into your work.[69] Furthermore, as an AI leader, it is important to continuously assess and improve your AI systems and algorithms to ensure they are aligned with ethical principles. This includes regularly reviewing and updating your AI practices to ensure they are fair, transparent, and responsible.[70]

C17 (Lack of ethics audits) ranks first within the 'people' category and the second most important challenge for developing ethically sound AI systems. Ethics audits are a useful tool to ensure that algorithms and AI systems are created

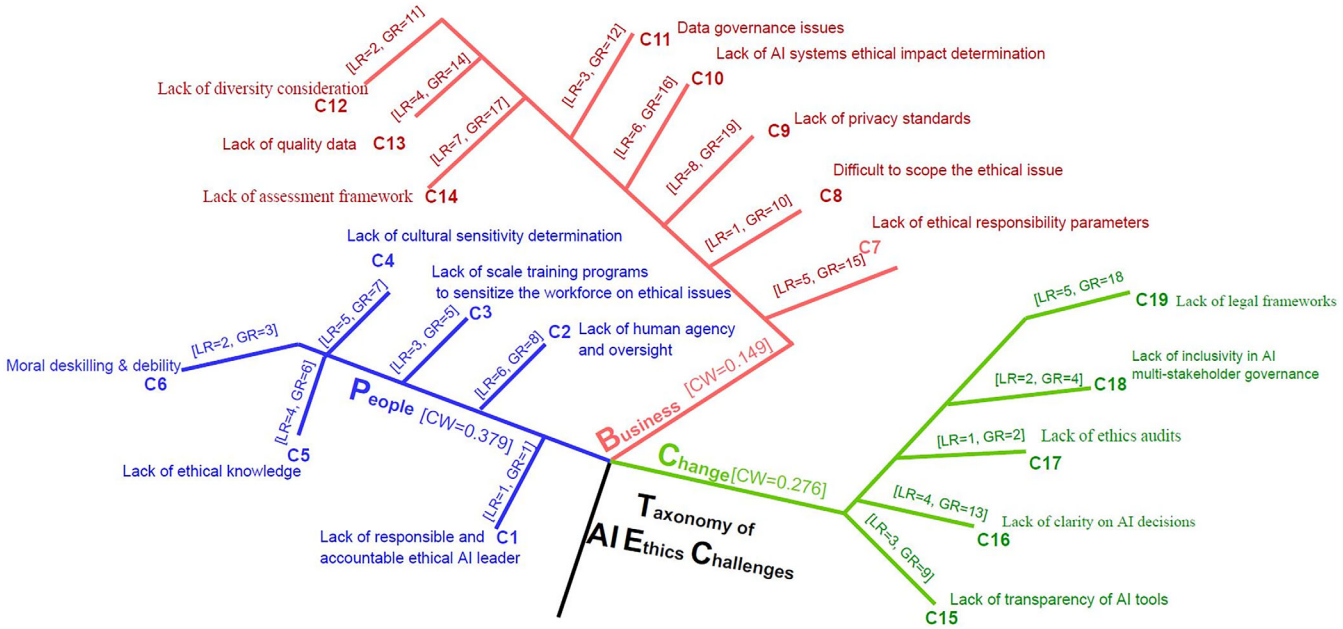

FIGURE 10 Prioritization-based taxonomy.

and used ethically and responsibly. An ethics audit systematically reviews an AI system or algorithm to identify potential ethical concerns or implications. This can help organizations identify and address potential ethical issues before they become a problem and ensure that their AI systems align with their values and ethical principles.[71] There are several reasons why some organizations may not conduct ethics audits: (i) Lack of resources conducting an ethics audit can be resource-intensive, and some organizations may not have the necessary resources (e.g., staff, time, budget) to conduct one. (ii) Lack of expertise- Conducting an ethics audit requires specialized knowledge and expertise, and some organizations may not have the necessary in-house expertise. (iii) Some organizations may not be aware of the importance of conducting ethics audits or may not understand the process or how to conduct one.

In some cases, there may be limited legal or regulatory requirements for conducting ethics audits, which may discourage organizations from doing so. Organizations must prioritize conducting ethics audits to ensure that their AI systems and algorithms are designed and used ethically and responsibly.[72] Thus, organizations need to plan their resources as a prerequisite to making an effective ethics audit. The taxonomy results show that C6 (Moral deskilling & debility), C18 (Lack of inclusivity in AI multistakeholder governance), and C3 (Lack of scale training programs to sensitize the workforce on ethical issues) are the third, fourth, and fifth ranked challenges for developing ethically sound AI systems.

5 | SUMMARY AND DISCUSSION

AI has become indispensable for banking, manufacturing, retail, and health industries. As a result, there is debate over developing laws for various technologies, such as manufacturing and nuclear power, to name a few, to limit the potential ethical harm they could cause. AI systems likewise carry the same ethical risk of harm; more particularly, they might abolish human control. AI ethics is a field that deals with the ethical issues and challenges that arise from the development and use of AI. It involves examining the potential consequences of AI systems and algorithms and developing strategies and guidelines to ensure they are used ethically and responsibly. AI ethics challenges can arise at various stages of the AI development process, including during the design, training, testing, deployment, and maintenance of AI systems. Addressing AI ethics challenges is important because it helps ensure that AI is used fairly and is beneficial to society. It also helps build trust in AI and avoid negative outcomes that could severely affect individuals and society. Literature shows just a few of the many issues explored in the field of AI ethics.[72] It is a rapidly evolving field, and researchers, policymakers, and the general public need to stay informed about the latest developments and engage in ongoing dialogue about the ethical implications of AI. This study aims to explore and prioritize AI ethics challenges by answering the following research questions.

5.1 | RQ1: What are the AI ethics challenges against the SPI manifesto, reported in multivocal literature?

We used the multivocal literature review method to examine the ethical issues with AI raised by the scientific community. By executing the steps of the MLR approach, 19 challenging factors were identified that are important for developing ethically trustworthy AI software systems. We further mapped the identified challenges against the core areas of the SPI manifesto (i.e., people, change, and business). The purpose of this mapping is to discover the problematic areas for the creation of morally sound AI systems in the SPI manifesto.

5.2 | RQ2: How strongly are the identified challenges important to consider in SPI manifesto categories for developing ethically sound AI software?

We performed the empirical study with practitioners to check the practicality of the identified AI ethics challenges. Using a questionnaire survey approach, we received feedback from 156 experts. The frequency analysis shows that most respondents agree that the reported challenging factors are important to be considered by practitioners aiming to develop an ethically sound AI system. Based on the frequency analysis, "Lack of AI systems ethical impact determination" was considered by 86% of the survey participants. The impact assessments involve identifying and evaluating the potential ethical impacts of an AI system, such as its potential biases, privacy implications, and potential for autonomous decision-making.

It is also important to consider the perspectives and needs of different stakeholders when developing and using AI systems. This may involve consulting with experts, engaging with community groups, and considering the views of those affected by the AI system. Thus, developing guidelines and best practices for developing and using AI can help ensure that AI systems are used ethically and responsibly. Overall, it is important to consider the ethical implications of AI systems carefully and to take steps to mitigate any negative impacts that may arise. We further noted that Data governance issues and Lack of assessment framework are considered by 82% of the survey participants as the most significant challenges to consider for developing ethically sound AI systems. Notably, more than 65% of the survey participants marked all the challenging factors enlisted as essential areas to address.

5.3 | RQ3: What would be the ranked-based taxonomy of AI ethics challenges?

Prioritization is identifying and ordering tasks or projects based on their importance or urgency. It is essential for effectively managing time and resources. It allows one to focus on the most important tasks and avoid wasting time on less important or lower-priority tasks. Hence, we develop a prioritization-based taxonomy of the identified AI ethics challenges. The results presented in the taxonomy show the priority levels of each challenge within the categories of the SPI manifesto, as well as their priorities on the overall SPI manifesto. For instance, the determined priority ranking indicated that C19 (Lack of legal frameworks) is ranked fifth in the 'change' category and 18th in the overall SPI manifesto., The local and global ranking assists practitioners in making decisions and policies with the intent to develop ethically sound AI systems. Moreover, the results indicated that C1 (Lack of responsible and accountable ethical AI leader), C17 (Lack of ethics audits), C6 (Moral deskilling & debility), C18 (Lack of inclusivity in AI multistakeholder governance), and C3 (Lack of scale training programs to sensitize the workforce on ethical issues) are the top-ranked challenging factors to be considered in SPI manifesto.

6 | THREATS TO VALIDITY

This section describes the significant risks of this research and how they were mitigated utilizing the recommendations made by References 45,73,74.

6.1 | Internal validity

By making the first mandatory option in our questionnaire, we ensured participants have knowledge of ethics and AI to reduce the risk of misrepresentation and internal validity. The participants additionally had a lot of experience managing and developing software, which reduced the chance of participants lacking skills.

6.2 | Construct validity

One potential threat to this study's construct validity is the multivocal literature review's incompleteness. This paper's conclusions are based on literature that was culled using search terms from a few digital databases and search engines. We added the keyword alternatives to overcome this restriction to create a strong search string. Further, we searched through a variety of search engines and digital databases to find the most relevant material for the study's aims. Despite these measures, we acknowledge that a few studies may have been missed, and completeness cannot be ensured, as it often occurs in literature reviews.

Furthermore, we utilized numerous sources of evidence, including analysis of multiple stakeholders, and artifacts, and observations, maintaining a chain of evidence (e.g., the Grounded Theory approach coding scheme,[75] and had the results reviewed through member checking to address the threat to construct validity).

Another construct threat is the development of survey instruments that refer to the weather as not the variables of the survey instrument that are understandable for industry experts. Moreover, there is a threat regarding the measurement scale of the variables. To mitigate this threat, we conducted a pilot assessment study with industry experts and updated

the questionnaire by considering their suggestions and to assess the survey variables, we used the well-known standard five-scale Likert scale.

The biases of the researchers and survey participants may also influence the reliability of the study's conclusions. We conducted an inter-rater validity test to investigate and reduce the researchers' prejudices to address the researchers' biases. Furthermore, to mitigate the biases of the questionnaire survey study, we used the 'Neutral' option, which allows survey participants to be unbiased.

6.3 | External validity

This study's findings are largely based on empirical data gathered from a specific setting, namely the ethics of the AI system. Therefore, the findings of this study do not claim the generalization of the study for the rest of the software development stakeholders. However, we do not consider the findings definitive or final because they can be repeated and modified in other circumstances.[76,77]

Regarding data representativeness, the study only contains AI ethics information (not the development process). The dependability of the data was strengthened by gathering it from many businesses with multiple teams, and participants with a wide range of jobs, and participant triangulation was aided.[78] It's probable that despite our analysis of data spans, we overlooked certain changes in the results, specifically the context-specific causes and tactics. To broaden the breadth of our findings and confirm their explanatory value in other circumstances, we recommend that future studies on this topic include extra data sources, such as more cases or interviews.

6.4 | Reliability

All authors extensively debated and revised emerging codes and relationships to reduce the risk of prejudice and ensure the validity of the data collecting and analysis techniques. Additionally, about interpretative validity,[78] we carried out member checking with the participation of three authors to confirm the accuracy of our findings, further ensuring investigator triangulation.

7 | STUDY IMPLICATIONS

This section presents the implications of this study for both researchers and practitioners as follows:

***For researchers*:** The results of this study offer a comprehensive understanding of the challenging factors that may negatively impact the consideration of ethical practices during AI system development. This knowledge base enables the research community to explore critical areas in developing ethically robust AI systems. The identified list of AI ethics challenges reveals new research directions, such as the need for guidelines to address these challenging factors. The study also examines the identified challenges against the core categories of the SPI manifesto, aiding researchers in developing tools and strategies to address AI ethics challenges at the category level. Furthermore, a prioritization-based taxonomy of the identified challenges related to the SPI manifesto has been developed, helping researchers understand the criticality of these challenges at both the SPI manifesto category level and in overall software process improvement. We believe that providing a priority order for the identified challenging factors will aid researchers in addressing the most significant ones in future research.

***For practitioners*:** The findings of our study can be applied by practitioners in various ways. For example, the comprehensive multivocal literature review and empirical study serve as a knowledge base regarding the challenging factors involved in developing ethically sound AI systems. The findings identify 19 significant challenges, each of which urges professionals to focus on creating ethical AI systems. Prioritizing these challenges assists practitioners in determining the most crucial elements. By identifying and prioritizing these challenges, practitioners can develop and refine methods to better address ethical concerns in AI system development. Additionally, the study's findings include a prioritization-based taxonomy of the challenging factors, accounting for key SPI manifesto categories, regional rankings, and global rankings. This prioritization-based taxonomy informs practitioners about the criticality of each challenge for specific categories and the overall software improvement process. The novel application of the fuzzy AHP approach in this domain could also

aid industry experts in addressing multicriteria decision-making problems and handling the vague opinions of AI ethics experts.

8 | CONCLUSIONS AND FUTURE WORK

Considering ethics in designing and developing AI systems is essential to minimize negative impacts, maximize the benefits of AI, and ensure that AI is used ethically and responsibly. To develop ethically robust systems, this study explores the core challenging factors against the core dimensions of the SPI manifesto. This study's multivocal literature review approach revealed 19 challenging factors mapped against SPI manifesto categories. Furthermore, a questionnaire survey of 156 industry professionals was conducted to determine how realistic the highlighted difficulties were. The empirical results provide evidence that the challenging factors enlisted are essential to be considered in SPI manifesto categories for developing ethically robust AI systems. In addition, the ranking of identified challenges was determined with respect to their significance for developing ethically sound AI systems. The prioritization-based results revealed that lack of responsible and accountable ethical AI leaders, lack of ethics audits, moral deskilling & debility, lack of inclusivity in AI multistakeholder governance, and lack of scale training programs to sensitize the workforce on ethical issues are the highest priority challenges that need remarkable consideration by industry experts.

Our future research efforts will be devoted to the development of guidelines that can assist industry practitioners in addressing the challenging factors faced while developing ethically sound AI systems. To achieve this, we will conduct interviews and case studies with industry experts to understand experts' positions concerning AI ethics challenges. We will also mine the digital repositories to extract the best practices addressing the challenging aspects of AI ethics system development.

AUTHOR CONTRIBUTIONS

Methodologyand formal analysis done by Muhammad Azeem Akbar and Arif Ali Khan; Projectadministration and Data curation done by Sajjad Mahmood and Selina Demi; the finalrevision and English polishing done by Saima Rafi.

FUNDING INFORMATION

The work is supported by Software Engineering group of LUT University, Finland.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Muhammad Azeem Akbar https://orcid.org/0000-0002-4906-6495

REFERENCES

1. Kuziemski M, Misuraca GJT p. AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. *Telecommun Policy*. 2020;44:101976.
2. Müller VC. *Ethics of Artificial Intelligence and Robotics*. The Stanford Encyclopedia of Philosophy; 2020.
3. Greene D, Hoffmann AL, Stark L. Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. 2019.
4. Vakkuri V, Kemell K-K, Abrahamsson P. Implementing ethics in AI: initial results of an industrial multiple case study. Paper presented at: International Conference on Product-Focused Software Process Improvement. 2019:331-338.
5. Leikas J, Koivisto R, Gotcheva N. Ethical framework for designing autonomous intelligent systems. *J Open Innov Technol Market Complex*. 2019;5:18.
6. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach*. 2020;30:99-120.
7. Chatila R, Firth-Butterfield K, Havens JC. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems Version 2. University of southern California Los Angeles. 2018.
8. Vakkuri V, Kemell K-K, Abrahamsson P. AI ethics in industry: a research framework. *Comput Sci*. 2019. https://arxiv.org/ftp/arxiv/papers/1910/1910.12695.pdf
9. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1:389-399.

10. Pekka A, Bauer W, Bergmann U, et al. The European Commission's high-level expert group on artificial intelligence: ethics guidelines for trustworthy AI. 2018:1-37.
11. Bundy A. *Preparing for the Future of Artificial Intelligence*. Springer; 2017.
12. Zhongming Z, Wei L. *Beijing Academy of Artificial Intelligence*. BAAI Ecology; 2020.
13. ISO/IEC. ISO/IEC JTC 1/SC 42 Artificial intelligence. 2021.
14. Vakkuri V, Kemell K-K, Jantunen M, Abrahamsson P. This is just a prototype: how ethics are ignored in software startup-like environments. Paper presented at: International Conference on Agile Software Development. 2020:195-210.
15. European Commission *Ethics Guidelines for Trustworthy AI*. Website of the European Union; 2019.
16. Vakkuri V, Kemell K-K, Jantunen M, Halme E, Abrahamsson P. ECCOLA—A method for implementing ethically aligned AI systems. *J Syst Soft*. 2021;182:111067.
17. Holmes W, Bialik M, Fadel C. *Artificial Intelligence in Education*. Center for Curriculum Redesign; 2020.
18. Amershi S, Begel A, Bird C, et al. Software engineering for machine learning: a case study. Paper presented at: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). 2019:291-300.
19. Lwakatare LE, Raj A, Bosch J, Olsson HH, Crnkovic I. A taxonomy of software engineering challenges for machine learning systems: an empirical investigation. In Agile Processes in Software Engineering and Extreme Programming: 20th International Conference, XP 2019, Montréal, QC, Canada, May 21–25, 2019, Proceedings 20. Springer International Publishing;. 2019:227-243.
20. Akkiraju R, Sinha V, Xu A, et al. Characterizing machine learning processes: a maturity framework. Paper presented at: International Conference on Business Process Management. 2020:17-31.
21. Nguyen-Duc A, Sundbø I, Nascimento E, Conte T, Ahmed I, Abrahamsson P. A multiple case study of artificial intelligent system development in industry. Proceedings of the Evaluation and Assessment in Software Engineering. 2020:1-10.
22. Khan AA, Badshah S, Liang P, et al. Ethics of AI: a systematic literature review of principles and challenges. Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022. 2022:383-392.
23. Khan AA, Akbar MA, Fahmideh M, et al. AI ethics: an empirical study on the views of practitioners and lawmakers. *IEEE Trans Comput Soc Syst*. 2023:1-14.
24. Garousi V, Felderer M, Mäntylä MV. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf Softw Technol*. 2019;106:101-121.
25. Akbar MA, Khan AA, Mahmood S, Mishra A. SRCMIMM: the software requirements change management and implementation maturity model in the domain of global software development industry. *Inf Technol Manag*. 2022:1-25.
26. Akbar MA, Smolander K, Mahmood S, Alsanad A. Toward successful DevSecOps in software development organizations: a decision-making framework. *Inf Softw Technol*. 2022;147:106894.
27. Khan AA, Shameem M, Nadeem M, Akbar MA. Agile trends in Chinese global software development industry: fuzzy AHP based conceptual mapping. *Appl Soft Comput*. 2021;102:107090.
28. Akbar MA, Shameem M, Khan AA, Nadeem M, Alsanad A, Gumaei A. A fuzzy analytical hierarchy process to prioritize the success factors of requirement change management in global software development. *J Softw: Evol Process*. 2021;33:e2292.
29. Akbar MA, Khan AA, Mahmood S, Alsanad A, Gumaei A. A robust framework for cloud-based software development outsourcing factors using analytical hierarchy process. *J Softw: Evol Process*. 2021;33:e2275.
30. Garousi V, Felderer M, Hacaloğlu T. Software test maturity assessment and test process improvement: a multivocal literature review. *Inf Softw Technol*. 2017;85:16-42.
31. Adams RJ, Smart P, Huff AS. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. *Int J Manag Rev*. 2017;19:432-454.
32. Zhang H, Babar MA, Tell P. Identifying relevant studies in software engineering. *Inf Softw Technol*. 2011;53:625-637.
33. Jalali S, Wohlin C. Systematic literature studies: database searches vs. backward snowballing. Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement. 2012:29-38.
34. Badampudi D, Wohlin C, Petersen K. Experiences from using snowballing and database searches in systematic literature studies. Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. 2015:17.
35. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 2014:38.
36. Afzal W, Torkar R, Feldt R. A systematic review of search-based testing for non-functional system properties. *Inf Softw Technol*. 2009;51:957-976.
37. Niazi M, Mahmood S, Alshayeb M, Qureshi AM, Faisal K, Cerpa N. Toward successful project management in global software development. *Int J Proj Manag*. 2016;34:1553-1567.
38. Khan AA, Keung J, Niazi M, Hussain S, Ahmad A. Systematic literature review and empirical investigation of barriers to process improvement in global software development: client–vendor perspective. *Inf Softw Technol*. 2017;87:180-205.
39. Khan SU, Niazi M, Ahmad R. Factors influencing clients in the selection of offshore software outsourcing vendors: an exploratory study using a systematic literature review. *J Syst Softw*. 2011;84:686-699.
40. Niazi M, Mahmood S, Alshayeb M, et al. Challenges of project management in global software development: a client-vendor analysis. *Inf Softw Technol*. 2016;80:1-19.
41. Khan AA, Keung J, Hussain S, Niazi M, Kieffer S. Systematic literature study for dimensional classification of success factors affecting process improvement in global software development: client–vendor perspective. *IET Softw*. 2018;12:333-344.

42. Shameem M, Kumar C, Chandra B, Khan AA. Systematic review of success factors for scaling agile methods in global software development environment: a client-vendor perspective. Paper presented at: 2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW). 2017:17-24.
43. White VJ, Glanville JM, Lefebvre C, Sheldon TA. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. *J Inf Sci*. 2001;27:357-370.
44. Kelle U. The development of categories: different approaches in grounded theory. *Sage Handbook Ground Theory*. 2010;2:191-213.
45. Runeson P, Höst M. Guidelines for conducting and reporting case study research in software engineering. *Emp Softw Eng*. 2009;14:131-164.
46. Rafi S, Akbar MA, Mahmood S, Alsanad A, Alothaim A. Selection of DevOps best test practices: a hybrid approach using ISM and fuzzy TOPSIS analysis. *J Softw Evol Process*. 2022;34:e2448.
47. Punter T, Ciolkowski M, Freimut B, John I. Conducting on-line surveys in software engineering. Proceedings 2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. 2003:80-88.
48. Molléri JS, Petersen K, Mendes E. An empirically evaluated checklist for surveys in software engineering. *Inf Softw Technol*. 2020;119:106240.
49. Garousi V, Tarhan A, Pfahl D, Coşkunçay A, Demirörs O. Correlation of critical success factors with success of software projects: an empirical investigation. *Softw Qual J*. 2019;27:429-493.
50. Baltar F, Brunet F. Social research 2.0: virtual snowball sampling method using Facebook. *Internet Res*. 2012;22(1):57-74.
51. Kitchenham B, Pfleeger SL. Principles of survey research: part 5: populations and samples. *ACM SIGSOFT Softw Eng Notes*. 2002;27:17-20.
52. Ali S, Khan SU. Software outsourcing partnership model: an evaluation framework for vendor organizations. *J Syst Softw*. 2016;117:402-425.
53. Akbar MA, Sang J, Khan AA, et al. Statistical analysis of the effects of heavyweight and lightweight methodologies on the six-pointed star model. *IEEE Access*. 2018;6:8066-8079.
54. Keshta I, Niazi M, Alshayeb M. Towards implementation of requirements management specific practices (SP1. 3 and SP1. 4) for Saudi Arabian small and medium sized software development organizations. *IEEE Access*. 2017;5:24162-24183.
55. Mahmood S, Anwer S, Niazi M, Alshayeb M, Richardson I. Key factors that influence task allocation in global software development. *Inf Softw Technol*. 2017;91:102-122.
56. Pries-Heje JJ. SPI manifesto. Version A.1.2.2010. 2010.
57. Axinn WG, Link CF, Groves RM. Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*. 2011;48:1127-1149.
58. Hagendorff T. Blind spots in AI ethics. *AI Ethics*. 2022;2:851-867.
59. Ong DC. An ethical framework for guiding the development of affectively-aware artificial intelligence. Paper presented at: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII). 2021:1-8.
60. Shameem M, Kumar RR, Kumar C, Chandra B, Khan AA. Prioritizing challenges of agile process in distributed software development environment using analytic hierarchy process. *J Softw: Evol Process*. 2018;30:e1979.
61. Albayrak E, Erensal YC. Using analytic hierarchy process (AHP) to improve human performance: an application of multiple criteria decision making problem. *J Intell Manuf*. 2004;15:491-503.
62. Akbar MA, Khan AA, Huang Z. Multicriteria decision making taxonomy of code recommendation system challenges: a fuzzy-AHP analysis. *Inf Technol Manag*. 2022;1-17.
63. Akbar MA, Shameem M, Mahmood S, Alsanad A, Gumaei A. Prioritization based taxonomy of cloud-based outsource software development challenges: fuzzy AHP analysis. *Appl Soft Comput*. 2020;95:106557.
64. Cheng EW, Li H. Construction partnering process and associated critical success factors: quantitative investigation. *J Manag Eng*. 2002;18:194-202.
65. Youssef M, Webster B. A multi-criteria decision making approach to the new product development process in industry. 2022:83-93.
66. Stahl BC. Responsible innovation ecosystems: ethical implications of the application of the ecosystem concept to artificial intelligence. *Int J Inf Manag*. 2022;62:102441.
67. Eisenbeiß SA, Giessner SR. The emergence and maintenance of ethical leadership in organizations: a question of embeddedness? *J Personnel Psychol*. 2012;11:7.
68. Burr C, Leslie D. Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI Ethics*. 2022;3:1-26.
69. Schiff D, Biddle J, Borenstein J, Laas K. What's next for ai ethics, policy, and governance? A global overview. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020:153-158.
70. Dignum V. Responsibility and artificial intelligence. *The Oxford Handbook of Ethics of AI*. Vol 4698. Oxford University Press; 2020:215.
71. Zinda N. Ethics auditing framework for trustworthy AI: lessons from the IT audit literature. *The 2021 Yearbook of the Digital Ethics Lab*. Springer; 2022:183-207.
72. Khan AA, Akbar MA, Waseem M, et al. AI ethics: software practitioners and lawmakers points of view. *IEEE Trans Comput Soc Syst*. 2022.
73. Fielding NG, Lee NFRM, Lee RM. *Computer Analysis and Qualitative Research*. Sage; 1998.
74. Yin RK. *Case Study Research: Design and Methods*. Vol 5. Sage; 2009.
75. Rodriguez P, Urquhart C, Mendes E. A theory of value for value-based feature selection in software engineering. *IEEE Trans Softw Eng*. 2020;48:466-484.

76. Barney G, Strauss G, Anzelm L. *Theoretical Sensitivity*. University of California; 1978.
77. Strauss A, Corbin J. *Basics of Qualitative Research Techniques*. American Psychological Association; 1998.
78. Briggs DC. Comment: Making a argument for design validity before interpretive validity. *Measurement*. 2004;2:171-191.

How to cite this article: Akbar MA, Khan AA, Mahmood S, Rafi S, Demi S. Trustworthy artificial intelligence: A decision-making taxonomy of potential challenges. *Softw: Pract Exper*. 2023;1-30. doi: 10.1002/spe.3216

APPENDIX A

Literature review sources: https://tinyurl.com/4jshmme8

APPENDIX B

Questionnaire instrument: https://forms.gle/WEC9prPYfPHHos3U8

APPENDIX C

Pairwise comparison instrument: https://tinyurl.com/42tury5h