AI for All: Identifying AI incidents Related to Diversity and Inclusion

RIFAT ARA SHAMS*, CSIRO's Data61, Australia
DIDAR ZOWGHI, CSIRO's Data61, Australia
MUNEERA BANO, CSIRO's Data61, Australia

Abstract: The rapid expansion of Artificial Intelligence (AI) technologies has introduced both significant advancements and challenges, with diversity and inclusion (D&I) emerging as a critical concern. Addressing D&I in AI is essential to reduce biases and discrimination, enhance fairness, and prevent adverse societal impacts. Despite its importance, D&I considerations are often overlooked, resulting in incidents marked by built-in biases and ethical dilemmas. Analyzing AI incidents through a D&I lens is crucial for identifying causes of biases and developing strategies to mitigate them, ensuring fairer and more equitable AI technologies. However, systematic investigations of D&I-related AI incidents are scarce. This study addresses these challenges by identifying and understanding D&I issues within AI systems through a manual analysis of two AI incident databases, AI Incident Database (AIID) and AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC). The research develops a decision tree to investigate D&I issues tied to AI incidents and populate a public repository of D&I-related AI incidents. The decision tree was validated through a card sorting exercise and focus group discussions. The research demonstrates that almost half of the analyzed AI incidents are related to D&I, with a notable predominance of racial, gender, and age discrimination. The decision tree and resulting public repository aim to foster further research and responsible AI practices, promoting the development of inclusive and equitable AI systems.

JAIR Track: Fairness and Bias in AI

JAIR Associate Editor: Roberta Calegari

JAIR Reference Format:
Rifat Ara Shams, Didar Zowghi, and Muneera Bano. 2025. AI for All: Identifying AI incidents Related to Diversity and Inclusion. *Journal of Artificial Intelligence Research* 83, Article 3 (June 2025), 25 pages. DOI: 10.1613/jair.1.17806

1 Introduction

The rapid proliferation of Artificial Intelligence (AI) technologies has brought both remarkable advancements and significant challenges. Among these challenges, the issue of diversity and inclusion (D&I) has attracted considerable attention [55]. Diversity refers to human attributes and characteristics, both inherent (e.g., sex, age, or ethnicity) and acquired (e.g., education, socioeconomic status, cultural background, or lived experiences), that shape individuals and their perspectives. Inclusion, on the other hand, is the intentional effort or action to ensure that people with diverse attributes are actively represented, valued, and involved in processes, such as the design and development of AI systems. Thus, diversity describes the varied composition of individuals involved, while inclusion constitutes the deliberate actions and practices aimed at meaningfully engaging and leveraging these diverse perspectives. By incorporating a wide range of viewpoints and experiences, inclusion helps create AI systems that are more comprehensive and sensitive to the needs of different groups, possibly leading to fairer and less biased outcomes [17]. Fairness then serves as a guiding principle to evaluate whether these systems treat

[*] Corresponding Author.

Authors' Contact Information: Rifat Ara Shams, ORCID: 0000-0002-9426-3068, rifat.shams@csiro.au, CSIRO's Data61, Australia; Didar Zowghi, didar.zowghi@csiro.au, CSIRO's Data61, Australia; Muneera Bano, muneera.bano@csiro.au, CSIRO's Data61, Australia.

DOI: 10.1613/jair.1.17806

individuals equitably, preventing discrimination or unjust advantages based on protected attributes. Meanwhile, bias can arise from both social constructs and prejudices embedded in society and technical factors, such as imbalanced or incomplete data, which together can influence AI outputs in ways that disproportionately affect certain groups based on their diversity attributes [7], [56], [23].

Notably, the distinctions between inherent and acquired attributes can be nuanced and context-dependent, as social identities often intersect and interact in complex ways, influencing AI systems differently across contexts. Practical actions to enhance inclusion in AI development include participatory design methods, stakeholder engagement, and intentionally diverse recruitment practices.

Therefore, addressing D&I in AI is crucial for several reasons: it can reduce biases, increases fairness, enhances creativity, and prevents harmful societal impacts [36]. However, despite these pressing needs, D&I considerations are often overlooked in the design, development, and deployment of AI, resulting in unintended consequences and many AI incidents.

According to the OECD AI expert team, "AI incidents are defined as an event where the development or use of an AI system results in actual harm as an 'AI incident', while an event where the development or use of an AI system is potentially harmful is termed an 'AI hazard'." [29]. Based on this definition, we can define Diversity and Inclusion (D&I)-related AI incidents, which refer to situations where the functioning or outcomes of an AI system adversely affect individuals or groups based on attributes such as race, gender, age, disability, sexual orientation, ethnicity, or religion. These incidents involve outcomes that perpetuate bias, discrimination, or exclusion, thereby violating principles of equity, fairness, and inclusion in the context of AI systems.

In recent years, reported real-world AI incidents that demonstrate discrimination and bias show the necessity of integrating D&I principles in AI applications. Instances such as Google Images misrepresenting women's job roles [15] and Google Photos mistakenly categorizing images of African Americans inappropriately [22] highlight the built-in biases of AI and the complex ethical problems that come with it. Even Tinder Plus, a popular platform, encountered issues when it implemented a biased personalized pricing algorithm that disproportionately charged users over 30 and gay and lesbian users aged 18-29 [11]. Incidents of facial recognition errors leading to wrongful arrests [8], AI hiring tools biased against females [13], and medical algorithms that prioritize white patients over black patients [28] clearly indicate deep-rooted biases in AI systems. Historical biases in motion capture data, which predominantly favored able-bodied male subjects [31], further exhibit the systemic exclusion of the disabled present in AI applications. Recent incidents, such as OpenAI's ChatGPT displaying gender bias in recommendation letters [44], only underscore the pressing need for embedding D&I principles in AI. Beyond sex, age, or race, many AI incidents occur based on different diversity attributes such as ethnicity, language, religion, nationality, disability, culture, socio-economic status, geographic location and so on.

Analyzing AI incidents through a D&I lens becomes critical for several compelling reasons. Firstly, it enables us to identify and understand the underlying causes of biases and discriminatory practices in AI systems. Recognizing these causes is vital for developing strategies that mitigate such biases, ensuring that future AI technologies are fairer and more equitable. Inspired by the frequent monitoring and maintenance of systems like the Black-box flight data recorder in aviation industry [35], we posit that it is essential to adopt a similar approach in dealing with AI systems. By learning from the past mistakes, we can enhance the reliability and trustworthiness of AI systems and prevent similar incidents from occurring in the future.

Despite the need of investigating D&I-related AI incidents, to the best of our knowledge, no research has been conducted to identify D&I related AI incidents, nor to propose strategies to avoid them. An important question to ask is: What is Diversity and Inclusion in AI? Zowghi et al. defined D&I in AI with "inclusion of humans with diverse attributes and perspectives in the data, process, system, and governance of the AI ecosystem" [55]. They proposed a 5-pillar framework in the AI ecosystems to propose a holistic and sociotechnical approach to D&I in AI consideration. Shams et al. conducted a systematic literature review (SLR) to explore the challenges and corresponding solutions to address D&I in AI and enhance D&I practices by AI [36]. They utilised that 5-pillar framework of Zowghi et al. to structure and present the results of their SLR. Another research has emphasized making AI diverse and inclusive to better meet everyone's needs, respect rights, and match current societal values [21]. Chi et al. reported that big companies like Google, Microsoft, and Salesforce, while discussing diversity and inclusion in their AI ethics rules, are focusing more on technical aspects and that of fairness [12]. Another recent study identified that most guidelines for AI mainly focus on fairness, justice, and discrimination, while ignoring diversity, equity, and inclusion [10]. None of these recent research articles has focused on D&I issues in the reported AI incidents.

To fill this gap, we have worked on establishing a set of criteria to effectively identify D&I-related AI incidents. For this purpose, we manually analyzed the AI incidents from two databases (AIID[1] and AIAAIC[2]), to classify the incidents into three categories: "related to diversity and inclusion", "not related to diversity and inclusion", and "more information required" to decide. We developed a decision tree to investigate the diversity and inclusion issues in AI incidents. To validate our categorization and the decision tree, we conducted card sorting exercise and focus group discussions with artificial intelligence/machine learning (AI/ML) researchers and practitioners who also have sufficient knowledge about D&I. Finally, we have designed, populated, and made publicly available repository of D&I-related AI incidents, intended for use for future research. The key contributions of this research are:

1. **Identification of D&I-Related AI Incidents:** We established criteria and explored two AI incident databases (AIID and AIAAIC) to categorize incidents based on their relevance with D&I. We also provided the reason behind our categorization and the associated diversity attributes for each incident.
2. **Development of Analytical Tools:** We created a decision tree to investigate D&I issues in AI incidents and validated through participatory activities.
3. **Public Repository Creation:** We developed a publicly accessible repository of D&I-related AI incidents. This resource provides valuable insights for researchers, developers, and policymakers, guiding the responsible development of AI systems.
4. **Assistance of Exploring Underlying Causes of D&I-Related AI Incidents:** Our research sheds light on the causes of AI incidents related to D&I, that could assist in proposing potential strategies to avoid such incidents in future (more details in Section 5).

Paper Organization. Section 2 discusses the background of this research and the related work. In Section 3, we explain our research methodology. Section 4 reports the results of this study which we discuss in Section 5. Section 6 discusses the possible threats to validity of this research. Finally, we conclude our research with possible future research directions in Section 7.

2 Background and Related Work

The evolution of artificial intelligence (AI) [40] has permeated many domains, including health [34], education [52], transportation [2], and law [5], necessitating the development of ethical and responsible systems. Discrimination can be embedded into AI systems through various avenues such as data, design, implementation, and the absence of adequate legal frameworks [46]. Data used to train AI models may contain biases, leading to algorithmic discrimination [48]. The design of AI algorithms can inadvertently perpetuate discriminatory outcomes, even when human prejudices are intended to be eliminated. Furthermore, the implementation of AI systems without proper checks and balances can result in discriminatory decisions [19]. The lack of comprehensive legal frameworks to regulate AI and prevent discriminatory practices poses a significant challenge [49]. To address these issues, a multidisciplinary approach integrating legal and technological perspectives is crucial to develop fair and unbiased AI systems that comply with existing antidiscrimination laws [4]. According to Zhou et al., the widespread

[1] https://incidentdatabase.ai/
[2] https://www.aiaaic.org/aiaaic-repository

application of AI in various domains makes it imperative to align its operational principles with ethical standards [53]. This necessitates the establishment of guidelines and principles to ensure such systems are unbiased, trustworthy, and fair to all.

Principles of Diversity and Inclusion aim to tackle the challenges of bias and discrimination in society. Embedding D&I principles into the processes of designing, developing, and deploying AI systems is crucial to achieving equity and fairness [55]. An important aspect of integrating D&I into AI involves the identification and analysis of D&I-related AI incidents. These incidents expose the underlying biases and discrimination embedded within AI systems. Identifying these incidents enables us to understand their causes and develop strategies to mitigate them, thereby enhancing the fairness and inclusivity for future AI technologies.

2.1 Diversity and Inclusion in AI

D&I in AI is gaining increasing attention in research and practice. To achieve trustworthy AI, the importance of embedding diversity and Inclusion throughout the AI system development life cycle has been emphasized. [55]. While D&I and fairness are distinct concepts, fostering diversity can lead to fair outcomes, particularly in information access systems like recommendation systems and search engines [32]. Scholars highlight the risks of AI systems perpetuating existing inequalities, underscoring the need for responsible AI development that incorporates diversity, equity, and inclusion (DEI) principles and practices [10]. Guidelines for AI increasingly advocate for DEI principles, emphasizing the importance of addressing DEI risks through actions that influence AI actors' behaviors and awareness [10].

In order to have a comprehensive understanding of diversity in AI, it is vital to acquire an understanding of different diversity attributes (e.g., gender, age, ethnicity, race, socio-economic status, nationality, religion etc.) that necessitate careful consideration within AI systems. Zowghi et al. defined diversity attributes as "known facets of diversity, including (but not limited to) the protected attributes in Article 26 of the International Covenant on Civil and Political Rights (ICCPR) [1], as well as race, colour, sex, language, religion, national or social origin, property, birth or other status, and inter-sections of these attributes" [54]. AI's impact on various diversity attributes looks into how AI systems can either include or exclude diverse groups such as women, LGBTQI+ individuals, different races, age groups, and people with different abilities, depending on how the data is selected, how the AI is trained, and how it is ultimately used [41]. While several studies have explored gender diversity in AI, numerous other dimensions of diversity attributes have been largely overlooked in AI research [36]. Similarly, in practice, AI systems like facial recognition, voice recognition, and prediction systems have a high probability of impacting diversity attributes such as race, gender, dialects etc. For example, studies have shown that automated face recognition systems are significantly impaired by demographic attributes, that leads to a significant decrease in face recognition performance [9].

2.2 AI Incidents Related to Diversity and Inclusion

Since D&I in AI is a new and growing field of study, there is a paucity of research on this topic. This gap has led to a lack of awareness in implementing D&I in AI systems, consequently resulting in various AI incidents that could be attributed to the violation of D&I principles. There are publicly available AI incident databases such as AI Incident Database (AIID), AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC), and OECD AI Incident Database[3].

A number of recent studies have focused on AI incident databases for various purposes. For example, a recent research emphasized the importance of an AI incident database for recording and examining real-world AI failures, thus preventing recurring mistakes and ensuring AI's societal benefits [27]. Wei et al. presented a detailed analysis of real-world AI ethical issues, drawn from the AI Incident Database, identifying 13 prevalent application

[3] https://oecd.ai/en/catalogue/tools/ai-incident-database

areas and 8 forms of ethical issues with the aim to provide AI practitioners with a practice-oriented guideline for ethical AI deployment [50]. Similarly, another study addressed challenges to engineering trustworthy AI by analyzing 30 real-world incidents of trust loss from the AI incident database, offering practical recommendations to be incorporated into the development cycle in AI systems [43]. Feffer et al. suggested the AI incident database as an educational tool, highlighting its role in a study that enhanced students' understanding of AI harms and the requirement for safe and responsible AI [18]. A recent research utilized public AI incident databases to assess reporting techniques, aiming to enhance incident documentation, thus contributing to safer, fairer AI development [47].

A recent study by the Center for Security and Emerging Technology (CSET) proposed AI harm taxonomy based on AIID database that characterizes AI incidents and classifies harms of relevance to the public policy community [30]. This taxonomy addresses issues related to bias and fairness in AI and identified 11 intangible harms that closely align with our diversity attributes. However, there are fundamental differences between the CSET taxonomy and our approach. While the CSET study relied solely on the AIID database, our research incorporated both the AIID and AIAAIC databases, providing a broader perspective. Additionally, CSET's classification process involved annotators categorizing AI incidents into predefined categories, with a project lead assigning reviewers to validate these classifications. In contrast, our study adopts a more rigorous methodology. Before analyzing AI incidents, we conducted a systematic literature review (SLR) [36] to enhance our understanding of diversity and inclusion issues in AI systems. This foundational knowledge informed our approach and ensured a robust basis for our categorization process. Moreover, our categorization process was validated through participatory activities, enhancing its reliability. We also utilized a top-down approach to identify diversity attributes, beginning with 11 attributes derived from our SLR. Importantly, we remained open to identifying additional diversity attributes during the analysis of AI incidents, ultimately expanding our list to 18 attributes (e.g., age, gender, religion, culture, language). By comparison, the CSET taxonomy identified 11 diversity attributes from the AIID database, most of which overlap with our findings. This comprehensive approach highlights the greater depth and scope of our study in exploring diversity and inclusion-related AI incidents.

While several recent studies have used AI incident databases for a variety of research objectives, there appears to be a noticeable gap in current state of the art on investigating D&I issues in AI incidents. This deficit extends to inquiries into the causes of D&I-related AI incidents, as well as the formulation of strategies that could potentially prevent such incidents. To mitigate this gap, we undertook a study utilizing two AI incident databases (AIID and AIAAIC) with the aim of identifying AI incidents that are related to diversity and inclusion (D&I). Part of our initiative was to also propose methods for such identification. Further, we have developed a publicly accessible repository of D&I-related AI incidents. This under explored area of AI research, which combines D&I considerations with incident analysis, hence presents a significant opportunity for future study, and could contribute greatly to our understanding and management of D&I issues in AI.

3 Methodology

With the aim to identify AI incidents related to diversity and inclusion, we formulated the following two research questions.

RQ1. *How can we identify if an AI incident is related to diversity and inclusion issues?*

RQ2. *To what extent are the existing AI incidents related to diversity and inclusion issues?*

We conducted a mixed-methods empirical study to answer the research questions. We collected data from two public online AI incident databases, and through a card sorting exercise and two focus group discussions with artificial intelligence/machine learning (AI/ML) researchers and practitioners, who are enthusiast of D&I. Figure 1 shows an overview of our research method. As this study worked with humans, ethics approval was acquired from our organization's Human Research Ethics Committee on 19/03/2024.

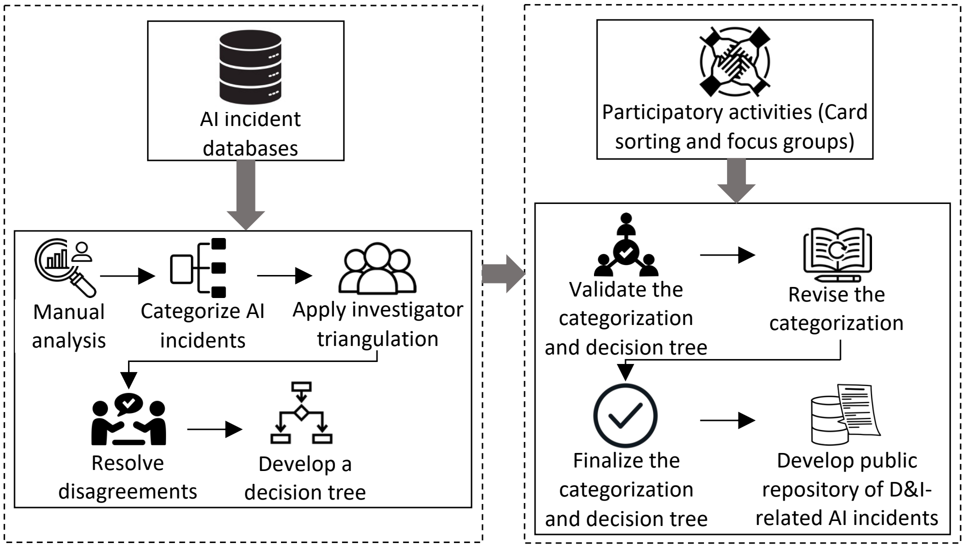

AI incident databases
Manual analysis
Categorize AI incidents
Apply investigator triangulation
Resolve disagreements
Develop a decision tree
Participatory activities (Card sorting and focus groups)
Validate the categorization and decision tree
Revise the categorization
Finalize the categorization and decision tree
Develop public repository of D&I-related AI incidents

Fig. 1. An overview of the research method

3.1 Data Collection

We collected data from AI incident databases, card sorting exercise and focus groups.

3.1.1 AI Incident Databases.

We collected AI incidents from two publicly available databases, AI Incident Database (AIID) and AI, algorithmic, and automation incidents and controversies (AIAAIC). These databases are dedicated archives, indexing the collective history of potential harms that have transpired in real-world scenarios as a result of deploying AI systems. We collected the data from AIID in July 2023 and from AIAAIC in March 2023. There were 551 AI incidents in AIID database and 575 incidents in AIAAIC. The AIID database includes the titles and summaries of the incidents including some news links, dates of the incidents, and information about the alleged/harmed/nearly harmed parties. On the other hand, AIAAIC includes the titles and summaries of the incidents along with the incident year, country, sector (e.g., media, sports, health), operator (e.g., Google, Microsoft), developer, system, technology (e.g., facial recognition, natural language processing) and corresponding news links. We conducted a detailed manual analysis of AI incidents (titles, summaries of the incidents, and alleged/harmed/nearly harmed parties) through the lens of D&I. Our goal was to determine whether any humans were harmed, if the incidents were solely due to technical failures, or if there were issues related to bias, discrimination, or violations of fairness. We also examined whether marginalized or minority groups were overlooked, and if there were any direct or indirect references to D&I issues. Specifically, we checked whether the impacted parties belonged to particular groups, such as women or children, and whether the names of the impacted individuals pointed to any diversity attributes like religion or ethnicity (e.g., Muslim, Asian). Based on this analysis, we categorized the incidents into three groups: (1) 'Related to D&I' (R), if they clearly involved diversity and inclusion factors; (2) 'Not Related to D&I' (NR), if no such factors were present; and (3) 'More Information Required' (MIR), for cases where the available data was insufficient or where there was a possibility that D&I issues might be involved but further investigation was needed. We also mentioned the reason supporting our determination for each incident and identified the diversity attributes for the D&I-related AI incidents.

3.1.2 Participants of Card Sorting and Focus Groups.

We conducted card sorting to validate our categorization process of AI incidents with their relevance with D&I issues, as card sorting is a reliable, fast, and inexpensive approach that provides comprehensive understanding on the subject matter [42]. Additionally, as focus groups are a valuable validation technique used in various research fields [26], [20], [45], we also conducted focus groups to validate our categorization process and the decision tree.

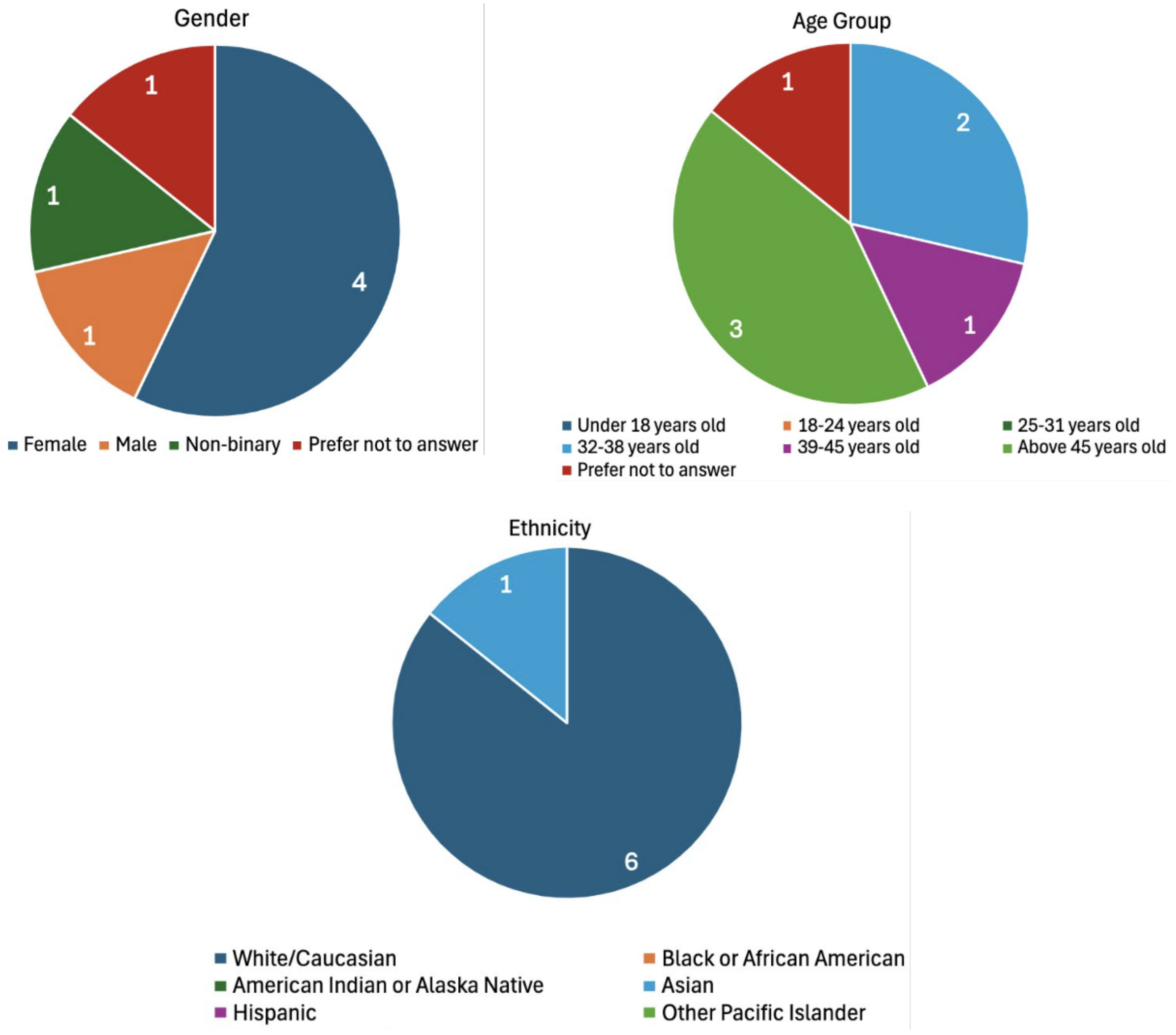


Fig. 2. Demographics of the participants

Initially, we formulated a specific set of recruitment criteria, facilitating the identification of suitable participants for the card sorting and focus groups. The main criterion to select the participants was finding practitioners who are AI researchers or AI practitioners, or information system and computer science researchers who have knowledge about diversity and inclusion. We recruited participants from diverse demographic backgrounds to ensure a broad range of perspectives and to promote unbiased, impartial feedback. This diversity helps capture varied insights and minimizes the risk of any single viewpoint dominating the results. As a result, their feedback is more representative and balanced. The demographics (gender, age group, and ethnicity) of the participants are shown in Figure 2. We shared our project objectives and participant recruitment criteria through social media (LinkedIn, Twitter (now X), Facebook) to invite participants in our study. We also invited potential participants through emails. Seven participants agreed to join the activities. All of them joined the card sorting, however, one participant did not attend the focus group. Therefore, we conducted two focus groups with three participants in each group.

At first, they were given participant information sheet to know the details of the project and their roles. It also included the risks and benefits of the project, withdrawal of participation from the project, confidentiality of their identities, data management, and contact details of the research team and human ethics research approval committee. Later, the participants were provided with a consent form and asked to give their consents by signing and returning the consent form.

3.1.3 Protocol of Card Sorting.

We conducted card sorting with seven participants. We prepared 10 cards for each participant, where each card contained one AI incident. We selected 5 incidents from AIID database and another 5 incidents from AIAAIC database that were particularly thought-provoking and represented different categories (R, NR, MIR) according to our manual analysis. As manual analysis takes time, analyzing more than 10 incidents in a participatory activity was difficult. This activity was conducted in our workplace that took approximately 60 minutes.

This method consisted of three parts. The first part was an icebreaker; the principal investigator of this study met the participants before data collection starts and spent some time with them chatting to help put them at their ease. The second part covered explaining our research, its objectives, expected outcomes, possible benefits and risks of this research, and the strategies adopted to ensure the confidentiality of the collected data. For ethical compliance, we provided the participants with a participant information sheet and sought their signed consent. The last part was the closed card sorting [25], where the participants were given the cards containing AI incidents. They were asked to group the incidents under three categories based on the relevance of the incidents with D&I issues. Each card was numbered so that they can be provided in the same order to everyone. The participants were asked to analyze each card and categorize it based on the given categories. At this stage, they were asked to make their own decision without any discussion with other participants. They were also asked to write the reasons behind their labelling decision for each incident.

3.1.4 Protocol of Focus Groups.

As focus groups are a valuable validation technique used in various research fields [26], [20], [45], we conducted two focus groups with three participants in each focus group to validate our categorization process and the decision tree. We arranged the focus groups at our workplace. The participants were provided a comfortable environment with round seating arrangements.

Each focus group was divided into two parts. In the first part, the participants were encouraged to discuss about the cards with AI incidents and the way they categorized the cards. In the second part, we provided the decision tree to the participants. We described how we developed the decision tree and how the decision tree works to identify the diversity and inclusion issues in AI incidents. Afterwards, the participants were encouraged to criticize the decision tree and provide recommendations on how to improve it based on their knowledge on the card sorting and the first part of the focus group discussion. We Asked the following questions to facilitate the focus groups.

- **Questions on Card Sorting Exercise:** How did you categorize the given AI incidents?, Have you faced any difficulties in categorizing them?, When you categorized an incident under "More information required", what information do you think should be provided to categorize them?
- **Questions on the Decision Tree:** Do you think the decision tree is aligned with your categorization? Do you think any step of the decision tree is not clear enough to understand? Do you have any comments/concerns/recommendations on the decision tree?

Given the nature of focus groups to encourage participants to have an in-depth discussion, these sessions provided an invaluable opportunity to extract the required data and validate our proposed decision tree. We arranged the two focus groups on two different days. The first focus group took 43 minutes and the second one took 47 minutes. We recorded the focus group discussions after obtaining participants' consent.

3.2 Data Analysis

Figure 1 shows the overview of the data analysis. The analysis was divided into the following steps to ensure maximum clarity and precision.

Collection and Initial Analysis of AI Incident Data. There were 551 incidents from the AI Incident Database (AIID), timestamped on 24/07/2023 and 575 incidents from AIAAIC, timestamped on 15/03/2023. The first author of this paper collected the database and manually analyzed the incidents to explore their correlation to D&I (details in Section 3.1.1). The incidents were categorized into three distinct groups based on their relation to D&I. To effectively analyze AI incidents through a D&I lens, annotators need a strong understanding of both AI and diversity and inclusion, along with relevant experience in this field. Our annotators possess these skills, ensuring the reliability of the categorization.

Investigator Triangulation. To promote confidence in the initial analysis, investigator triangulation was applied to the first 100 AIID incidents. Another researcher from our team was asked to perform an independent analysis to the incidents. We experienced 26 disagreements in categorization. However, they were resolved after having a discussion between the two investigators, allowing for a deeper understanding of the complex dynamics inherent within each case. By investigating these incidents from multiple perspectives, we increased confidence in the validity of our findings.

Re-Analyzing Incident Data and Decision Tree Development. The first author manually analyzed all the 551 AIID incidents once again taking into account the insights obtained from the discussion. As a result, the decision for 45 incidents were changed. The first author also analyzed the first 50 incidents from AIAAIC database with the knowledge developed from the analysis of AIID and the discussion with the second investigator. The structured flow of our manual analysis to determine whether AI incidents were linked to D&I issues and their categorization into three groups allowed us to the development of the initial version (V1) of a decision tree (see Figure 7 in Appendix A). The decision tree was designed to provide a clear, systematic, and step-by-step process for identifying whether an AI incident is tied to D&I issues. The decision tree's structure was informed by key categorization criteria used in our manual process. These criteria included: assessing whether any individuals were harmed by the AI system, identifying the presence of bias or discriminatory outcomes, determining if diversity-related attributes (e.g., gender, ethnicity, disability) were explicitly mentioned, and evaluating any indications or speculations of a potential relationship with D&I issues. By translating these conditions into a structured format, we created a decision tree that reflects our categorization logic. Similar to our manual categorization, our proposed decision tree also provides three categories of AI incidents based on their relevance with D&I: related to D&I, not related to D&I and more information required to establish relevance. This systematic classification aims to provide a robust tool for identifying and understanding the nuances of AI incidents through a D&I lens. More details of the decision tree is described in Section 4.

Updating Decision Tree. To ensure the validity and effectiveness of the decision tree, the preliminary analysis and the decision tree were shared with the co-authors of this paper who are also experts in D&I in AI. Their feedback and insights played a crucial role in refining the tool. Additionally, to test the practical utility of the decision tree, one of the co-authors applied it to a random sample of incidents, classifying them according to the tree's criteria. This exercise revealed areas for improvement and helped refine the decision tree further. Following several rounds of iterative discussions and updates informed by this practical testing and expert feedback, we developed an updated version (V2) of the decision tree. The revisions in V2 were guided by a deeper understanding of the incidents, ensuring the framework is both comprehensive and practical for evaluating the relationship between AI incidents and D&I concerns (see Figure 8 in Appendix A).

Organizing Participatory Activities. We organized card-sorting and focus group discussions on April 2024 with 7 and 6 participants respectively to validate our categorization and the decision tree. We randomly selected five incidents from the AIID, ensuring representation from all three categories: “related to D&I”, “not related to D&I”, or “more information required”. Additionally, we selected another five incidents from the AIAAIC, also covering all three categories. In the card sorting exercise, all participants' categorizations directly aligned with our predefined categories for three specific incidents, demonstrating a strong consensus. For each of these incidents, participants classified them in the same way we had, sorting them consistently under “related to D&I”, “not related to D&I”, or “more information required”. For five incidents, their categorizations closely aligned with our own. This alignment indicates that our categorizations were intuitive and that participants shared a common understanding of the distinctions within these categories. Different opinions came from different participants for the remaining two incidents. Despite divergent perspectives, our focus group discussions were insightful and productive, leading to enhanced clarity in our understanding of the varied perceptions. Both focus group discussions were recorded and transcribed for facilitating detailed data analysis.

Finalizing Categorization and Decision Tree. After conducting the card-sorting and focus groups, and subsequent analysis of qualitative data from participatory activities, the first author conducted a manual analysis of all 551 incidents for the third time and finalized the incident categorization. The decision for 17 incidents were updated again. Following the first focus group discussion, we updated the decision tree (V3) (see Figure 9 in Appendix A). After the second focus group discussion and the revision of the incident categorization, slight modifications were done to the decision tree once again (V4) (see Figure 10 in Appendix A). All authors engaged in another discussion after getting reviewers' feedback and decided to make slight updates to the decision tree to enhance clarity regarding its functionality. No major changes were made; only the order of the conditions and some wording were adjusted. This exercise helped us come up with the final version of the decision tree (V5) (see Figure 11 in Appendix A). Our process stopped when everyone was satisfied with the final decision tree.

Applying the Decision Tree on Another Database. The first author applied the decision tree manually for the first 310 incidents from AIAAIC database and identify the D&I-related AI incidents. However, we will analyze all the incidents from AIAAIC and populate our repository with new incidents in future.

Public Repository Creation. Finally, we created a public repository of D&I-related AI incidents for both AIID and AIAAIC. This open-source repository will serve as a valuable resource for other researchers in this field. However, the repository requires continuous updates as new incidents are constantly being added to the incident databases.

4 Results

This section presents the results derived from the analysis of the AI incident database. The results also reflects the analytical overview of the card sorting exercise and focus group discussions carried out with AI/ML researchers and practitioners.

4.1 RQ1: Identifying AI Incidents Related to D&I Issues

Figure 3 shows the results of RQ1. To identify Diversity and Inclusion (D&I) related AI incidents, we developed a decision tree after a rigorous analysis.

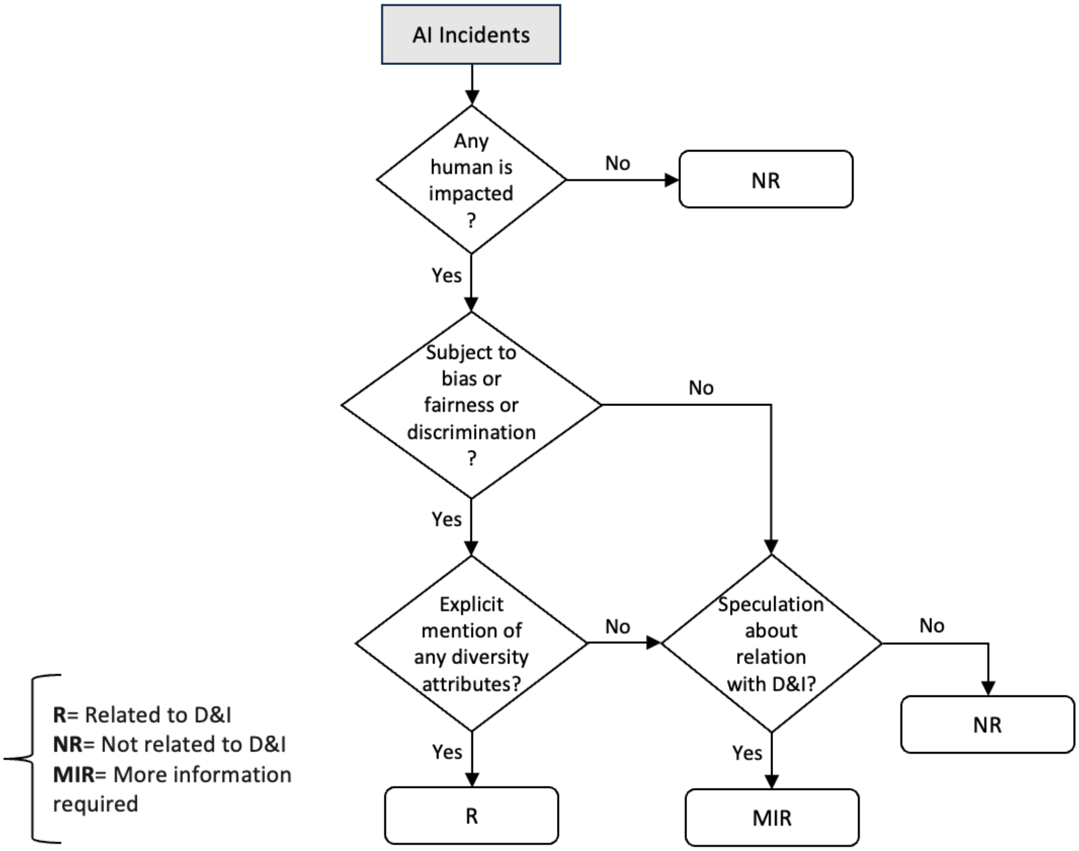

AI Incidents
Any human is impacted ?
No
NR
Yes
Subject to bias or fairness or discrimination ?
No
Yes
Explicit mention of any diversity attributes?
No
Speculation about relation with D&I?
No
NR
Yes
R
Yes
MIR
R= Related to D&I
NR= Not related to D&I
MIR= More information required

Fig. 3. Decision tree to detect D&I-related AI incidents

Our proposed decision tree has four conditions. The first condition checks if any human is directly or indirectly impacted by the AI incident. The impact can be physical, psychological, financial and so on. If the response is negative, the incident is not related to D&I. On the other hand, a positive response directs the process to the second condition, which examines whether the AI incident is subject to bias or violates fairness or results in discrimination. If the result is 'yes,' there is a strong likelihood that the AI incident is related to D&I. Therefore, the incident is scrutinized under the third condition, which examines whether the AI incident explicitly mentions about any diversity attributes which is responsible for bias or discrimination, including but not limited to gender, sex, sexual orientation, age, skin tone, race, ethnicity, religion, language, literacy, disability, neurodiversity, facial features, physical features, nationality, accent, culture, geographic location, socio-economic status, and political ideology. If the answer of this condition is a 'yes', the incident is clearly related to D&I issues. However, a negative determination on both the second and third conditions points towards a lower likelihood of direct D&I applicability. Nonetheless, a final check must be conducted on whether we can speculate any relationship of the incident with D&I issues. For example, if the impacted parties represent any diverse communities such as a specific gender or a particular age group or if the name(s) of the impacted parties highlight any diversity attributes, such as association with a particular religious faith, then we need more information to determine if the incident occurred due to the violation of these diversity attributes. Therefore, a positive outcome of this condition needs more investigation to ascertain any potential correlation with D&I issues, whereas a negative implies an absence of a D&I implication.

For instance, if a child experiences harm solely because of a malfunction in the AI system, without any element of bias or discrimination targeting children, the second criterion would not be met. In such a case, the incident would be regarded as an unintended technical failure rather than an issue of discriminatory impact. Consequently, the evaluation process would bypass this condition and proceed to examine the fourth condition, whether we have any speculation of the relationship of this incident with D&I. If we get positive result for this condition, we need additional information before drawing a definitive conclusion about its relevance or irrelevance to D&I. However, if the answer is a 'no', then the incident is not related to D&I issues.

Figure 4 presents several illustrative examples of AI incidents to elucidate the procedure of identifying their association with D&I issues. Figure 4 (a) demonstrates the trajectory of an AI incident within the decision tree that ultimately leads to the conclusive result of "related to D&I". The AI incident states:

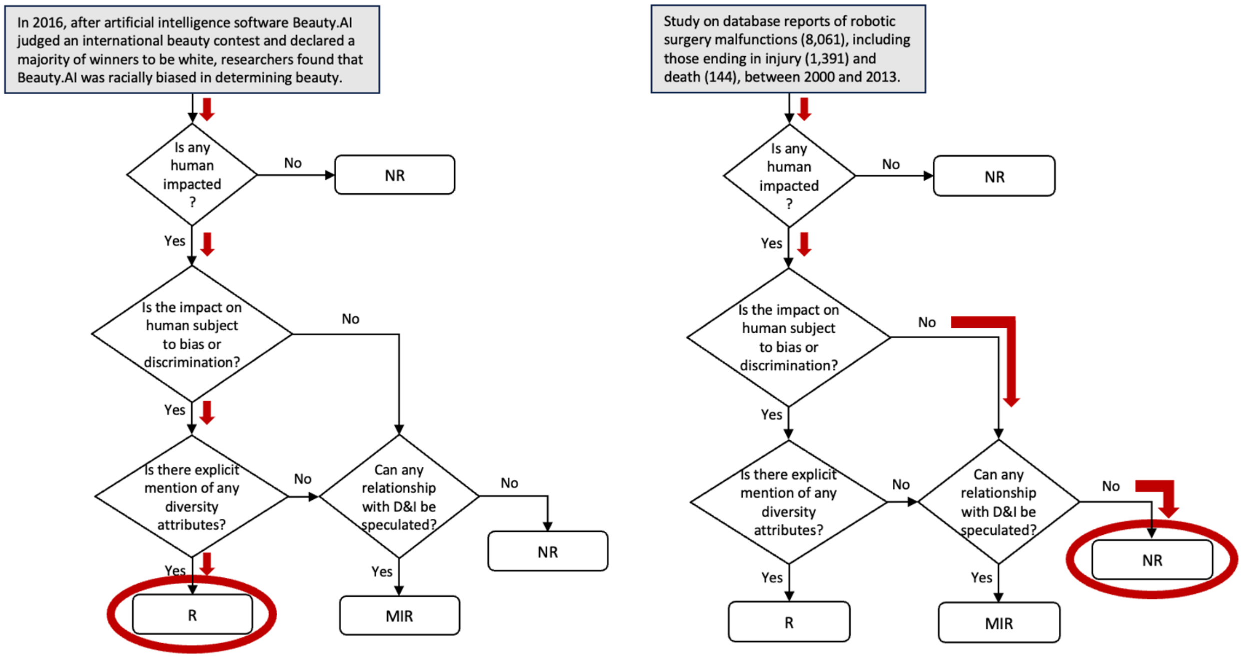
> *"In 2016, after artificial intelligence software Beauty.AI judged an international beauty contest and declared a majority of winners to be white, researchers found that Beauty.AI was racially biased in determining beauty."*

As humans are impacted through this AI incident, the process moves forward to the second condition and examines whether this incident is subject to bias or discrimination. Since the incident clearly indicates the presence of bias, the answer will be 'yes', and the decision pathway moves forward to the third condition to check if this incident is explicitly related to any diversity attributes. Since this incident clearly has racial undertones, we can conclude that the incident has relevance to D&I issues. Similarly, Figure 4 (b) shows an example of an AI incident that is not related to D&I issues.

> *"Study on database reports of robotic surgery malfunctions (8,061), including those ending in injury (1,391) and death (144), between 2000 and 2013."*

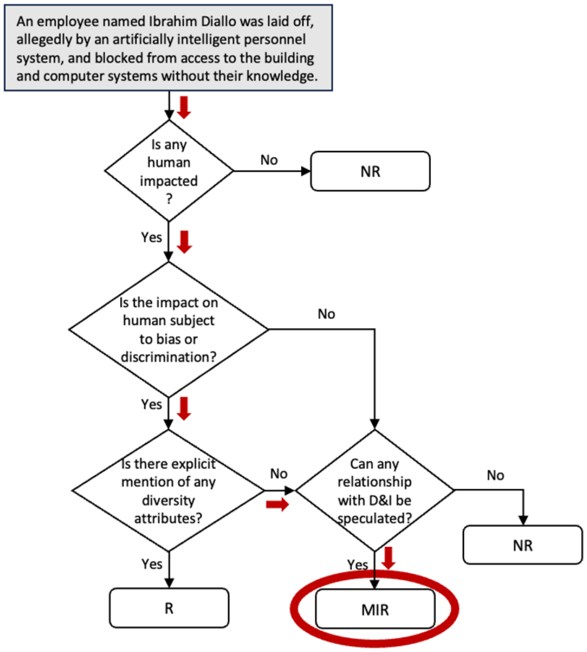
The above-mentioned AI incident evidently caused harm on humans, therefore, the procedural control moves towards the next condition of the decision tree. Here, the incident is examined if it represents any bias, breaches fairness or carries discriminatory implications. In absence of clear indications, the control moves towards the fourth condition to assess whether the incident can be speculated to be related to D&I issues. Since the outcome remains negative, the final determination categorizes the incident as "not related to D&I". Similarly, Figure 4 (c) demonstrates the navigation of another AI incident through the decision tree, leading to an outcome of "more information required".

> *"An employee named Ibrahim Diallo was laid off, allegedly by an artificially intelligent personnel system, and blocked from access to the building and computer systems without their knowledge."*


In 2016, after artificial intelligence software Beauty.AI judged an international beauty contest and declared a majority of winners to be white, researchers found that Beauty.AI was racially biased in determining beauty.
Is any human impacted ?
No
NR
Yes
Is the impact on human subject to bias or discrimination?
No
Yes
Is there explicit mention of any diversity attributes?
No
Can any relationship with D&I be speculated?
No
NR
Yes
R
Yes
MIR
Study on database reports of robotic surgery malfunctions (8,061), including those ending in injury (1,391) and death (144), between 2000 and 2013.
Is any human impacted ?
No
NR
Yes
Is the impact on human subject to bias or discrimination?
No
Yes
Is there explicit mention of any diversity attributes?
No
Can any relationship with D&I be speculated?
No
NR
Yes
R
Yes
MIR

(a) Illustration of the functioning procedure of the decision tree for an AI incident that is related to diversity and inclusion.

(b) Illustration of the functioning procedure of the decision tree for an AI incident that is not related to diversity and inclusion.


An employee named Ibrahim Diallo was laid off, allegedly by an artificially intelligent personnel system, and blocked from access to the building and computer systems without their knowledge.
Is any human impacted ?
No
NR
Yes
Is the impact on human subject to bias or discrimination?
No
Yes
Is there explicit mention of any diversity attributes?
No
Can any relationship with D&I be speculated?
No
NR
Yes
R
Yes
MIR

(c) Illustration of the functioning procedure of the decision tree for an AI incident that requires more information to understand its relevance to diversity and inclusion.

Fig. 4. Illustration of the functioning procedure of the decision tree to understand its relevance to diversity and inclusion with some examples.

The indicated incident certainly impacted humans, prompting the control to proceed towards the next condition of the decision tree to examine whether the incident exhibits bias, violates fairness, or results in discrimination. As this incident clearly breaches fairness, the process moves to the next condition to assess any direct relevance with diversity attributes. When no explicit association is uncovered concerning diversity attributes, the control advances to the next condition to explore whether there is any speculation about a possible association of the incident with D&I issues. Since the impacted individual's name is Ibrahim Diallo, which is commonly associated with a specific religious group (particularly Muslims), we speculate that there might be a possibility this incident occurred due to religious reasons. Therefore, it is essential to investigate whether the incident is genuinely related to religion or if it is merely coincidental. Hence, further information is needed to accurately assess the incident's relevance to D&I.

4.2 RQ2: Extent to Which the Existing AI Incidents are Related to D&I issues

Of the 551 AI incidents extracted from the AIID database, our analysis identified 189 AI incidents related to D&I issues (see Table 1). 80 incidents require further information to establish their correlation with D&I issues. The rest of the 282 incidents are not related to D&I. Similarly, from AIAAIC database, we identified almost half of the incidents are related to D&I issues, 144 D&I-related AI incidents among the 310 incidents we analyzed (see Table 2). 50 incidents require more information to make determination. The rest of the 116 incidents are not related to D&I issues.

We developed the first version of D&I-related AI incidents repository based on the AIID and AIAAIC databases [37]. Each record in this repository encompasses details such as the incident ID, title, description, date, alleged

Table 1. The extent to which the existing AI incidents are related to D&I issues from AIID Database

Status	No. of AI incidents	Percentage
Related to D&I	189	34.3%
Not related to D&I	282	51.18%
More information required	80	14.52%

Table 2. The extent to which the existing AI incidents are related to D&I issues from AIAAIC Database

Status	No. of AI incidents	Percentage
Related to D&I	144	46.45%
Not related to D&I	116	37.42%
More information required	50	16.13%

deployer of AI system, alleged developer of AI system, and alleged harmed or nearly harmed parties and so on. Each incident ID is linked to a corresponding article for a more detailed account of the incident. Further, in the repository, we added the status in relation to D&I, diversity attributes such as age, gender, ethnicity, race and so on, and the reason supporting our determination. We have outlined two categories under the "status" column to categorize the incidents: related to D&I (R) and more information required (MIR). Let us consider an example.

> *"Google's Perspective API, which assigns a toxicity score to online text, seems to award higher toxicity scores to content involving non-white, male, Christian, heterosexual phrases."*

We labeled the above-mentioned AI incident as being "related to D&I". The reason for our decision was "Google's API provided higher toxicity scores to non-white, male, Christian, heterosexual phrases. This is a clear breach of D&I". The diversity attributes involved in this incident encompassed race, gender, sexual orientation, and religion. Similarly, we have another instance of an AI incident stating:

> *"An algorithm used to rate the effectiveness of school teachers in New York has resulted in thousands of disputes of its results."*

The aforementioned AI incident is classified as "more information required". Our reasoning for this decision was articulated as such: "The situation might involve bias or fairness problems if the algorithm was trained with skewed data or overlooked factors like class size or student backgrounds, possibly leading to unfair teacher ratings. However, disputes might stem from disagreement with outcomes, not algorithm unfairness or discrimination, so more details are needed to confirm if the algorithm is biased".

Figure 5 shows the proportion of 16 different diversity attributes implicated in AI incidents correlated with D&I from the database, AIID. In our approach, we employed a top-down taxonomy to identify diversity attributes systematically. Drawing from one of our recent studies, we initially established a framework comprising 11 diversity attributes identified through a comprehensive systematic literature review [36]. These attributes provided a solid foundation for our analysis. However, while examining AI incident databases, we adopted a flexible and open-ended coding approach. This allowed us to remain receptive to uncovering additional diversity attributes beyond those identified in our prior work. By combining a structured taxonomy with exploratory analysis, we aimed to ensure a thorough and nuanced understanding of diversity attributes as they emerge in real-world AI incidents. With this process, we identified seven additional diversity attributes. They are 'religion', 'culture', 'facial features', 'nationality', 'accent', 'political ideology', and 'literacy'. All the attributes have their unique characteristics. For example, 'facial features' refer to the distinct parts or characteristics of a person's face that

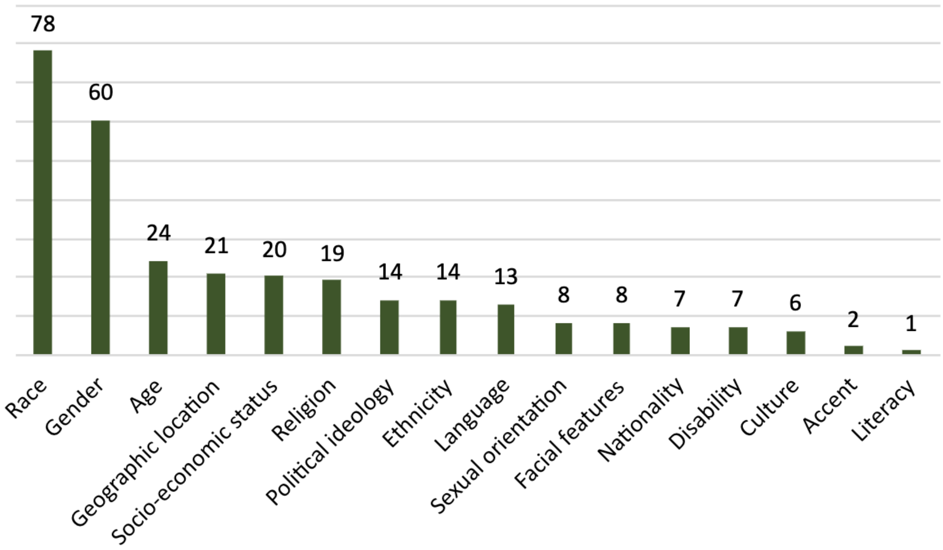


Fig. 5. The ratio of diversity attributes in AI incidents from AIID database

contribute to their appearance. These include both structural (e.g., eyes, nose, mouth, ears) and expressive elements (e.g., smile lines, eyebrows and their movement) [16].

The majority of the AI incidents associated with D&I arose from discrimination based on 'race'. In total, 'race' is directly linked to 78 distinct incidents. The second highest diversity attributes is 'gender', which is indicated in 60 AI incidents. Substantial proportions of AI incidents are also related to attributes such as 'age', 'geographic location', 'socio-economic status', and 'religion', associated with 24, 21, 20, and 19 incidents, respectively.

Similarly, the ratio of 15 different diversity attributes for the D&I-related AI incidents from AIAAIC database are shown in Figure 6. Similar to AIID, this database also has the maximum incidents based on 'race'; 62 incidents occurred due to racial bias. The second highest diversity attribute for AIAAIC is 'age', associated with 40 incidents. Bias based on 'gender' and 'ethnicity' linked to same number of incidents (36). Additionally, significant number of incidents occurred due to bias on 'socio-economic status' and 'geographic location', 26 and 23 incidents respectively.

5 Discussion and Implications

This section discusses the findings of this study as well as the implications for research and practice.

First Step Towards Identifying Cause of D&I-related AI Incidents. Identifying D&I-related AI incidents is the first step towards further investigations on these incidents to explore the underlying causes of them and identify potential strategies to avoid them. To accurately identify D&I-related AI incidents, we iteratively constructed a decision tree to systematically assess the D&I issues of an AI incident. This will guide further research trying to better understand the causes of these incidents. However, the decision tree can only provide preliminary insights unless it is complemented by a thorough causal analysis as well as strategies to avoid future AI incidents based on D&I issues.

The Role of AI in D&I Issues. Our research indicates a significant number of AI incidents linked to D&I challenges. Out of 551 incidents from AIID, 189 (34.3%) were found to be related to D&I issues, and out of 310 incidents from AIAAIC, 144 (46.45%) were related to D&I issues. This finding underscores the role AI plays

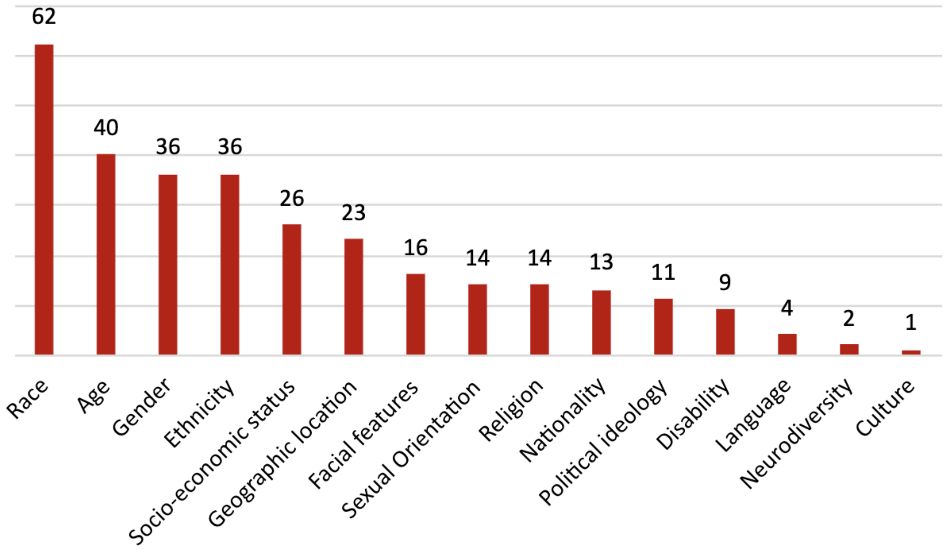


Fig. 6. The ratio of diversity attributes in AI incidents from AIAAIC database

in amplifying or generating D&I concerns. Careful attention must be paid to D&I during AI system design, development and deployment. Otherwise, these systems may have the potential to cause further harm due to biases inherent in their design and deployment.

The Uncertainties Around AI and D&I. Despite our decision tree, considerable uncertainty remains, as 80 out of 551 incidents from AIID and 50 out of 310 incidents from AIAAIC still require additional information to classify. This shows the complexity involved in dealing with D&I in AI, affirming the need for deeper analysis and further research.

Necessary Information for Categorizing AI Incidents. Due to the lack of adequate information, significant number of incidents were categorized under "more information required". Thus, the summaries of the AI incidents need to be more informative and comprehensive to facilitate the determinations regarding any potential connections to D&I issues. If data privacy rules allow it, the summaries should incorporate the names of the accused or harmed parties that might signal elements of diversity, such as religion, ethnicity, race, gender, or age. The objective is to determine whether the incident was unintentional or incited by biases. Equally, if an AI system wrongly impacted an individual, it is crucial to establish whether any biases played a role in this incident.

Bias in AI systems can manifest through various pathways, significantly impacting the fairness and reliability of the outcomes. One primary source of bias is the training data [24]. Another critical aspect is the potential bias within the models or algorithms [38]. Design choices, such as the selection of features or the weighting mechanisms in algorithms, can introduce or exacerbate biases [33]. Furthermore, there are instances where issues within AI systems arise from genuine bugs independent of any biases. These bugs could result from coding errors, inadequate system testing, or unforeseen interactions within the AI components, impacting the model's functionality without being inherently related to diversity and inclusion concerns [3]. Therefore, while biases are often linked to D&I, it is important to recognize that not all problems in AI systems come from these biases. Some could be purely technical defects, necessitating comprehensive debugging and optimization processes. However, if there is evidence of bias, discrimination, or unfairness, the next step is to pinpoint if these are grounded in D&I

concerns. If the incident is linked to D&I issues, it is necessary to determine whether it was the result of AI error or human interference operating via the AI platform.

Regular Updates of the Repository. Developing a public repository of the categorized D&I-related AI incidents is important to provide insights into the implications and assist in raising awareness of D&I issues among AI researchers and developers. As AI continues to evolve and be utilized across various domains, the frequency of incidents involving AI is also rising. Moreover, AI incident databases are regularly updated with new occurrences. Therefore, this repository should be continuously updated to remain relevant and beneficial.

Significant Proportion of Racial, Gender, and Age Discrimination. The attribute most frequently linked to discrimination is 'race', with it being implicated in a total of 78 distinct incidents from AIID and 62 incidents from AIAAIC. The prominence of race-related issues in AI applications raises serious concerns about discriminatory practices and cultural biases taking root in the development phase. Following race, 'gender' is the second most significant attribute from AIID with respect to discrimination, with a total of 60 documented cases. It demonstrates that AI systems might not be treating all genders equally, which can lead to serious repercussions in systems where accuracy or fairness is imperative. 'Age' is also another diversity attribute which is mentioned 40 times in AIAAIC. The high proportion of AI incidents involving racial, gender, and age discrimination highlights an important aspect of how these attributes are represented and examined within datasets and algorithmic fairness research [39]. Attributes such as gender, age, and race are not only more readily available in most datasets but are also more frequently the focus of fairness studies. Consequently, these characteristics are more likely to reveal patterns of discrimination, both because they are more explicitly recorded and because they are regularly scrutinized in fairness evaluations. Consequently, these attributes tend to reveal violations more prominently, making them some of the most frequently identified categories of discrimination in AI systems. While it is important to consider all diversity attributes we identified in this study to prevent future occurrences of D&I-related AI incidents, particular emphasis should be placed on addressing 'race', 'gender', and 'age' diversity within AI systems.

Roadmap for Future Research. Mature industrial sectors (e.g., aviation) capture their real-world failures in incident databases to inform safety improvements. AI systems currently cause real-world harm without a collective and systematic analysis of the causes of their failures. Many of these incidents are attributed to bias, unfairness, and violations of diversity and inclusion principles. As a result, companies repeatedly make the same mistakes in the design, development, and deployment of AI systems. What is lacking is a systematic approach for the thorough analysis of AI incidents to interrogate the causes, learn from mistakes, and provide actionable recommendations for all the relevant stakeholders in the AI ecosystem.

We propose to address this challenge by developing an evidence-based framework for operationalizing diversity and inclusion in AI. This framework consists of three stages: monitoring, analysis, and investigation. The main aim is to develop methodologies to analyze and report the causes of AI incidents stemming from D&I violations. The monitoring stage refers to the development of a portal that allows (a) others to report D&I-related AI incidents, (b) monitors and collects D&I-related incidents from news media and other social and public discourse, and (c) monitors the online AI incident databases for such entries. The analysis stage involves the development of sophisticated methods for deeper analysis of the D&I-related AI incidents and preparing a structured report that contains all the pertinent information for future developers. The last stage is about development of specific criteria for investigating the incident and developing a repository of actionable recommendations. This framework promotes transparency and trust in AI systems and supports an evidence-based AI impact assessment for the risks associated with D&I violations in AI lifecycle. Inspired by the Black Box flight recorder concept in the aviation industry, we aim to create a similar approach in the AI ecosystem with our proposed framework to monitor, analyze, and investigate AI incidents to offer clear recommendations for improving AI safety and to continuously contribute to avoiding similar AI incidents.

In this paper, we have presented the first step in developing our framework. Our work establishes a robust and repeatable technique for identifying D&I-related AI incidents and creating and making an online repository of the already-reported AI incidents available for researchers and practitioners. By making the repository accessible, it is hoped that the research community can extend the analysis beyond this initial examination. It further opens the door for community contributions and validations, facilitating a more robust understanding of D&I in AI. We plan to use the power of generative AI to automate activities in various stages of our framework for reporting and recommendations.

Actionable Next Steps. The next phase of this research project will focus on leveraging the decision tree and repository to develop a comprehensive framework that provides actionable guidelines for addressing D&I concerns in AI system design and deployment. By systematically analyzing the underlying causes of past incidents and identifying trends, this framework aims to inform preventative measures, improve risk assessment, and support the creation of inclusive AI technologies. The specific next steps include:

(1) ***Refinement of the Decision Tree:*** Building on the current validation through card sorting and focus group discussions, we plan to refine the decision tree further with additional expert input. This step will ensure its robustness and usability for analyzing a broader range of D&I-related AI incidents.
(2) ***Expansion of the Repository:*** We aim to continuously expand the public repository of D&I-related AI incidents by incorporating new cases from diverse contexts and domains. This repository will act as a knowledge base for identifying patterns, root causes, and systemic issues in AI design and deployment.
(3) ***Framework Development:*** We will develop a comprehensive framework for addressing D&I concerns in AI systems using insights derived from the decision tree and repository. This framework will include practical guidelines [55], tools, and methodologies [6] to support developers and organizations in designing inclusive and equitable AI technologies.
(4) ***Engagement with Stakeholders:*** To promote adoption, we will collaborate with AI practitioners, policymakers, and researchers to disseminate the framework and decision-making tools. This collaboration will include workshops, training sessions, and outreach activities to ensure the tools and guidelines are actionable and widely implemented.
(5) ***Testing and Iteration:*** The framework and tools will be tested in real-world AI projects to assess their effectiveness in mitigating D&I issues and preventing incidents. Feedback from these implementations will inform iterative improvements to the framework.

By following this roadmap, we aim to bridge the gap between analyzing past incidents and proactively preventing future ones.

6 Threats to Validity

This section discusses the possible threats arising from this research based on the four validation criteria: credibility, confirmability, dependability, and transferability [14].

Credibility. One potential threat to credibility could stem from the fact that the AI Incident Databases (AIID and AIAAIC), which were our primary data source, may not be an exhaustive list of AI incidents related to D&I issues. Unreported incidents, and incidents that were incorrectly reported or categorized in the database could be overlooked in the study. Furthermore, incidents reported in the database may carry a certain level of bias as they may disproportionately represent incidents from specific sectors, regions, or communities. We could overcome this threat by including other databases such as OECD AI incident database in the second version of our repository, which is our future work.

Confirmability. Confirmability refers to the degree of neutrality in the research findings. A significant threat to confirmability could arise from the manual analysis of incidents. Personal bias, misunderstanding or misinterpretation of information could influence our analysis and the development of the decision tree. Personal interests and preconceptions may have influenced the assessments and categorizations of incidents. Similarly, the decision-tree may have been influenced by the investigators' understanding and interpretation of D&I issues. However, investigator triangulation and focus groups have been conducted to minimize potential bias. Furthermore, the investigators come from diverse ethnic, cultural, racial, and age groups, which enhances their perspective on diversity and inclusion.

Another potential threat could be raised from the number of participants in the card sorting and focus groups. Although the number of the participants was small, they represented some aspects of diversity as shown in Figure 2. In general, in card sorting exercises, there are not many participants, however, they are conducted for validation of the categorization and the usefulness of the decision tree. We aim to engage more practitioners to test our decision tree in our future work.

Dependability. Dependability concerns the repeatability of the research findings in similar contexts. There exists a threat to the dependability of our study owing to it's largely qualitative nature. Our categorization, decision tree development, and subsequent analysis are driven mostly by interpretation, which may vary among different researchers. Additionally, the general dynamics of AI and D&I issues are rapidly evolving, which may invalidate our decision tree and categorization over time. Therefore, it is essential that we persistently update and enrich our repository with new AI incidents. Additionally, we are planning to create the second version of our D&I-related AI incidents database by merging data from other AI incident databases such as OECD.

Transferability. Transferability refers to the extent to which the findings can be applied in different contexts. The D&I issues are not universally defined and may vary across different cultures, communities, and regions. Therefore, the AI incidents identified as related or unrelated to D&I issues in our study may not be perceived the same way in other contexts. However, we proposed the methods to identify D&I issues in AI incidents, that mitigate the threat. The methods can be applied in any AI incidents from different cultural settings.

Moreover, the absence of a universal D&I or bias/fairness framework makes it challenging to create a decision tree that can comprehensively guide the ethical assessment of incidents across diverse cultural and moral contexts. Therefore, our decision tree operates under certain implicit assumptions about what constitutes discrimination or bias, which may not fully capture the nuances of different societal values or ethical perspectives. While we have considered AI incidents from diverse global contexts and used them to inform the development of the decision tree, we recognize that cultural and moral differences could still influence the interpretation of discrimination. To further address this, future iterations of the decision tree may benefit from the addition of region-specific guidelines or cultural adaptability features that ensure a more context-aware ethical assessment framework.

In this paper, we have not addressed any subjective or philosophical discussions related to moral concepts. Instead, our focus is rooted in addressing tangible and universally recognized principles outlined in the Charter of Human Rights [51]. These principles encompass critical aspects such as disability, sexual orientation, gender, age, race, ethnicity, and religion. Our primary objective is to ensure that AI systems are designed, developed, and deployed in a manner that upholds these rights and avoids perpetuating or amplifying discrimination against any individual or group. With this objective, we aim to contribute to the creation of equitable and inclusive AI systems that respect and protect the dignity and rights of all individuals.

7 Conclusions and Future Work

With the aim to develop more inclusive, unbiased, and trustworthy AI systems, this study is a critical first step in identifying D&I-related incidents in AI and understanding the extent to which D&I issues exist within AI systems. We developed a decision tree as a preliminary framework for identifying and categorizing AI incidents in terms of their relation to D&I issues. We also proposed and populated a public repository on D&I-related AI incidents. The AI systems were found to have a significant association with D&I issues, with 34.3% and 46.45% of the analyzed incidents being related to D&I from AIID and AIAAIC databases respectively. This emphasizes the need for careful attention to D&I during AI system design, development, and deployment. Moreover, the lack of information in some cases also complicates the task of categorizing AI incidents, making it difficult to draw definitive conclusions regarding their D&I implications. Hence, a more informative and comprehensive representation of AI incidents is required. Despite the comprehensive consideration of 16 diversity attributes, findings indicated a concerned prominence of racial, gender, and age discrimination in AI incidents. It underlines the urgency to address ‘race’, ‘gender’, and ‘age’ biases in AI system development, while not undermining the criticality of other attributes.

This study also provides a roadmap for future research in D&I within AI. Given the dynamic nature of AI, there is a continuous need to populate and revise the repository with new incidents for timely research investigations. Additionally, we must analyze all the 575 existing AIAAIC incidents in future and include the D&I-related AI incidents in our repository. Future studies should also try to understand why some AI incidents needed “more information”. This could help figure out exactly what challenges or complications are making it difficult to clearly classify these incidents. Additionally, further research is essential to develop comprehensive guidelines, and concrete strategies to prevent the occurrence of D&I-related AI incidents. Therefore, comprehensive guidelines are also necessary to embed D&I principles into AI’s design, development, and deployment.

References

[1] 2014. Australian Human Rights Commission: A quick guide to Australian discrimination laws. (2014).

[2] Rusul Abduljabbar, Hussein Dia, Sohani Liyanage, and Saeed Asadi Bagloee. 2019. Applications of artificial intelligence in transport: An overview. *Sustainability* 11, 1 (2019), 189.

[3] Lavisha Aggarwal and Shruti Bhargava. 2023. Fairness in AI Systems: Mitigating gender bias from language-vision models. *arXiv preprint arXiv:2305.01888* (2023).

[4] Alba Soriano Arnanz et al. 2023. Creating non-discriminatory Artificial Intelligence systems: balancing the tensions between code granularity and the general nature of legal rules. *IDP. Revista de Internet, Derecho y Política* 38 (2023), 1–12.

[5] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. *Artificial Intelligence* 289 (2020), 103387.

[6] Muneera Bano, Didar Zowghi, Fernando Mourao, Sarah Kaur, and Tao Zhang. 2024. Diversity and Inclusion in AI for Recruitment: Lessons from Industry Workshop. *arXiv preprint arXiv:2411.06066* (2024).

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities.* MIT press.

[8] J. Brodkin. 2023. Black man wrongfully jailed for a week after face recognition error, report says. https://arstechnica.com/tech-policy/2023/01/facial-recognition-error-led-to-wrongful-arrest-of-black-man-report-says/.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency.* PMLR, 77–91.

[10] Gaelle Cachat-Rosset and Alain Klarsfeld. 2023. Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. *Applied Artificial Intelligence* 37, 1 (2023), 2176618.

[11] Chiara Cavaglieri. 2022. Tinder’s unfair pricing algorithm exposed. https://www.which.co.uk/news/article/tinders-unfair-pricing-algorithm-exposed-adCwG8b7VRYo.

[12] Nicole Chi, Emma Lurie, and Deirdre K Mulligan. 2021. Reconfiguring diversity and inclusion for AI ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 447–457.

[13] J. Cook. 2018. Amazon ditches AI recruitment tool that ’learnt to be sexist’. https://www.afr.com/world/europe/amazon-ditches-ai-recruitment-tool-that-learnt-to-be-sexist-20181011-h16h8p.

[14] Daniela S Cruzes and Tore Dyba. 2011. Recommended steps for thematic synthesis in software engineering. In *2011 international symposium on empirical software engineering and measurement.* IEEE, 275–284.

[15] Andrew Van Dam. 2019. Searching for images of CEOs or managers? The results almost always show men. https://www.washingtonpost.com/business/2019/01/03/searching-images-ceos-or-managers-results-almost-always-show-men/.

[16] Liya Ding and Aleix M Martinez. 2010. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 11 (2010), 2022–2038.

[17] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (2022), 20539517221082027.

[18] Michael Feffer, Nikolas Martelaro, and Hoda Heidari. 2023. The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. In *Proceedings of the 3rd ACM Conference on Equity*

and Access in Algorithms, Mechanisms, and Optimization. 1–11.

[19] Xavier Ferrer, Tom Van Nuenen, Jose M Such, Mark Coté, and Natalia Criado. 2021. Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine* 40, 2 (2021), 72–80.

[20] Derek Fisk, Ben Clendenning, Philip St. John, and Jose Francois. 2024. Multi-stakeholder validation of entrustable professional activities for a family medicine care of the elderly residency program: A focus group study. *Gerontology & Geriatrics Education* 45, 1 (2024), 12–25.

[21] Eduard Fosch-Villaronga and Adam Poulsen. 2022. Diversity and inclusion in artificial intelligence. *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice* (2022), 109–134.

[22] Alex Hern. 2018. Google's solutions to accidental algorithmic racism: ban gorillas. https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people.

[23] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–16.

[24] J Kavitha, J Sasi Kiran, Srisailapu D Vara Prasad, Krushima Soma, G Charles Babu, and S Sivakumar. 2022. Prediction and Its Impact on Its Attributes While Biasing MachineLearning Training Data. In *2022 Third International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE).* IEEE, 1–7.

[25] Joanna Kerr, Katerina Hilari, and Lia Litosseliti. 2010. Information needs after stroke: What to include and how to structure it on a website. A qualitative study using focus groups and card sorting. *Aphasiology* 24, 10 (2010), 1170–1196.

[26] Peter A Lichtenberg and Susan E Macneill. 2000. Prospective validity study of a triaging method for mental health problems: The MacNeill-Lichtenberg Decision Tree (MLDT). *Clinical gerontologist* 21, 1 (2000), 11–19.

[27] Sean McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15458–15463.

[28] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[29] OECD. 2024. Defining AI Incidents and Related Terms. *OECD Artificial Intelligence Papers, No. 16, Paris* (2024).

[30] Kevin Paeth, Daniel Atherton, Nikiforos Pittaras, Heather Frase, and Sean McGregor. 2024. Lessons for Editors of AI Incidents from the AI Incident Database. *arXiv preprint arXiv:2409.16425* (2024).

[31] J. Pepitone. 2024. AI is being built on dated, flawed motion-capture data, IEEE Spectrum. https://spectrum.ieee.org/motion-capture-standards.

[32] Lorenzo Porcaro, Carlos Castillo, Emilia Gómez, and João Vinagre. 2023. Fairness and diversity in information access systems. *arXiv preprint arXiv:2305.09319* (2023).

[33] Valentina Pyatkin, Frances Yung, Merel CJ Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics* 11 (2023), 1014–1032.

[34] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature medicine* 28, 1 (2022), 31–38.

[35] Jeremy Sear. 2001. The ARL 'Black Box'Flight Recorder–Invention and Memory. *Bachelor of Arts (Honours). The University of Melbourne* (2001).

[36] Rifat Ara Shams, Didar Zowghi, and Muneera Bano. 2023. AI and the quest for diversity and inclusion: a systematic literature review. *AI and Ethics* (2023), 1–28.

[37] Rifat Ara Shams, Didar Zowghi, and Muneera Bano. 2024. *Diversity and Inclusion (DI)- Related AI Incidents Repository.* https://doi.org/10.5281/zenodo.11639709

[38] Donghee Shin and Emily Y Shin. 2023. Data's Impact on Algorithmic Bias. *Computer* 56, 6 (2023), 90–94.

[39] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy data practices harm fairness research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 642–659.

[40] RM Singari and PK Kankar. 2022. Contemporary Evolution of Artificial Intelligence (AI): An Overview and Applications. *Advanced Production and Industrial Engineering: Proceedings of ICAPIE 2022* 27 (2022), 130.

[41] Roger Søraa. 2023. *AI for Diversity.* CRC Press.

[42] Donna Spencer and Todd Warfel. 2004. Card sorting: a definitive guide. *Boxes and arrows* 2, 2004 (2004), 1–23.

[43] Jeff C Stanley and Stephen L Dorton. 2023. Exploring Trust With the AI Incident Database. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 67. SAGE Publications Sage CA: Los Angeles, CA, 489–494.

[44] C. Stokel-Walker. 2023. ChatGPT replicates gender bias in recommendation letters. https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/.

[45] Beki Subaeki, Aedah Abd Rahman, Khaerul Manaf, Riffa Haviani Laluma, Agung Wahana, and Nur Lukman. 2022. Assessing Tax Online System Success: A Validation of Success Model with Focus Group Study. In *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED).* IEEE, 1–6.

[46] Xianghui Tao, Lujia Li, Jianjun He, et al. 2022. Research on discrimination and regulation of artificial intelligence algorithm. In *2nd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2022)*, Vol. 12348. SPIE, 79–84.

[47] Violet Turri and Rachel Dzombak. 2023. Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 576–583.

[48] Antje Von Ungern-Sternberg. 2021. Discriminatory AI and the Law–Legal Standards for Algorithmic Profiling. *Draft Chapter, in: Silja Vöneky, Philipp Kellmeyer, Oliver Müller and Wolfram Burgard (ed.) Responsible AI, Cambridge University Press (Forthcoming)* (2021).

[49] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.

[50] Mengyi Wei and Zhixuan Zhou. 2022. Ai ethics issues in real world: Evidence from ai incident database. *arXiv preprint arXiv:2206.07635* (2022).

[51] George Williams. 2006. The Victorian charter of human rights and responsibilities: Origins and scope. *Melb. UL Rev.* 30 (2006), 880.

[52] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* 2021 (2021), 1–18.

[53] Jianlong Zhou, Fang Chen, Adam Berry, Mike Reed, Shujia Zhang, and Siobhan Savage. 2020. A survey on ethical principles of AI and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 3010–3017.

[54] Didar Zowghi and Muneera Bano. 2024. AI for all: Diversity and Inclusion in AI. https://doi.org/10.1007/s43681-024-00485-8. , 4 pages.

[55] Didar Zowghi and Francesca da Rimini. 2023. Diversity and Inclusion in Artificial Intelligence. *arXiv preprint arXiv:2305.12728* (2023).

[56] Frederik J Zuiderveen Borgesius. 2020. Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights* 24, 10 (2020), 1572–1593.

A Appendix A. Evolution of the Decision Tree

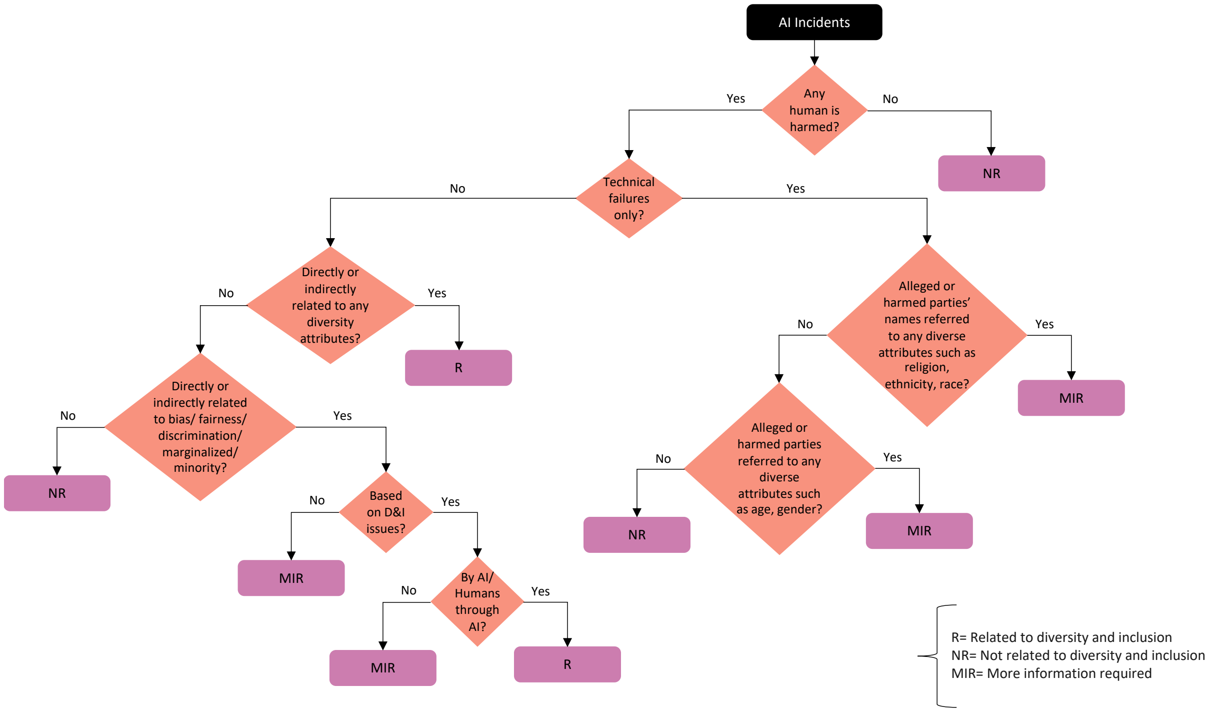

AI Incidents
Any human is harmed?
Yes
No
NR
Technical failures only?
No
Yes
Directly or indirectly related to any diversity attributes?
No
Yes
R
Alleged or harmed parties' names referred to any diverse attributes such as religion, ethnicity, race?
No
Yes
MIR
Directly or indirectly related to bias/ fairness/ discrimination/ marginalized/ minority?
No
Yes
NR
Based on D&I issues?
No
Yes
MIR
By AI/ Humans through AI?
No
Yes
MIR
R
Alleged or harmed parties referred to any diverse attributes such as age, gender?
No
Yes
NR
MIR
R= Related to diversity and inclusion
NR= Not related to diversity and inclusion
MIR= More information required

Fig. 7. Decision tree: Version 1

Received 14 June 2024; revised 2 April 2025; accepted 1 May 2025

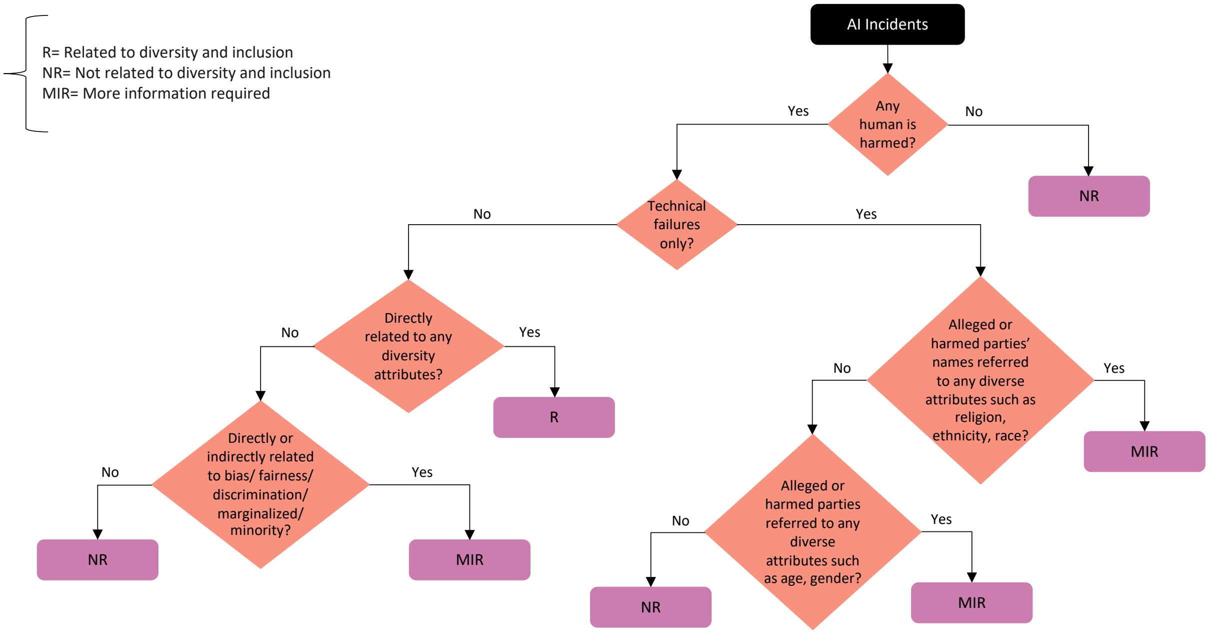

R= Related to diversity and inclusion
NR= Not related to diversity and inclusion
MIR= More information required
AI Incidents
Any human is harmed?
Yes
No
NR
Technical failures only?
No
Yes
Directly related to any diversity attributes?
No
Yes
R
Directly or indirectly related to bias/ fairness/ discrimination/ marginalized/ minority?
No
Yes
NR
MIR
Alleged or harmed parties' names referred to any diverse attributes such as religion, ethnicity, race?
No
Yes
MIR
Alleged or harmed parties referred to any diverse attributes such as age, gender?
No
Yes
NR
MIR

Fig. 8. Decision tree: Version 2

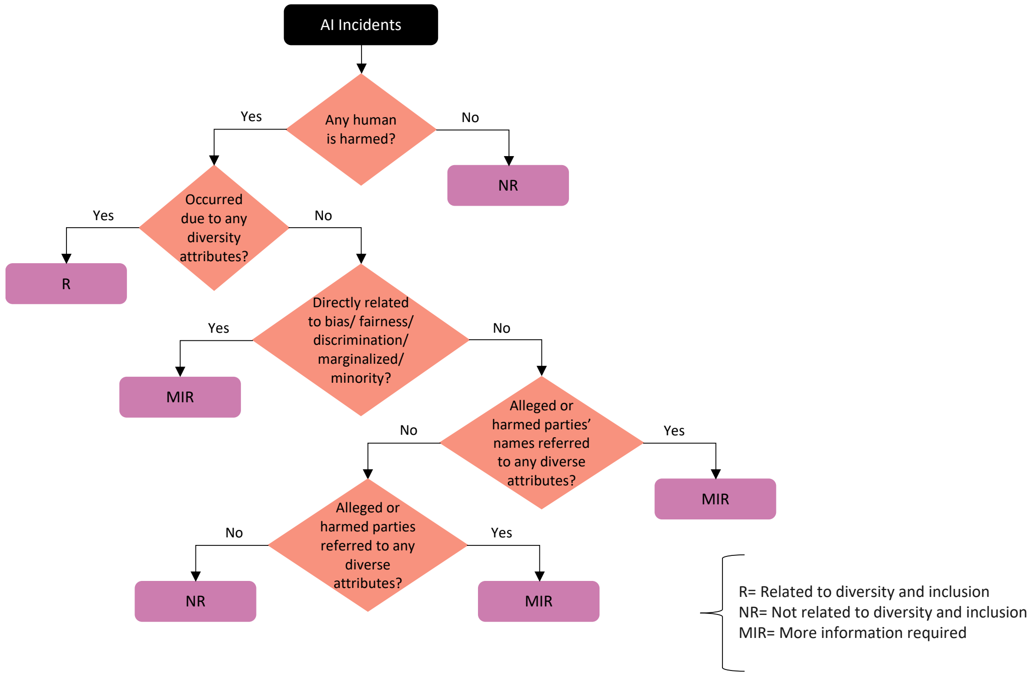

AI Incidents
Any human is harmed?
Yes
No
NR
Occurred due to any diversity attributes?
Yes
No
R
Directly related to bias/ fairness/ discrimination/ marginalized/ minority?
Yes
No
MIR
Alleged or harmed parties' names referred to any diverse attributes?
No
Yes
MIR
Alleged or harmed parties referred to any diverse attributes?
No
Yes
NR
MIR
R= Related to diversity and inclusion
NR= Not related to diversity and inclusion
MIR= More information required

Fig. 9. Decision tree: Version 3

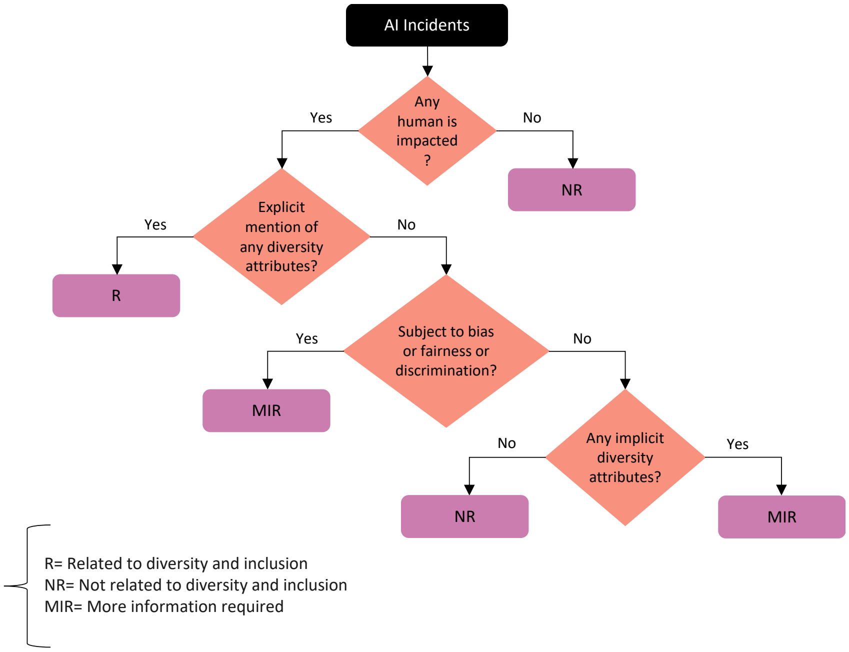

AI Incidents
Any human is impacted ?
Yes
No
NR
Explicit mention of any diversity attributes?
Yes
No
R
Subject to bias or fairness or discrimination?
Yes
No
MIR
Any implicit diversity attributes?
No
Yes
NR
MIR
R= Related to diversity and inclusion
NR= Not related to diversity and inclusion
MIR= More information required

Fig. 10. Decision tree: Version 4

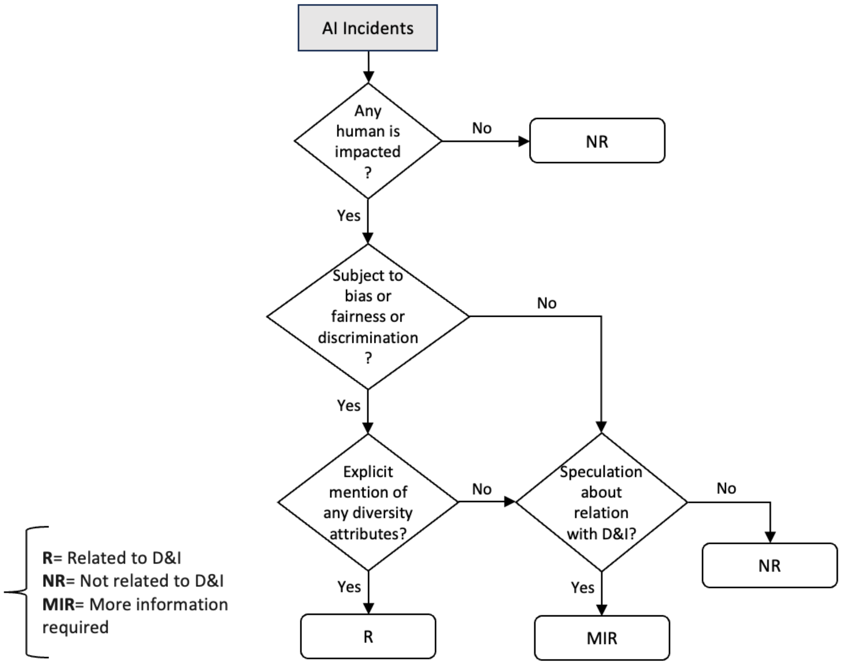

AI Incidents
Any human is impacted ?
No
NR
Yes
Subject to bias or fairness or discrimination ?
No
Yes
Explicit mention of any diversity attributes?
No
Speculation about relation with D&I?
No
NR
Yes
R
Yes
MIR
R= Related to D&I
NR= Not related to D&I
MIR= More information required

Fig. 11. Decision tree: Version 5