Why AI Harms Can't Be Fixed One Identity at a Time: What 5300 Incident Reports Reveal About Intersectionality

EDYTA PAULINA BOGUCKA, Nokia Bell Labs, United Kingdom
SANJA ŠĆEPANOVIĆ, Nokia Bell Labs, United Kingdom and University of Oxford, United Kingdom
DANIELE QUERCIA, Nokia Bell Labs, United Kingdom and Politecnico di Torino, Italy

AI risk assessment is the primary tool for identifying harms caused by AI systems. These include intersectional harms, which arise from the interaction between identity categories (e.g., class and skin tone) and which do not occur, or occur differently, when those categories are considered separately. Yet existing AI risk assessments are still built around isolated identity categories, and when intersections are considered, they focus almost exclusively on race and gender. Drawing on a large-scale analysis of documented AI incidents, we show that AI harms do not occur one identity category at a time. Using a structured rubric applied with a Large Language Model (LLM), we analyze 5,300 reports from 1,200 documented incidents in the AI Incident Database, the most curated source of incident data. From these reports, we identify 1,513 harmed subjects and their associated identity categories, achieving 98% accuracy. At the level of individual categories, we find that age and political identity appear in documented AI harms at rates comparable to race and gender. At the level of intersecting categories, harm is amplified up to three times at specific intersections: adolescent girls, lower-class people of color, and upper-class political elites. We argue that intersectionality should be a core component of AI risk assessment to more accurately capture how harms are produced and distributed across social groups.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **HCI design and evaluation methods**; • **Computing methodologies** → *Artificial intelligence*; • **Social and professional topics** → *Codes of ethics.*

Additional Key Words and Phrases: intersectionality, incidents, risk assessment, responsible AI, ethical AI, AI governance

ACM Reference Format:
Edyta Paulina Bogucka, Sanja Šćepanović, and Daniele Quercia. 2026. Why AI Harms Can't Be Fixed One Identity at a Time: What 5300 Incident Reports Reveal About Intersectionality. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3805689.3812347

1 Introduction

In 2018, the Argentinian province of Salta approved the deployment of a machine-learning system to predict adolescent pregnancy. Trained on health, education, and housing data, the system generated individual risk scores. Its deployment triggered public backlash after news reports showed that the resulting harms were not evenly distributed [9]. Girls were singled out for risk scoring and follow-up, while boys were excluded entirely, placing responsibility for pregnancy prevention on girls. Girls in rural and Indigenous communities were more likely to receive in-person follow-ups such as household visits than girls in urban areas. Pre-adolescent girls were labeled years in advance, even though they had not yet engaged in adolescent behavior.

Authors' Contact Information: Edyta Paulina Bogucka, Nokia Bell Labs, Cambridge, United Kingdom, edyta.bogucka@nokia-bell-labs.com; Sanja Šćepanović, Nokia Bell Labs, Cambridge, United Kingdom and University of Oxford, Oxford, United Kingdom, sanja.scepanovic@nokia-bell-labs.com; Daniele Quercia, Nokia Bell Labs, Cambridge, United Kingdom and Politecnico di Torino, Turin, Italy, quercia@cantab.net.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812347

This unequal distribution of harms is commonly explained using the theory of intersectionality. The concept was originally introduced by Crenshaw to describe how legal and policy frameworks failed to account for the combined effects of race and gender [48]. Subsequent scholarship emphasizes that intersectionality describes not only interactions among identity categories, but also the forms of social power that structure inequality such as racism, sexism, and classism [53, 55, 66]. The Salta case illustrates this logic in practice. The system caused harm not because of age, gender, ethnicity, or geography on their own, but because of how these factors played out within powerful institutions: public health bodies, welfare agencies, and regional governments. This dynamic is captured in a widely cited definition of intersectionality by Brah and Phoenix [42], who describe it as "the complex, irreducible, varied, and variable effects which ensue when multiple axes of differentiation — economic, political, cultural, psychic, subjective and experiential — intersect in historically specific contexts". Throughout this paper, we refer to these axes of differentiation as *identity categories* (such as race, gender, or age), and to their specific instantiations as *identity values* (such as Black, woman, or adolescent).

Despite broad recognition of intersectionality in theory, empirical research on AI-related harms has applied it narrowly. Prior work has largely focused on a small number of isolated identity categories [38], most commonly race and gender [40, 80, 81]. While this literature has established that intersectional harms exist [86], it offers limited insight into how they are distributed across many identity categories and AI uses. As a result, the empirical bases used in AI risk assessment such as fairness benchmarks [36, 57, 60] and incident taxonomies [49] may overlook identity configurations that are common in AI incidents. To make progress on this gap, we ask:

(RQ_1) Which identity categories and their corresponding values appear most often in AI incidents reported in the news, and what kinds of harm are linked to them?

(RQ_2) How do intersecting identity categories affect the likelihood and types of harm caused by AI systems?

To address these questions, we make three main contributions:

(1) **We develop a systematic approach for identifying and characterizing intersectional AI harms in incident reports (§3).** The approach specifies a structured rubric for identifying harmed subjects, relevant identity categories and values, and associated harm types. It can be applied either manually or with the support of Large Language Models (LLMs) to analyze heterogeneous and unstructured reports at scale.

(2) **We apply this approach to a large set of documented AI incidents to produce an empirical characterization of intersectional AI harms (§4).** We analyze 5,300 incident reports across 1,200 unique incidents from the AI Incident Database [75], identifying 1,513 harmed subjects with an accuracy of 98%. Across single identity categories implicated in these harms, age (32%) and political identity (27%) appear as frequently as race (25%) and gender (24%). However, we show that AI harms do not occur one identity category at a time. The most common intersectional harms involve nationality and political identity through deepfake targeting (12%), age and gender through sexualized profiling (10%), and nationality and class through denial of services (6%). At these intersections, people identified as upper-class political elites, adolescent girls, and lower-class people of color appear up to three times more often than would be expected if each identity category contributed to harm on its own.

(3) **We identify recurring patterns and the institutional contexts through which intersectional AI harms arise (§4).** Through thematic analysis, we find six such patterns (six specific ways AI systems act on identity attributes): sexualizing, steering, matching, inferring, gating, and manipulating (§4.1). These interact with institutions of social power to produce intersectional harms: algorithmic suspicion, sexualized exploitation, environmental exposure, political manipulation, and militarized state violence (§4.2).

Building on these contributions, we discuss how intersectional AI harms emerge across both high-visibility and everyday systems, consider the implications for AI risk assessment, and discuss the limitation of using news as sources of our incidents (§5). To support researchers in advancing this research direction, we have publicly released our approach at **https://social-dynamics.net/ai-risks/intersectional**.

2 Related Work

Next, we review prior work on AI-related harms through two complementary lenses: how harms manifest across intersecting identities (§2.1), and how they vary across AI use cases (§2.2).

2.1 Studying AI Harms Across Identity Intersections

Researchers in human-computer interaction (HCI) and AI ethics have increasingly drawn on intersectionality theory [48, 53, 55, 66] to examine how AI systems distribute harms unevenly across marginalized and privileged groups and reinforce forms of social power such as racism, sexism, and classism [80, 86]. In practice, however, this work has largely focused on intersections of only two identity categories [38], most often race and gender [40, 80, 81]. For example, Buolamwini and Gebru [45] showed that commercial facial analysis systems misclassified darker-skinned women at substantially higher rates than lighter-skinned men. Similarly, audits of CV screening systems found that LLMs favored CVs with white-sounding names 85% of the time, while never favoring Black male names over white male names [88]. These findings illustrate how harms at specific intersections such as allocative harms affecting Black men in hiring, remain invisible when race or gender is analyzed in isolation.

Across 526 papers published between 2018 and 2021 at FAccT and AIES, race and gender each appear in nearly 20% of papers, while categories such as disability, socioeconomic status, and religion appear in fewer than 3% [40]. Analyses of widely used fairness datasets further show that attributes related to disability, religion, health and belief are largely absent, and that socioeconomic variables, while more available, are rarely used empirically [82]. Recognizing this narrow focus, prior work has called for greater attention to underexamined intersections, such as class and race [80] or disability and age [86], though addressing them remains challenging in practice [89]. For example, the Political Deepfakes Incident Database documents harms at intersections of class and political identity using manually collected cases involving public figures. [87]. Such harms range from discrimination to targeted violence to financial loss.

Recent work has begun to overcome constraints on studying intersections involving more than two identity categories by leveraging LLMs. Phutane et al. [73] generated hiring conversations for school teacher and software engineer positions across six LLMs, constructing candidate profiles that explicitly combine disability, gender, and caste. They found that adding disability alone increased harmful conversation content by up to 58 times relative to baseline profiles, while further adding gender and caste increased harm by an additional 10 to 51%. They also showed that candidates at these combined intersections were exposed to new types of harm: being framed as inspirational because of their identity rather than their qualifications ("inspiration porn"), portrayed as possessing exaggerated abilities ("superhumanization"), or valued primarily for fulfilling workplace diversity goals rather than for their professional merit ("tokenism").

2.2 Studying AI Harms Across Use Cases

Much of the empirical literature on AI harms focuses on use cases that would be classified as high risk under the EU AI Act, that is, systems deployed in sensitive domains where failure or misuse can cause significant harm [70]. Examples include hiring [88], criminal justice [55], benefit allocation [43, 47, 74] and targeted advertising [79]. While this focus has yielded rich analyses, it has concentrated research attention on a narrow set of benchmark datasets and tasks. For example, much fairness research relies on COMPAS dataset for recidivism assessment [36], German Credit for credit risk assessment [57], and UCI Adult for income prediction [60]. This reliance on a few benchmarks limits coverage of the diversity of AI deployments and the harms they produce [40, 82].

A growing body of qualitative work has examined harms in seemingly low risk, everyday algorithmic systems, where users encounter AI quietly and repeatedly in routine interactions. Van Nuenen et al. [86] documented nearly 100 firsthand accounts of unfair treatment across everyday systems such as shopping platforms, with harms reported along axes of gender, sexual orientation, race and ethnicity, disability, and geographic location.

Across these identities, participants most frequently described being forced into inaccurate or adjacent categories by automated systems ("enforced categorization"), or being rendered unrecognizable altogether ("symbolic annihilation" [59]). Basoah et al. [37] examined AI-supported writing technologies and showed that even mundane tools such as grammar checkers and autocorrect systems can generate identity-related harms. They documented these harms along race and language: Black users reported that these tools failed to recognize African American Vernacular English, flagged culturally common expressions as errors, and framed their linguistic practices as unprofessional, leading to experiences of exclusion and cultural erasure.

In parallel with qualitative research, recent quantitative, data-driven work has examined AI harms across use cases at scale using large collections of incidents [41, 50, 81]. These studies draw on volunteer-submitted incident reports curated in public databases, reflecting the limited availability of structured top-down incident reporting mechanisms such as those envisioned under the EU AI Act [70]. For example, Shams et al. [81] analyzed incident reports from the AI Incident Database [75] and the AI, Algorithmic, and Automation Incidents and Controversies database [34] collected by mid-2023. These studies identified 18 diversity attributes implicated in AI incidents, including facial features, culture, and accent. However, across both databases, reported identity-related harms most commonly involved bias or discrimination along the same identity categories that recur across much of the literature: race and gender.

Research gap. Although intersectional AI harms are well established in theory, existing approaches to AI risk assessment focus on a narrow subset of isolated identity categories; when intersections are considered, they are almost exclusively limited to race and gender. We address this limitation by analyzing 5,300 AI incident reports to examine identity-related harms across a broader set of identity categories and less-explored intersections.

3 Methods for Analyzing Identity-Related AI Harms

To analyze identity-related AI harms, we followed a four-step approach (Figure 1, Steps 1–4). We first collected AI incident reports (§3.1), then identified harmed subjects and their identity categories within these reports (§3.2). Because an incident may involve an identity category without being caused by it, we assessed category relevance using counterfactual evaluation (§3.3), and finally compared relevant categories and their intersections across incidents (§3.4).

3.1 Collecting AI Incident Reports

Identity-related AI harms typically become visible not in controlled settings, but when deployed systems fail or harm populations. To identify harmed subjects and implicated identity categories, we rely on AI incident data, following De Miguel Velázquez et al. [50] and Shams et al. [81]. We define three criteria for selecting data sources: (C1) preservation of original materials such as serious incident reports [70], (C2) sufficient detail to support intersectional analysis, and (C3) collection and maintenance with human oversight. Appendix A explains how these criteria were applied.

We evaluated three data sources commonly used in studies of AI harms [41, 50, 81]: the OECD AI Incidents Monitor (AIM [68]; $K_{\text{incidents}} = 12,500$), the AI, Algorithmic, and Automation Incidents and Controversies database (AIAAIC [34]; $K_{\text{incidents}} = 2,100$), and the AI Incident Database (AIID [75]; $K_{\text{incidents}} = 1,200$). We selected the AIID as the only source satisfying all three criteria.

The AIID is a curated repository of AI incident reports submitted by community members, researchers, and industry practitioners. Submissions are collected through a public web form where contributors provide incident metadata and the text of underlying report(s). Reports must originate from public online sources such as news media, court records, or company disclosures, though, in practice, most derive from news coverage. After submission, AIID editors review the materials and determine whether to include the incident in the database.

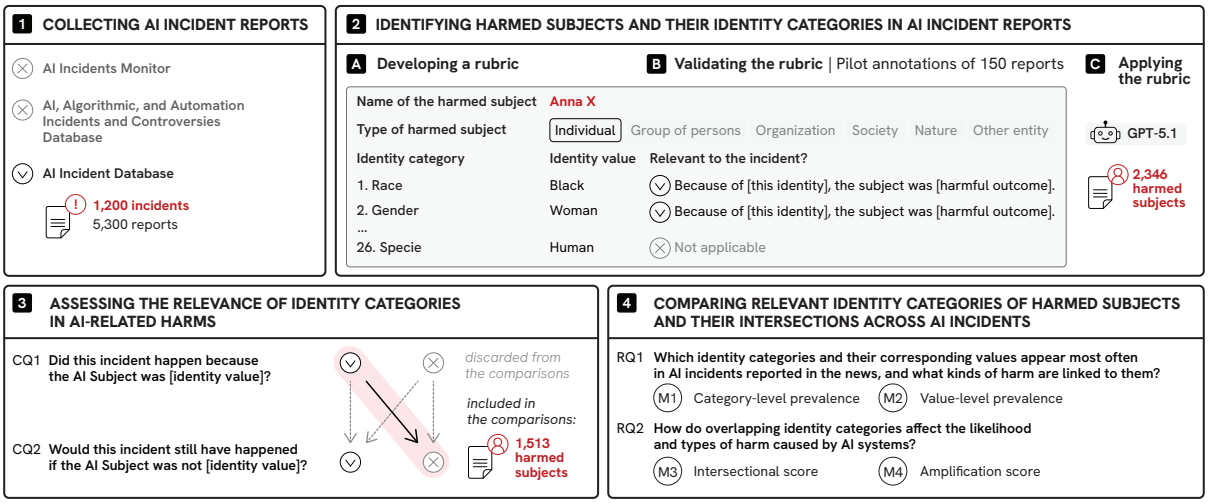

Fig. 1. **Overview of our four-step methodology for identifying and analyzing intersectional AI harms in AI incidents.** The approach combines large-scale incident collection (1), a rubric for identifying the identities of harmed subjects (2A–B), LLM-assisted extraction of these identities using the rubric (2C), and counterfactual relevance assessment (3). Together, these steps enable the systematic identification of harmed subjects, their identity categories and intersections, and the concrete harms attributable to identity-specific factors across the diverse AI uses covered in the incident reports (4).

This reporting pipeline introduces four main biases: (1) concentration in media-visible domains (e.g., communication, transportation) [77]; (2) reliance on liberal or center-left news sources [50]; (3) amplification of incidents involving public figures or politically identifiable groups through repeated media coverage [87]; and (4) a focus on incidents occurring in the United States [50]. We discuss these biases when interpreting our results in §5.4.

We then obtained a complete dataset from the AIID as of September 1, 2025. The dataset included incident identifiers, incident titles, descriptions, the number of associated reports, and the full text of each report, spanning 1,200 unique incidents and 5,300 reports. Incidents were documented by 1–58 reports ($\mu = 4$), with 61% associated with more than 1 report. This structure provided sufficient detail to identify harmed subjects and extract identity categories while enabling comparison across incidents documented by multiple reports.

3.2 Identifying Harmed Subjects and Their Identity Categories in AI Incident Reports

We proceeded to identify harmed subjects and their identity categories in each report. To do so, we developed a rubric, applied it at scale to the full dataset using an LLM, and validated the results.

3.2.1 Developing a Rubric. We developed the rubric by reviewing foundational and subsequent research on intersectionality [38, 48, 53, 55, 71, 84], representation of users in HCI [62, 73, 80], and existing regulatory classifications relevant to AI harms [54, 70]. The rubric consists of five components.

First, it identifies harmed subjects using their exact name or descriptor as it appears in the report text [75].

Second, the rubric defines a typology of harmed subjects (i.e., individuals, groups, organizations, societies, and nature), adapted from the EU AI Act [70], a leading framework for AI governance.

Third, the rubric specifies a fixed set of 26 identity categories for each harmed subject and requires assigning a concrete value within a category when applicable. For example, the caste category can include values such as "Brahmin" and "Dalit" [73], and the education category can include values such as "student" or "vocational trainee". The category set was developed using an intercategorical approach [62], treating identity categories as

provisional analytical labels rather than fixed or exhaustive descriptions [80]. Categories were sourced through iterative review of seminal intersectionality and HCI literature [38, 46, 48, 53, 55, 71, 84]. Where categories could take a wide range of values, the rubric specifies explicit grouping rules to ensure consistency in value assignment. For class, for example, the rubric adopts the definition from Ames et al. [35] as "a nexus of income level, educational attainment, and type of employment". Since education is captured by its own category, the rubric requires grouping only income and occupation values into lower, middle, and upper class (e.g., "gig worker" to lower, "small business owner" to middle, and "politician" to upper class). The full list of categories, their exemplary values, grouping rules, and methodological process are provided in Appendix B.

Fourth, the rubric outlines how to assess whether an identity category is relevant to an incident using two counterfactual questions: (CQ1) whether the incident occurred because the subject had this identity category, and (CQ2) whether the incident would likely have occurred if the subject were identical in all respects except this identity category.

Fifth, the rubric requires a description of the harm experienced by the subject due to their identity category.

Prior to its application at scale, we improved the rubric through pilot annotations. Two members of the research team independently applied preliminary versions of the rubric to a pilot set of 50 incident reports sampled to span different AI subject types and harm contexts. These annotations were compared and discussed in joint review sessions. Disagreements were mostly about relevance judgments for borderline cases where identity information was ambiguously related to harm. These were used to refine category definitions and improve the phrasing of the counterfactual relevance questions. This process was repeated iteratively until the rubric yielded stable annotations across pilot reports.

3.2.2 Applying the Rubric. We then applied the rubric at scale to the incident reports by translating it into a structured prompt that can be executed by LLMs (provided in Appendix C). To analyze our dataset, we used OpenAI's GPT-5.1 via batch API processing. At the time of analysis, this model showed the strongest performance on capabilities relevant to our rubric, including instruction following (SWE-bench Verified benchmark), extraction of information from long documents (BrowseComp Long Context benchmark), and multi-step reasoning (GPQA Diamond benchmark).

For each incident report, the model applied the specified prompt. The model produced structured extractions aligned with the rubric, including harmed subjects, subject types, identity categories and values, and harm descriptions (see Figure 2 for an exemplary incident report and the corresponding extractions). The resulting dataset comprised 5,300 reports covering 1,200 unique incidents and 2,346 harmed subjects.

3.2.3 Validating the Rubric Results. To assess extraction quality, we conducted a double-annotation exercise on a sample of 50 incident reports. Two annotators independently identified harmed subjects, assigned identity categories and values, and evaluated relevance using the rubric's counterfactual criteria. Agreement was high for harmed subject identification (92% agreement, PABAK = 0.84) and causal relevance assessment (88% agreement, PABAK = 0.76).[1] Identity category and value assignment showed substantial agreement (Cohen's $\kappa = 0.63$), with annotators matching on 82% of judgments. Most disagreements between the research team and the LLM arose in three situations.

First, when identity attributes were implied but not always explicitly stated in incident reports. For example, the LLM failed to extract that Céline Dion is a woman [33]. Second, when the LLM mistook quoted claims about a subject for factual identity attributes. For example, it incorrectly inferred a political identity of "communist dictator" for Kamala Harris from an AI-generated disinformation post such as quoted in the incident report [29]. Third, disagreements occurred when distinguishing between directly harmed subjects and affected bystanders.

[1]For these highly skewed distributions where both annotators predominantly selected "Yes", we report the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) rather than Cohen's κ, which underestimates agreement under class imbalance.

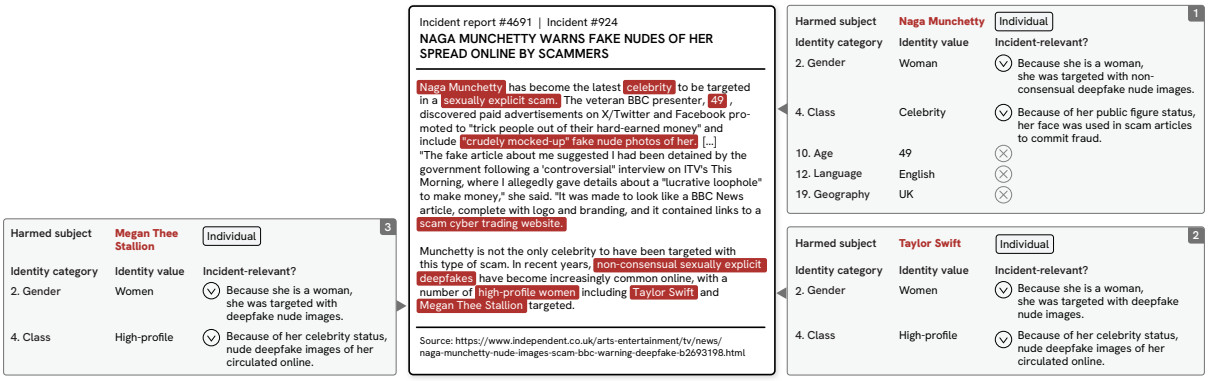

Fig. 2. **Example of an incident report illustrating harmed subjects, their identity categories and values, and associated harm descriptions extracted by an LLM using our rubric.** The example of report 4691 highlights three harmed subjects whose identities combine female gender with upper-class roles such as media personalities and celebrities.

For example, in the Tay chatbot incident [6], different members of the research team identified different harmed subjects, such as Black people or women, while the LLM labeled a broader group, such as U.S. social media users. After resolving disagreements through discussion between annotators, we computed LLM accuracy against the human gold standard. The LLM achieved 98% accuracy on subject identification, 97% accuracy on identity category value assignments (across 24 categories per subject), and 92% accuracy on causal relevance judgments. To explore these disagreements, we examined the distribution of LLM misattributions across identity categories (i.e., cases where the LLM's identity category assignments differed from those of the annotators; Appendix D, Figure 7). Misattributions were low across most categories (0–6%), with the highest rate for gender (28%). This rate reflects the first situation described above (i.e., the presence of implicit gender cues such as gender-coded titles or names that are inferred by human annotators but not by LLMs).

3.3 Assessing the Relevance of Identity Categories in AI-Related Harms

Not all identity categories extracted from incident reports might be causally relevant to the harms described. Journalistic accounts often include personal details for context, and guidance on reporting on AI advises against reading such details as explanations for what went wrong, in order to avoid amplifying harm through misattribution [1]. We therefore assessed the relevance of each extracted identity category using the counterfactual evaluation defined in the rubric.

Specifically, for each identity category associated with a harmed subject, we analyzed the rubric's counterfactual component, which recorded two binary judgments: (CQ1) whether the incident occurred because the subject had this identity category, and (CQ2) whether the same incident would likely have occurred if the subject were identical in all respects except for this identity category. An identity category was retained only if it materially contributed to the AI system's behavior (CQ1 = "Yes") and if changing the category would plausibly have altered the outcome (CQ2 = "No"). Identity categories that did not meet this condition were removed from the subject record. If a harmed subject had no remaining relevant identity categories after this step, the subject was removed. If an incident had no remaining harmed subjects with relevant identity categories, the incident was removed.

To illustrate this step, consider an incident report describing the biased pregnancy prediction system deployed in Argentina, as introduced at the opening of this paper [9]. The system generated individual risk scores from administrative data to guide follow-up interventions. Two identity categories are associated with the harmed subjects: gender (girls) and nationality (Argentine). For *gender*, the LLM determined that the incident plausibly occurred because the subjects were girls (CQ1 = "Yes"), and that the outcome would likely have differed had the

subjects been boys (CQ2 = "No"), since girls were singled out for risk scoring while boys were excluded entirely. The gender category was therefore retained. For *nationality*, the LLM determined that the incident did not occur because the subjects were Argentine (CQ1 = "No"), and that the same outcome would likely have occurred had the girls held a different nationality (CQ2 = "Yes"). Because nationality did not plausibly influence the system's behavior, this category was removed from the subject record.

After applying this relevance filtering, the resulting dataset comprised 711 unique incidents and 1,513 harmed subjects. This filtered dataset forms the basis for the analyses reported in the following section.

3.4 Comparing Relevant Identity Categories of Harmed Subjects and Their Intersections Across AI Incidents

Using the filtered dataset, we address our two research questions by defining appropriate metrics and applying complementary quantitative and qualitative methods. We first analyze identity categories and values individually to assess their prevalence and associated harms (RQ_1), and then examine their intersections (RQ_2).

RQ_1: Which identity categories and their corresponding values appear most often in AI incidents reported in the news, and what kinds of harm are linked to them?
We answered this question both quantitatively and qualitatively. Quantitatively, we defined 2 metrics:

(1) *Category-level prevalence.* To determine which identity categories are most represented in AI harm incidents, we calculated the prevalence of each category as:

$$\text{prevalence}_{\text{c}_i} = \frac{n_{\text{c}_i}}{N} \tag{1}$$

where n_{c_i} is the number of incidents involving at least one subject of identity category c_i (e.g., gender, race), and N is the total number of incidents in our dataset.

(2) *Value-level prevalence.* To examine which specific identity values within each category are most frequently associated with harm, we calculated:

$$\text{prevalence}_{\text{v}_i} = \frac{n_{\text{v}_i}}{N} \tag{2}$$

where n_{v_i} is the number of incidents involving at least one subject with identity value v_i (e.g., female for gender, Black for race).

Qualitatively, we thematically analyzed harm examples associated with specific identities. We followed a four-step iterative coding process [44, 63, 64, 78]. First, two coders from the research team independently coded a random sample of 50 incident reports by writing down the AI behavior, the identity categories and their specific values it targeted, and the resulting harm (e.g., behavior: AI generated sexualized images; identities: gender — women, age — girls; harm: sexual objectification). We discussed these codes among all authors, and refined them into a shared codebook. Second, the same two coders applied this codebook independently to the full dataset, and resolved any disagreements through discussion. Third, we reviewed the codes and clustered AI actions describing similar harms into a single pattern (e.g., sexualizing women and girls; mismatching people of color). Fourth, we refined these clusters into six final patterns, which we report with representative incidents in §4.1.

RQ_2: How do intersecting identity categories affect the likelihood and types of harm caused by AI systems?
We answered this question both quantitatively and qualitatively. Quantitatively, we defined 2 metrics:

(1) *Intersectional score.* To identify which combinations of identity categories co-occur most frequently in AI harm incidents, we calculated the joint prevalence of each category pair as:

$$\text{intersectional\_score}_{\text{c}_1,\text{c}_2} = \frac{n_{\text{c}_1,\text{c}_2}}{N} \tag{3}$$

where n_{c_1,c_2} is the number of incidents involving both identity categories (e.g., gender and age) and N is the total number of incidents. We visualize co-occurrences in a matrix to reveal intersectional harm patterns.

(2) *Amplification score.* To assess whether for particular intersecting categories, specific identity value combinations are disproportionately represented among AI-harmed subjects beyond what would be expected by chance, we computed a conditional amplification score:

$$\text{amplification\_score}_{v_1,v_2} = \frac{n_{v_1,v_2}}{\mathbb{E}[n_{v_1,v_2}]} \tag{4}$$

Here, n_{v_1,v_2} is the number of incidents involving both identity category values v_1 and v_2 (e.g., a Black woman), and $\mathbb{E}[n_{v_1,v_2}]$ is the expected number of incidents under the assumption that v_1 and v_2 are independent:

$$\mathbb{E}[n_{v_1,v_2}] = \frac{n_{v_1} \times n_{v_2}}{N} \tag{5}$$

A score greater than 1 indicates that the value combination occurs more frequently than expected by chance, reflecting amplified vulnerability to AI harm; a score below 1 indicates lower-than-expected occurrence.

Qualitatively, we thematically analyzed harms associated with intersecting identity values. First, we selected harms involving multiple identity categories and values (e.g., harm: sexual objectification; identities: gender — women *and* age — girls). Second, we coded and grouped these harms into patterns at each intersection. Third, we clustered patterns across intersections into five intersectional harm themes, which we report with representative incidents in §4.2.

4 Results

We report results in two steps, first examining identity categories and values in isolation (§4.1), and then analyzing how intersecting identities shape the likelihood and forms of AI-related harm (§4.2).

4.1 *RQ$_1$:* Which identity categories and their corresponding values appear most often AI incidents reported in the news, and what kinds of harm are linked to them?

Identity-related AI harms span all identity categories, with age and political identity appearing at rates comparable to race and gender. We found 1,513 harmed subjects across 711 unique incidents. Among these, age emerges as the most prevalent identity category, appearing in 32% of incidents, followed by political identity (27%), class (25%), race (25%), nationality (25%), and gender (24%) (Figure 3).

Within identity categories, AI harms concentrate on those who are structurally exposed rather than numerically dominant. In age-based incidents, adolescents (14%) and children (8%) appear more frequently than adults (5%), despite comprising a smaller share of the populations affected by many AI systems (Figure 4). Political identity illustrates this pattern clearly: harms concentrate on structurally exposed roles such as political elites (13%), voters (7%), activists (4%), and party candidates (4%), rather than on larger ideological groups like left-wing individuals (5%). Class-based harms disproportionately affect lower-class subjects (12%). Race-related harms primarily involve people of color (22%), while incidents involving white subjects are far less common (5%). Gender-related harms overwhelmingly affect females (21%), compared to males (4%).

Single identity category AI harms recur through a small set of mechanisms by which systems act on one socially salient value and scale its consequences. The thematic analysis shows that these harms follow six recurring patterns in how AI systems act on a single identity value, namely by *sexualizing*, *steering*, *matching*, *inferring*, *gating*, or *manipulating* individuals. *Sexualizing* is most visible in harms affecting women and girls, where generative systems are used to produce and circulate non-consensual sexual imagery such as deepfakes. This turns femininity into a persistent site of reputational and psychological attack (in incidents [18, 21, 26]), with especially acute effects when such material circulates in school or peer contexts (see incident [24]).

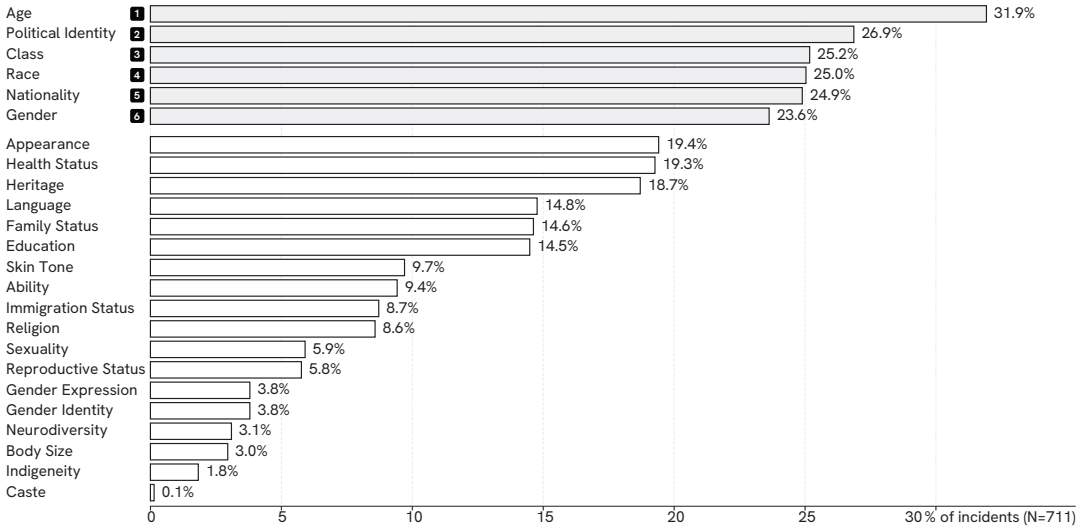

Fig. 3. **Percentage of incidents in which the identity category was causally relevant to the incident (prevalence of the identity category as per Equation 1).** The figure shows the percentage of incidents ($N = 711$) in which each identity category was identified as causally relevant to harm, counted once per incident; the six most prevalent categories (1–6) each appear in over 20% of incidents. Age and political identity appear as frequently as race and gender, followed by class and nationality. Least frequently documented categories include body size, indigeneity, and caste.

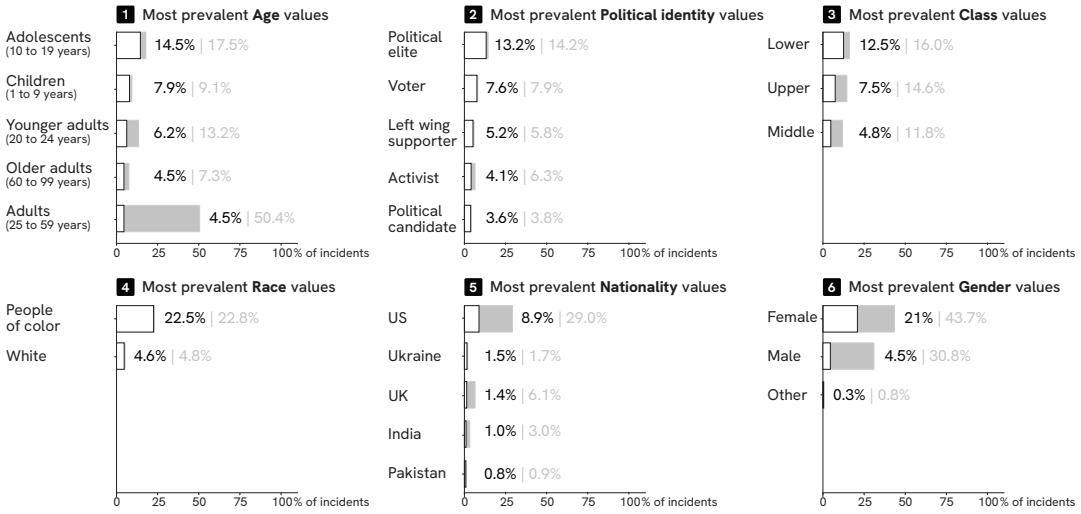

Fig. 4. **Percentage of incidents for the most prevalent values within each of the six most prevalent categories (Figure 3).** Each panel shows the percentage of incidents in which a given identity value appears (prevalence of the identity value as per Equation 2. Grey rectangles show overall prevalence across incidents, and white rectangles show prevalence when the identity value was causally relevant to the harm counted once per incident). Adolescents, political elites, individuals with lower class, people of color, U.S. nationals, and females appear most frequently within their respective categories.

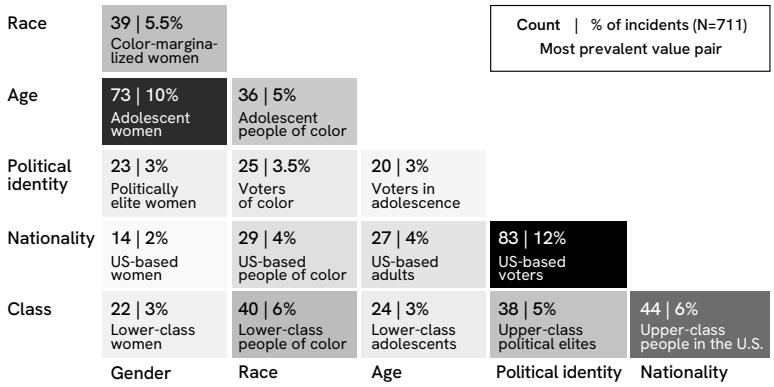

Fig. 5. **Prevalence of intersecting identity categories in AI incidents with identity-related harm.** Each cell shows the number and share of incidents ($N = 711$) involving subjects with both categories, counted once per incident. Darker shading indicates higher prevalence. The most frequent intersections involve nationality and political identity, age and gender, and nationality and class.

Steering operates most clearly in harms affecting men and boys, where recommender systems systematically funnel them toward misogynistic or extremist content, shaping norms around dominance, hostility, and political violence rather than simply reflecting pre-existing preferences (see incidents [4, 17]). *Matching* becomes harmful where biometric identification systems misrecognize racialized individuals: facial recognition technologies disproportionately misidentify Black people and convert technical error into wrongful stops or arrests (see incidents [11, 13, 15, 20]). *Inferring* produces distinct harms when AI systems attempt to predict or reveal sexual orientation, exposing individuals to risks of outing, targeting, and physical danger without consent (see incidents [8, 19]). *Gating* emerges where access to work or services is conditioned on rigid identity verification, as in biometric systems that assume binary gender presentation and treat change as anomaly, leading to automated exclusion and income loss for gender-diverse people (see incident [10]). Finally, *manipulating* captures harms tied to political identity, where generative media and personalization systems fabricate imagery or narratives attributed to specific political groups, shaping beliefs by exploiting identity-based trust (see incidents [22, 25]).

4.2 RQ_2: How do intersecting identity categories affect the likelihood and types of harm caused by AI systems?

Intersectional harms are not evenly distributed across category combinations. Instead, a few pairings recur frequently across AI incidents (Figure 5). The most common intersection is nationality and political identity (12%), followed by age and gender (10%) and nationality and class (6%).

Female gender, adolescence, lower class status, and political elite roles stand out for sharply amplifying harm when they intersect, often appearing at least twice as often as expected. Intersectional effects are not merely additive. Several identity value combinations appear in incidents far more often than would be expected based on their individual prevalence alone (Figure 6). The intersection of female gender with adolescent age appears at more than twice the rate expected under independence of the two categories. A contrasting gender and age pattern emerges for male subjects. Male gender intersects not with adolescent age but with adulthood, appearing nearly three times more often than expected. This shows that amplification also arises when socially dominant categories align. Socioeconomic and racial intersections also show strong amplification. Lower class status intersecting with being a person of color appears at nearly twice the expected rate, illustrating how marginalized identities compound one another. Upper class status intersecting with political elite membership

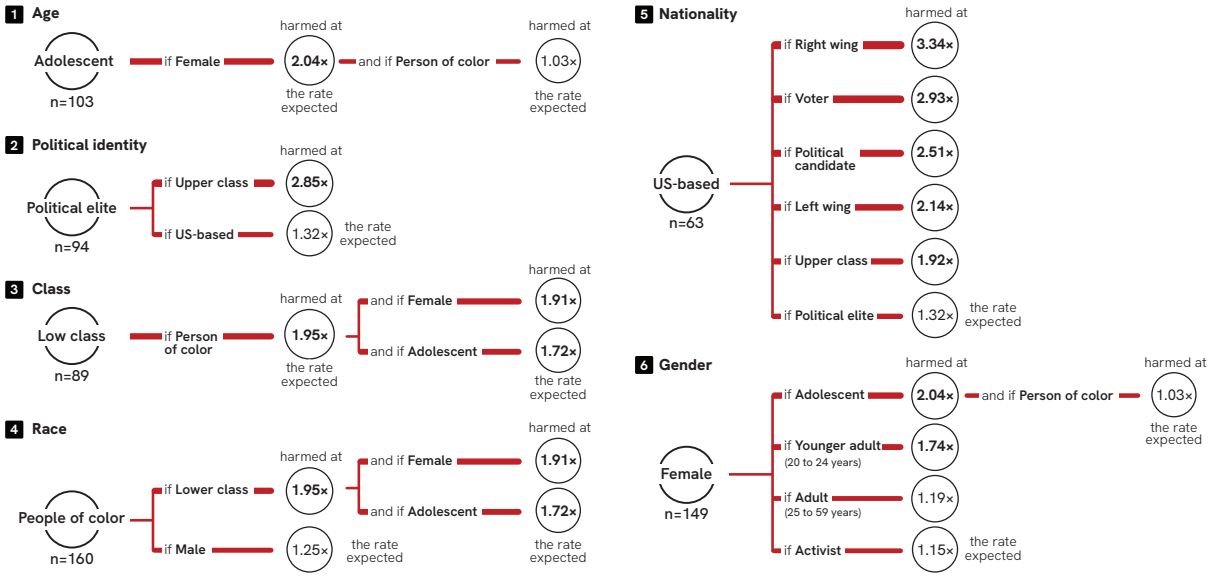

Fig. 6. **The amplification scores for the six most prevalent categories in Figure 3.** Line width and labels inside the circles with an × amplification factor report the amplification score between a pair of two identity values, calculated as per Formula 4. Scores greater than 1 indicate that a given intersection occurs at ×-times the rate expected under independence of the two categories. The strongest amplification is observed for US and `right-wing` voters (3.34×–2.93×): from point 5, the observed number of harmed incidents involving subjects who are US and `right-wing`, $n_{\text{US, right-wing}}$, is 3.34× the expected number under the assumption that US and `right-wing` are independent. This value pair is followed by intersections of political elites with upper-class status (2.85×), and adolescent age with female gender (2.04×).

appears nearly three times more often than expected, similarly amplifying already dominant positions. By contrast, several identity values that are common in isolation show little or no amplification when intersected. For example, male gender intersecting with race yields amplification score close to one, indicating no elevated harm. These patterns show that intersectional amplification emerges at both ends of the power spectrum, compounding harm for marginalized groups like women and adolescents, and concentrating risk even among dominant, highly visible groups such as political elites.

Intersectional AI harms arise through processes that link AI systems to social power. We identified five dominant intersectional harm themes (Table 1, Appendix E).

First, intersectional harms cluster around *algorithmic suspicion and criminalization*. These harms are driven by risk-scoring, fraud-detection, and classification systems in welfare administration, immigration control, and criminal justice. They disproportionately affect low-income, racialized, and migrant families. For example, in the Dutch childcare benefits system, dual-nationality families were flagged as fraudulent, leading to debt and family separation (see IDs 40, 101, and 335 in Table 1). In this theme, AI systems operationalize historically racialized and classed logics of suspicion, reflecting racism, classism, and nativism.

Second, a recurring pattern of *sexualized exploitation* arises from generative media systems and platform-scale recommender infrastructures. They disproportionately affect adolescent girls and young women, particularly those from lower-class, Indigenous, or migrant communities. Harms include non-consensual deepfakes, "nudification", grooming, and invasive prediction of reproductive futures. An opening example of this paper concerns a predictive analytics system in Argentina that classified low-income Indigenous girls as at risk" of adolescent pregnancy [9]. These girls were often subjected to intensive data collection and surveillance linked to access to social services

(see IDs 188, 904, and 924 in Table 1). Here, AI systems amplify sexism and ageism, reinforcing patriarchal control over women's bodies.

Third, intersectional harms emerge through *environmental violence*. These harms are produced by AI-enabled technical infrastructure and allocation systems in energy, urban planning, and healthcare. They disproportionately burden racialized and lower class communities with children. A clear example is the siting of xAI's supercomputer facility in South Memphis, which concentrated pollution in historically Black neighborhoods, increasing health risks for children and families without local consent or governance input (see ID 1144 in Table 1). In this theme, AI harms reflect environmental racism and classism.

Fourth, we identify *political manipulation and democratic erosion* enabled by AI-generated media, synthetic personas, and personalization systems. These systems operate in electoral and political communication contexts. They target young voters and racialized communities with tailored disinformation. AI-generated images falsely depicting Black political support illustrate this dynamic (see IDs 202, 650, and 972 in Table 1). These cases exploit forms of social power such as ageism, racism and nationalism to shape participation and public perception.

A final theme involves *militarized and state violence* enabled by AI. These harms are driven by surveillance, targeting, and threat-classification systems in policing, security, and warfare. They disproportionately affect civilians at the intersections of nationality, race, and political identity. For example, the Lavender AI system used during Israel's war in Gaza (see ID 672 in Table 1) classified Palestinian men as suspected militants and identified their family homes as targets for airstrikes, often with minimal human review, resulting in large-scale civilian deaths. These cases reflect extreme forms of nationalism, racism, and militarism where automated classification directly mediates life-and-death outcomes.

5 Discussion

This study examines identity-related AI harms at a scale and breadth rarely achieved in prior work. Combining large-scale AI incident analysis with an intersectional rubric and thematic analysis, we surface both established and underexplored harm patterns. We interpret these findings in relation to prior literature (§5.1), discuss contributions to intersectionality-informed AI research (§5.2), and outline implications, limitations, and future directions (§5.3, §5.4).

5.1 In-line with Previous Literature

Our findings align with prior research in two key areas: (1) persistent disparities along race and gender lines in AI harm, and (2) the disproportionate targeting of structurally vulnerable groups.

First, we replicate well-documented racial and gender disparities in AI harms. Prior work showed that algorithmic systems disproportionately misrepresent or disadvantage racialized and feminine-coded individuals [45, 55]. Our findings reinforce this pattern, with AI-related harm incidents disproportionately involving people of color and women.

Second, we confirm that AI harms concentrate on specific social groups rather than on numerically dominant populations. Low-income families experience disproportionate harm from automated decision-making systems that govern access to welfare, where risk scores can worsen financial instability [51]. Racialized and minority groups are disproportionately affected by surveillance and classification systems that intensify unequal monitoring [39]. Children are frequently harmed by AI systems in digital environments despite having limited capacity to opt out from automated decisions. These patterns support critiques that fairness frameworks often overlook how harm is shaped by power, governance, and unequal exposure to institutional systems [56]. Our results reinforce this critique by showing elevated harm among children and low-income subjects, indicating that AI systems amplify existing social vulnerabilities.

5.2 Contributions to Previous Literature

Our findings challenge prevailing assumptions in responsible AI research in two key ways: (1) by broadening the landscape of identity categories implicated in AI risk assessments beyond the field's dominant focus on race and gender, and (2) by showing that intersectional harms arise not only from the compounding of marginalizations but also through concentrated exposure at the high-end spectrum of power.

First, we find that AI harms related to age and political identity occur as frequently as those based on race and gender. This challenges the dominant framing in algorithmic fairness research, which has historically focused on race and gender as the primary axes of harm [45, 67]. Age has been largely neglected as a core dimension of algorithmic discrimination. Researchers have recently called attention to AI ageism [83], where older adults face exclusion, misclassification, or stereotyping in AI systems across domains like employment and health care [69]. Age-based vulnerability also extends to younger populations, as AI systems embedded in children's toys, tutors, and media ecosystems deliberately personalise learning, play, and social interaction, raising age-specific risks of reduced exposure to diverse experiences, dependency on compliant AI companions, and early immersion in highly personalised algorithmic environments [85]. Political identity has emerged as a significant axis of vulnerability. Peters [72] argue that political bias in algorithmic systems can operate like racial or gender bias but with fewer social or institutional safeguards. Our findings support this concern, showing that activists, candidates, and voters are regularly harmed by AI-mediated impersonation.

Second, we find that intersectional harms are not confined to the most marginalized groups. Previous work on intersectionality in AI has appropriately focused on how intersecting systems of oppression compound risk for subjects at the margins such as women of color or low-income migrants [61]. Our findings do not contradict this focus but extend it by showing that intersectional harms can also arise through concentrated exposure at positions of institutional visibility and power. Political elites are one such example. In our dataset, they frequently appear in incidents where AI systems are used to manipulate public narratives that undermine trust and participation [76]. To see how this occurs, consider how Cara Hunter was targeted by a deepfake pornographic video during an election campaign, an incident widely reported to have nearly ended her political career [58]. Similarly, older affluent individuals can face exclusion through age-based algorithmic profiling in domains such as credit scoring and hiring, where automated systems treat age itself as a risk or liability, overriding otherwise advantaged class positions and professional credentials [83].

5.3 Implications

Our findings show that addressing AI harms requires moving beyond risk occurring for specific identities in isolation toward identifying and mitigating risks occurring at specific identity intersections. In practice, this means that intersectionality can be used to: (1) identify which groups are structurally exposed during system design; (2) evaluate systems against intersection-specific harm patterns at deployment; and (3) monitor how these harms emerge and accumulate over time. We elaborate these implications across three phases of the AI lifecycle: design, deployment, and monitoring.

Design Phase. The rubric introduced in this work (Appendix B, C) can support early risk anticipation by moving beyond single-axis notions of intended users. Design teams can use it internally to stress-test assumptions about "typical" users and to identify which identity intersections are likely to be structurally exposed within a given domain. For example, in public-sector systems, categories such as language or migration status may be more consequential than the race–gender combinations emphasized in many Western fairness frameworks. At the same time, our findings caution against uncritical use of identity-based personas during design. When applied superficially, intersectionality can reinforce stereotypes rather than reveal how institutional systems unevenly affect different groups.

Deployment Phase. Our analysis informs guidance for audits, procurement, and regulatory oversight. For example, the concentration of non-consensual image generation involving adolescent girls calls for heightened scrutiny of AI tools used in schools and messaging apps. Similarly, harms involving adult men being algorithmically steered toward misogynistic or extremist content indicate that recommender systems should be assessed for how they shape behavior across specific gender–age cohorts, rather than solely through aggregate engagement metrics. Procurement teams and regulators can use these empirically observed harm patterns to request documentation on how vendors identify and mitigate intersection-specific risks. Such evaluations should consider not only incident frequency but also harm severity and downstream impact. As we show in our work, some intersections appear less often yet carry disproportionate consequences such as long-term reputational damage.

Monitoring Phase. Our findings underscore the value of participatory end-user auditing for detecting intersectional harms that emerge or intensify post-deployment. Many harms become visible only through sustained use and lived experience, particularly for structurally exposed groups. The rubric can support participatory audits by providing a shared structure for documenting identity-relevant harms and their concrete impacts, without requiring technical expertise. Integrating such audits into ongoing monitoring helps surface rare but severe, cumulative, or normalized harms that may remain invisible to developer-centric metrics such as fairness benchmark performance.

5.4 Limitations and Future Work

Our work comes with four main limitations that point to directions for future research.

First, it relies on publicly documented AI incidents shaped by news media visibility and disclosure practices. This introduces reporting bias: harms affecting less visible populations or occurring in closed systems are likely underrepresented [50]. Future work could incorporate regulatory filings or whistleblower reports.

Second, incident reports primarily describe individuals or groups directly affected by AI uses and rarely document ripple effects on other populations. Our analysis therefore focuses on explicitly named subjects and cannot capture indirect harms across social, organizational, or institutional contexts. Future work could extend incident reporting frameworks to capture secondary harms and affected stakeholders [2], or use strategic foresight to anticipate them [52].

Third, the incident reports are skewed toward U.S.-based subjects, limiting generalizability to regions where AI harms may be underreported, differently framed, or embedded in distinct sociopolitical contexts. As a result, the prevalence and amplification patterns reflect locally documented harms rather than the global distribution of AI-related harm. Future work should prioritize incident collection in non-U.S. and non-English-speaking contexts and examine how intersectional harm manifests across different regulatory, cultural, and infrastructural settings.

Fourth, our use of LLMs enables scalable analysis but reflects the interpretive frame of the rubric and model responses. Although we applied counterfactual filtering and validation to improve reliability, future work could incorporate complementary approaches such as expert or community-based annotation, to surface alternative interpretations.

6 Conclusion

This paper delivers an empirical analysis to date of intersectional AI harms by examining 5,300 reports across 1,200 incidents. It identifies recurring intersectional harm patterns that show how AI systems amplify both structural vulnerability and institutional visibility across diverse domains. These contributions broaden the empirical foundations of responsible AI research and demonstrate why AI risk assessment must move beyond narrow fairness categories to address the identity configurations most commonly harmed in practice.

7 Endmatter Statements

7.1 Generative AI Usage Statement

We used generative AI tools for three tasks: data analysis, manuscript development, and accessibility support, as described below.

Use in Data Analysis. We used a state-of-the-art, commercially available LLM, OpenAI GPT-5.1, via the API, to systematically extract information about harmed subjects, identity categories, and associated harms from AI incident reports in the AI Incident Database. The model was used to apply a structured rubric through prompt-based batch processing, including counterfactual relevance assessment, as described in the Methods sections (§3.2 and §3.3). All model outputs were reviewed by the authors and filtered according to predefined criteria before inclusion in the final dataset.

Use in Manuscript Development. We used another LLM, Google Gemini 3 Pro, via the chat interface to support grammar checking and stylistic edits in selected paragraphs of the manuscript. These tools were not used to generate original research content, results, or interpretations.

Use in Accessibility Support. We also used Google Gemini 3 Pro via the chat interface to assist with drafting initial alt-text descriptions for figures and tables. These descriptions were subsequently reviewed and revised by the authors.

7.2 Ethical Considerations Statement

Our work raises four ethical considerations.

First, our reliance on news-based incident reports means that our findings reflect which harms are deemed newsworthy rather than which occur most frequently. Harms affecting communities with less media visibility may therefore be underrepresented. We caution against interpreting the absence of certain identity configurations as evidence that such harms do not occur.

Second, categorizing individuals into discrete identity values (e.g., "Black", "woman", "lower-class") risks reifying categories that are fluid and socially constructed. We adopted this approach to enable systematic analysis while recognizing that it can obscure within-group heterogeneity.

Third, the use of LLMs to assist in coding introduces potential biases, as models may reflect stereotypes embedded in their training data. We mitigated this risk through a structured rubric and manual validation, but acknowledge that automated identification of sensitive attributes is inherently imperfect.

Fourth, our analysis emphasizes the frequency and co-occurrence of identity categories in documented incidents, but frequency alone does not capture the seriousness or impact of harm. Some harms may appear less often in incident data yet carry disproportionate social, psychological, or material consequences. Our results should therefore not be interpreted as ranking harms by importance or severity.

References

[1] Charlie Beckett, Edward Finn, Fredrik Heintz, Frederic Heymans, Suren Jayasuriya, Sayash Kapoor, Santosh Kumar Biswal, Arvind Narayanan, Agnes Stenbom, and Jenny Wiik (Eds.). 2023. *Reporting on Artificial Intelligence: A Handbook for Journalism Educators*. UNESCO. https://doi.org/10.58338/HSMK8605 Partial contributors: Jenny Bergenmar, Ammina Kothari, Bernhard Dotzler, Teemu Roos, Nicolas Kayser-Bril, Steve Woolgar.

[2] Avinash Agarwal and Manisha J. Nene. 2024. Standardised Schema and Taxonomy for AI Incident Databases in Critical Digital Infrastructure. In *IEEE Pune Section International Conference (PuneCon)*. 1–6. doi:10.1109/PuneCon63413.2024.10895867

[3] AI Incident Database. 2014. *Incident 9: NY City School Teacher Evaluation Algorithm Contested*. AI Incident Database. https://incidentdatabase.ai/cite/9/

[4] AI Incident Database. 2015. *Incident 263: YouTube Recommendations Implicated in Political Radicalization of User*. AI Incident Database. https://incidentdatabase.ai/cite/263/

[5] AI Incident Database. 2016. *Incident 40: COMPAS Algorithm Reportedly Performs Poorly in Crime Recidivism Prediction*. AI Incident Database. https://incidentdatabase.ai/cite/40/

[6] AI Incident Database. 2016. *Incident 6: Microsoft's TayBot Allegedly Posts Racist, Sexist, and Anti-Semitic Content to Twitter*. AI Incident Database. https://incidentdatabase.ai/cite/6/

[7] AI Incident Database. 2017. *Incident 13: High-Toxicity Assessed on Text Involving Women and Minority Groups*. AI Incident Database. https://incidentdatabase.ai/cite/13/

[8] AI Incident Database. 2017. *Incident 167: Researchers' Homosexual-Men Detection Model Denounced as a Threat to LGBTQ People's Safety and Privacy*. AI Incident Database. https://incidentdatabase.ai/cite/167/

[9] AI Incident Database. 2018. *Incident 188: Argentinian City Government Deployed Teenage-Pregnancy Predictive Algorithm Using Invasive Demographic Data*. AI Incident Database. https://incidentdatabase.ai/cite/188

[10] AI Incident Database. 2018. *Incident 396: Transgender Uber Drivers Mistakenly Kicked off App for Appearance Change during Gender Transitions*. AI Incident Database. https://incidentdatabase.ai/cite/396/

[11] AI Incident Database. 2019. *Incident 288: New Jersey Police Wrongful Arrested Innocent Black Man via FRT*. AI Incident Database. https://incidentdatabase.ai/cite/288/

[12] AI Incident Database. 2020. *Incident 101: Dutch Families Wrongfully Accused of Tax Fraud Due to Discriminatory Algorithm*. AI Incident Database. https://incidentdatabase.ai/cite/101/

[13] AI Incident Database. 2020. *Incident 244: Colorado Police's Automated License Plate Reader (ALPR) Matched a Family's Minivan's Plate to That of a Stolen Vehicle Allegedly, Resulting in Detainment at Gunpoint*. AI Incident Database. https://incidentdatabase.ai/cite/244/

[14] AI Incident Database. 2020. *Incident 335: UK Visa Streamline Algorithm Allegedly Discriminated Based on Nationality*. AI Incident Database. https://incidentdatabase.ai/cite/335/

[15] AI Incident Database. 2020. *Incident 74: Detroit Police Wrongfully Arrested Black Man Due To Faulty FRT*. AI Incident Database. https://incidentdatabase.ai/cite/74/

[16] AI Incident Database. 2022. *Incident 202: A Korean Politician Employed Deepfake as Campaign Representative*. AI Incident Database. https://incidentdatabase.ai/cite/202/

[17] AI Incident Database. 2022. *Incident 300: TikTok's "For You" Algorithm Allegedly Abused by Online Personality to Promote Anti-Women Hate*. AI Incident Database. https://incidentdatabase.ai/cite/300/

[18] AI Incident Database. 2022. *Incident 314: Stable Diffusion Abused by 4chan Users to Deepfake Celebrity Porn*. AI Incident Database. https://incidentdatabase.ai/cite/314/

[19] AI Incident Database. 2022. *Incident 431: Robbers Accessed Drugged Gay Men's Bank Accounts Using Their Phones' Facial Recognition*. AI Incident Database. https://incidentdatabase.ai/cite/431/

[20] AI Incident Database. 2023. *Incident 592: Facial Recognition Misidentifies Pregnant Woman Leading to False Arrest in Detroit*. AI Incident Database. https://incidentdatabase.ai/cite/592/

[21] AI Incident Database. 2023. *Incident 610: Deepfake Technology Was Used to Generate Naked Pictures of Underage Girls in Spanish Town*. AI Incident Database. https://incidentdatabase.ai/cite/610/

[22] AI Incident Database. 2024. *Incident 650: AI-Generated Images of Trump with Black Voters Spread as Disinformation Before U.S. Primary Elections*. AI Incident Database. https://incidentdatabase.ai/cite/650/

[23] AI Incident Database. 2024. *Incident 672: "Lavender"' and "The Gospel"' AI Systems Reportedly Used in Gaza Targeting Operations with Civilian Harm Allegations*. AI Incident Database. https://incidentdatabase.ai/cite/672/

[24] AI Incident Database. 2024. *Incident 717: Fake AI-Generated Law Firms Sent Fake DMCA Notices to Increase SEO*. AI Incident Database. https://incidentdatabase.ai/cite/717/

[25] AI Incident Database. 2024. *Incident 862: Purportedly AI-Generated Video Allegedly Depicts Martin Luther King Jr. Supporting Donald Trump*. AI Incident Database. https://incidentdatabase.ai/cite/862/

[26] AI Incident Database. 2024. *Incident 874: 1 in 6 Congresswomen Have Reportedly Been Targeted by AI-Generated Nonconsensual Intimate Imagery.* AI Incident Database. https://incidentdatabase.ai/cite/874/

[27] AI Incident Database. 2024. *Incident 904: Kate Isaacs, Advocate Against Image-Based Abuse, Reports Being Deepfaked.* AI Incident Database. https://incidentdatabase.ai/cite/904/

[28] AI Incident Database. 2024. *Incident 924: Alleged Deepfake Scam Uses BBC Presenter Naga Munchetty's Image to Promote Fraudulent Investment Scheme.* AI Incident Database. https://incidentdatabase.ai/cite/924/

[29] AI Incident Database. 2024. *Incident 972: Russian Influence Operation Allegedly Uses AI to Create Fake Kamala Harris Campaign Website and Rhino-Hunting Hoax.* AI Incident Database. https://incidentdatabase.ai/cite/972/

[30] AI Incident Database. 2025. *Incident 1075: New Orleans Police Reportedly Used Real-Time Facial Recognition Alerts Supplied by Project NOLA Despite Local Ordinance.* AI Incident Database. https://incidentdatabase.ai/cite/1075/

[31] AI Incident Database. 2025. *Incident 1077: FBI Reports Ongoing Vishing and Smishing Campaign Allegedly Targeting Government Officials Using Purportedly AI-Generated Voices.* AI Incident Database. https://incidentdatabase.ai/cite/1077/

[32] AI Incident Database. 2025. *Incident 1144: xAI Allegedly Operates Unpermitted Methane Turbines in Memphis to Power Supercomputer Colossus to Train Grok.* AI Incident Database. https://incidentdatabase.ai/cite/1144/

[33] AI Incident Database. 2025. *Incident 980: AI-Generated Songs Allegedly Imitating Céline Dion Circulate Online Without Authorization.* AI Incident Database. https://incidentdatabase.ai/cite/980/

[34] AI Now Institute. 2025. *AI, Algorithmic, and Automation Incidents and Controversies.* https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents Accessed: 2025-09-01.

[35] Morgan G. Ames, Janet Go, Joseph 'Jofish' Kaye, and Mirjana Spasojevic. 2011. Understanding Technology Choices and Values Through Social Class. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. Association for Computing Machinery, New York, NY, USA, 55–64. doi:10.1145/1958824.1958834

[36] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. ProPublica.

[37] Jeffrey Basoah, Jay L. Cunningham, Erica Adams, Alisha Bose, Aditi Jain, Kaustubh Yadav, Zhengyang Yang, Katharina Reinecke, and Daniela Rosner. 2025. Should AI Mimic People? Understanding AI-Supported Writing Technology Among Black Users. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 51 pages. doi:10.1145/3757423

[38] Greta R. Bauer, Siobhan M. Churchill, Mayuri Mahendran, Chantel Walwyn, Daniel Lizotte, and Alma Angelica Villa-Rueda. 2021. Intersectionality In Quantitative Research: A Systematic Review Of Its Emergence And Applications Of Theory And Methods. *SSM - Population Health* 14 (2021), 100798. doi:10.1016/j.ssmph.2021.100798

[39] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity Press.

[40] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 948–958. doi:10.1145/3531146.3533157

[41] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* 12, 1 (2024), 33–43. doi:10.1609/hcomp.v12i1.31598

[42] Avtar Brah and Ann Phoenix. 2004. Ain't I a Woman? Revisiting Intersectionality. *Journal of International Women's Studies* 5, 3 (2004), 75–86. https://vc.bridgew.edu/jiws/vol5/iss3/8

[43] Justin-Casimir Braun, Eva Constantaras, Aung Htet, Gabriel Geiger, Dhruv Mehrotra, and Daniel Howden. 2023. *Suspicion Machine Methodology.* https://www.lighthousereports.com/methodology/suspicion-machine/

[44] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[45] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[46] Dasom Choi. 2025. Designing Inclusive AI Interaction for Neurodiversity. In *Companion Publication of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*. Association for Computing Machinery, 31–34. doi:10.1145/3715070.3747338

[47] Eva Constantaras, Gabriel Geiger, Justin-Casimir Braun, Dhruv Mehrotra, and Aung Htet. 2023. *Inside the Suspicion Machine.* https://www.wired.com/story/welfare-state-algorithms/

[48] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241. doi:10.2307/1229039

[49] CSET. 2024. AI Harm Taxonomy for AIID. https://incidentdatabase.ai/taxonomy/csetv1. Accessed: 2026-03-20.

[50] Julia De Miguel Velázquez, Sanja Šćepanović, Andrés Gvirtz, and Daniele Quercia. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer* 57, 11 (2024), 71–81. doi:10.1109/MC.2024.3432492

[51] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[52] Leon Fröhling, Alessandro Giaconia, Edyta Paulina Bogucka, and Daniele Quercia. 2026. Agent-Supported Foresight for AI Systemic Risks: AI Agents for Breadth, Experts for Judgment. arXiv:2602.08565 [cs.HC] https://arxiv.org/abs/2602.08565

[53] Maya Goodwill, Roy Bendor, and Mieke van der Bijl-Brouwer. 2021. Beyond Good Intentions: Towards a Power Literacy Framework for Service Designers. *International Journal of Design* 15, 3 (2021), 45–59.

[54] Thilo Hagendorff, Leonie N. Bossert, Yip Fai Tse, and Peter Singer. 2022. Speciesist Bias In AI: How AI Applications Perpetuate Discrimination And Unfair Outcomes Against Animals. *AI and Ethics* 3, 3 (2022), 717–734. doi:10.1007/s43681-022-00199-9

[55] Patricia Hill Collins. 2002. *Black Feminist Thought.* Routledge. doi:10.4324/9780203900055

[56] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, And The Limits Of Antidiscrimination Discourse. In *Information, Communication & Society*, Vol. 22. Taylor & Francis, 900–915.

[57] Hans Hofmann. 1994. Statlog (German Credit Data). https://archive.ics.uci.edu/dataset/144. doi:10.24432/C5NC77

[58] Cara Hunter and Anna Moore. 2025. 'It Was Extremely Pornographic': Cara Hunter On The Deepfake Video That Nearly Ended Her Political Career. *The Guardian* (01 Dec. 2025). https://www.theguardian.com/society/ng-interactive/2025/dec/01/it-was-extremely-pornographic-cara-hunter-on-the-deepfake-video-that-nearly-ended-her-political-career Accessed: 2026-01-13.

[59] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, Article 305 (2021), 44 pages. doi:10.1145/3476046

[60] Ron Kohavi and Barry Becker. 1996. UCI Adult Data Set. https://archive.ics.uci.edu/ml/datasets/adult. https://archive.ics.uci.edu/ml/datasets/adult UCI Machine Learning Repository.

[61] Paola Lopez. 2024. More Than the Sum of its Parts: Susceptibility to Algorithmic Disadvantage as a Conceptual Framework. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 909–919. doi:10.1145/3630106.3658944

[62] Leslie McCall. 2005. The Complexity of Intersectionality. *Signs: Journal of Women in Culture and Society* 30, 3 (2005), 1771–1800. doi:10.1086/426800

[63] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3 (2019). doi:10.1145/3359174

[64] Matthew Miles and Michael Huberman. 1994. *Qualitative Data Analysis: A Methods Sourcebook.* Sage.

[65] Gustavo Moreira, Edyta Paulina Bogucka, Marios Constantinides, and Daniele Quercia. 2025. The Hall of AI Fears and Hopes: Comparing the Views of AI Influencers and those of Members of the U.S. Public Through an Interactive Platform. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, Article 1093, 27 pages. doi:10.1145/3706598.3714117

[66] Jennifer C. Nash. 2019. *Black Feminism Reimagined: After Intersectionality.* Duke University Press. doi:10.1215/9781478002253

[67] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

[68] Organisation for Economic Co-operation and Development. 2025. *OECD AI Incidents Monitor.* https://oecd.ai/en/incidents Accessed: 2025-09-01.

[69] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, White, 27? Bias Evaluation on Data and Algorithms for Affect Recognition in Faces. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 973–987. doi:10.1145/3531146.3533159

[70] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj

[71] K. M. Pauly. 1996. Describing the Emperor's New Clothes: Three Myths of Education (In)Equality. In *The Gender Question in Education: Theory, Pedagogy and Politics*, Ann Diller et al. (Eds.). Westview Press, Boulder, CO.

[72] Uwe Peters. 2022. Algorithmic Political Bias in Artificial Intelligence Systems. *Philosophy and Technology* 35, 2 (2022). doi:10.1007/s13347-022-00512-8

[73] Mahika Phutane, Hayoung Jung, Matthew Kim, Tanushree Mitra, and Aditya Vashistha. 2025. ABLEIST: Intersectional Disability Bias in LLM-Generated Hiring Scenarios. arXiv:2510.10998 [cs.CL] https://arxiv.org/abs/2510.10998

[74] Joanna Redden, Lina Dencik, and Harry Warne. 2020. Datafied Child Welfare Services: Unpacking Politics, Economics And Power. *Policy Studies* 41, 5 (2020), 507–526. doi:10.1080/01442872.2020.1724928

[75] Responsible AI Collaborative. 2025. *AI Incident Database.* https://incidentdatabase.ai/ Accessed: 2025-09-01.

[76] Rest of World Staff. 2024. *AI Elections Tracker.* https://restofworld.org/2024/elections-ai-tracker/ Accessed: 2026-01-13.

[77] Isabel Richards, Claire Benn, and Miri Zilka. 2025. From Incidents to Insights: Patterns of Responsibility Following AI Harms. In *Proceedings of the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery, 151–169. doi:10.1145/3757887.3763018

[78] Johnny Saldaña. 2015. *The Coding Manual for Qualitative Researchers.* Sage.

[79] Princess Sampson, Ro Encarnacion, and Danaë Metaxa. 2023. Representation, Self-Determination, and Refusal: Queer People's Experiences with Targeted Advertising. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for

Computing Machinery, 1711–1722. doi:10.1145/3593013.3594110

[80] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, 5412–5427. doi:10.1145/3025453.3025766

[81] 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 (2025), 25 pages. doi:10.1613/jair.1.17806

[82] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 642–659. doi:10.1145/3630106.3658931

[83] Justyna Stypinska. 2022. AI Ageism: A Critical Roadmap For Studying Age Discrimination And Exclusion In Digitalized Societies. *AI and Society* 38, 2 (2022), 665–677. doi:10.1007/s00146-022-01553-5

[84] The Combahee River Collective. 1982. A Black Feminist Statement. In *But Some of Us Are Brave: All the Women Are White, All the Blacks Are Men, But Some of Us Are Brave*, Gloria T. Hull, Patricia Bell Scott, and Barbara Smith (Eds.). Feminist Press, Old Westbury, NY, 13–22.

[85] The Economist. 2025. How AI is Rewiring Childhood. https://www.economist.com/leaders/2025/12/04/how-ai-is-rewiring-childhood.

[86] Tom Van Nuenen, Jose Such, and Mark Cote. 2022. Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems. *Proceedings of the ACM on Human-Computer Interaction* 6, Article 445 (2022), 30 pages. doi:10.1145/3555546

[87] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. 2024. Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database. In *Proceedings of the Conference on Artificial Intelligence (AAAI) and Conference on Innovative Applications of Artificial Intelligence (IAAI) and Symposium on Educational Advances in Artificial Intelligence (EAAI)*. AAAI Press, Article 2626, 6 pages. doi:10.1609/aaai.v38i21.30349

[88] Kyra Wilson and Aylin Caliskan. 2025. *Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval.* AAAI Press, 1578–1590.

[89] Pamela J. Wisniewski, Neha Kumar, Christine Bassem, Sarah Clinch, Susan M. Dray, Geraldine Fitzpatrick, Cliff Lampe, Michael Muller, and Anicia N. Peters. 2018. Intersectionality as a Lens to Promote Equity and Inclusivity within SIGCHI. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI)*. Association for Computing Machinery, 1–6. doi:10.1145/3170427.3186324

Appendix

A Criteria for Selecting the Source of AI Incident Reports

We defined three criteria for selecting a source of AI incident reports: (C1) whether the source preserves original materials such as serious incident reports [70], (C2) whether those materials contain sufficient detail for intersectional analysis, and (C3) whether the materials are collected and maintained with human oversight.

To evaluate *preservation of original materials*, we examined whether a source retains the full text of incident reports or instead replaces them with summaries. To evaluate *analytical detail*, we assessed whether the incident data provide enough information to identify who was harmed and how identity-related information is disclosed, either explicitly or implicitly, following the distinction introduced in prior systematic reviews [38, 80]. To evaluate *human oversight*, we examined whether the source describes editorial review during data submission and validation. These sources performed as follows across our criteria:

C1: Preservation of original materials. AIM represents incidents using LLM-generated summaries rather than preserving full report texts. AIAAIC records incidents primarily as short descriptions written by contributors and accompanied by links to external sources, but does not retain the full text of original materials. In contrast, the AIID preserves the full text of publicly available incident reports by requiring contributors to submit report content and verify that it matches the linked source.

C2: Analytical detail. AIM aggregates both AI hazards (where harm is only potential) and incidents (where harm has occurred), requiring an additional filtering step to determine whether harm occurred and who was affected. AIAAIC entries vary in detail and are often too brief to support consistent subject-level analysis, depending on how contributors describe incidents. The AIID provides the richest textual detail by preserving full report texts.

C3: Human oversight. AIM relies largely on automated processes to generate summaries, with limited documented human editorial review. AIAAIC depends on volunteer submissions with minimal standardization beyond link provision. The AIID applies documented editorial review to report submission and incident linking, with automated tools used to detect duplicate reports, and final decisions made through human judgment.

B List of Identity Categories, Their Exemplary Values, and Rules for Grouping Identity Values Within Identity Categories

B.1 Listing Identity Categories and Their Exemplary Values

We constructed this list iteratively by reviewing both identity categories and their exemplary values. We began with race, gender, and class, as identified in earliest intersectionality scholarship [48, 84], drawing directly on the values used in that work such "people of color" and "white people", "women" and "men", and "working-class" and "upper-class". We then added categories and values introduced in later work, including sexuality and nationality [55], age, disability, religion, education, language, appearance, and reproductive status [71]. To include categories relevant to contemporary AI-related harms, we further expanded the list to include body size, gender identity, immigration status [53], indigeneity, caste, political identity, health status, family status, geographic location [38], and neurodiversity [46]. Finally, to capture incidents affecting non-human entities, we included specie category related to environmental and ecological harm [54]. This process resulted in a final list of 26 identity categories paired with exemplary values (see next page).

List of Identity Categories and Their Exemplary Values

(1) **Race** (e.g., White, Black)
(2) **Gender** (e.g., Male, Female)
(3) **Gender Identity** (e.g., Cisgender, Trans)
(4) **Class** (e.g., Upper class, Working class)
(5) **Sexuality** (e.g., Heterosexual, Gay)
(6) **Nationality** (e.g., German, Syrian)
(7) **Ability** (e.g., Able-bodied, Disabled)
(8) **Gender Expression** (e.g., Masculine, Feminine, gender nonconforming)
(9) **Heritage** (e.g., European descent, African American, Indigenous heritage, diasporic)
(10) **Age** (e.g., Teenager, Adult, Middle-aged, Senior)
(11) **Appearance** (e.g., Conventionally attractive, perceived as unattractive)
(12) **Language** (e.g., Anglophone, English as a second language)
(13) **Skin Tone** (e.g., Light, Dark)
(14) **Religion** (e.g., Christian, Muslim)
(15) **Reproductive Status** (e.g., Fertile, Infertile)
(16) **Body Size** (e.g., Thin, fat, obese)
(17) **Education** (e.g., Student, professor, vocational trainee, graduate of an elite university, self-taught)
(18) **Immigration Status** (e.g., Citizen, permanent resident, temporary visa holder, undocumented migrant)
(19) **Geography** (e.g., Urban, rural, remote region, informal settlement, university town, capital city)
(20) **Indigeneity** (e.g., Indigenous person, settler descendant, colonizer lineage)
(21) **Family Status** (e.g., Single, married, divorced, single parent, caregiver)
(22) **Caste** (e.g., Brahmin, Dalit)
(23) **Political Identity** (e.g., Progressive, conservative, libertarian, socialist)
(24) **Neurodiversity** (e.g., Neurotypical, autistic, ADHD)
(25) **Health Status** (e.g., Mentally well, living with depression, chronically ill)
(26) **Species** (e.g., Human, nonhuman animal, plant, insect)

B.2 Defining Rules for Grouping Identity Values Within Identity Categories

To ensure consistency and comparability across incidents, we defined rules for grouping identity values within four identity categories: race, gender, class, and age. For race, we grouped values into two categories, distinguishing between "people of color" and "white subjects", following conventions in prior work on AI harms [65]. For gender, we grouped values into three categories (female, male, and other), where "other" captures non-binary gender identities. For class, we followed the definition by Ames et al. [35], which conceptualizes class as "a category defined by a nexus of income level, educational attainment, and type of employment". Because education was treated as a separate identity category in our rubric (see point 17 above), we grouped income and occupation values into three categories: lower, middle, and upper class. For example, "gig worker" was mapped to lower, "small business owner" to middle, and "politician" to upper class. For age, we grouped values into five ranges: children (1-9 years), adolescents (10-19 years), younger adults (20-24 years), adults (25-59 years), and older adults (60-99 years).

C LLM Prompt Operationalizing the Rubric

Part 1/3

Persona: You are an expert AI Incident Analyst. Your core expertise is the application of **Kimberle Crenshaw's intersectionality theory** to analyze AI incident reports. You are precise, context-sensitive, and you avoid flattening identities into isolated categories. To assess how identity contributes to harm, you reason causally and structurally. You often work backwards from the observed harm to trace contributing design choices or detection failures. This approach is similar to Fault Tree Analysis, where analysts start with a failure and identify underlying conditions or assumptions that allowed it to occur.

Introduction: I will provide you with a set of AI incident reports. You will perform four tasks on each report.

Tasks:
TASK 1: Extract AI Subject Details Analyze each report to identify and categorize unique AI Subjects - living entities subjected to or affected by the AI system use. Each AI Subject is identified by identity markers, which are specific attributes (e.g., "White" for Race, "Female" for Gender) drawn from 26 predefined identity categories rooted in intersectionality theory. To do the task, follow the subtasks 1.1 and 1.2.
If multiple reports refer to the same AI Subject, merge them into a single one by comparing names and identity context. Consider two AI Subjects the same if: - Their names refer to the same group (even with different phrasing, e.g., "Users of Alice", "Users of Yandex's Alice"). - Their identity markers are identical or overlapping, e.g., two reports list the AI Subjects as "Young, Spanish-speaking user" and "Teenage users who speak Spanish". Since both share the same markers for Age ("Teenager/Young") and Language ("Spanish"), they can be treated as the same AI Subject.

Subtask 1.1: Identify the AI Subject Extract the exact name of the living entity (human individuals, groups, societies, organizations, or nature) affected by the AI, verbatim as it appears in the report. Exclude inanimate objects (e.g., AI systems, websites, recommender systems, AI agents).

Subtask 1.2: Classify AI Subject Type Assign each AI Subject to one category: 1. An individual - a single person or named entity (e.g., "John Doe", "22-year-old") 2. A group of persons - a subset of people within a specific context (e.g., "Australian students", "employees", "protesters", "Amazon delivery drivers") 3. Society - a broad, regional, national, or global population (e.g., "Australians", "Canadian public", "Moldovan citizens"). If the reference is to a broad national or regional public, it belongs here. If it refers to a specific subgroup, classify under "a group of persons". 4. Organization - an institution, corporation, company, NGO, government body, university, or media agency (e.g., "Google", "UNICEF", "The Guardian", "Harvard University", "Amazon") . 5. Nature - animals, plants, or ecological entities (e.g., "rhinos", "Amazon rainforest"). 6. Other - ambiguous or unclear living entities not fitting the above categories.

TASK 2: Extract Identity Markers For each AI Subject, systematically extract identity markers explicitly or implicitly appearing in the incident across 26 predefined identity categories rooted in intersectionality theory. These markers are specific attributes within a category. Example: "White" is a marker within the category of "Race".

[List of identity categories and their exemplary values] as shown in Appendix B

Extraction Rules:
Rule 1: Explicit Markers If a marker is explicitly mentioned in the report, return: - Category: The predefined identity category (e.g., "Gender") - Marker: The exact marker wording from the report (e.g., "Non-binary") - Marker type: "Explicit" - Source: Direct excerpt from the report
Rule 2: Inferred Markers If a marker is not explicitly mentioned but can be reasonably inferred from context, return: - Category: The predefined identity category (e.g., "Family Status") - Marker: e.g., "Caregiver" - Marker type: "Inferred" - Source: Brief reasoning (1-2 sentences) explaining the inference, citing specific report details.
Do not generalize or modify markers; use report-specific evidence only. Example: For "applicants with young children were denied flexible work options" infer "Family Status": "Caregiver", with reasoning: "The phrase 'applicants with young children' indicates that the AI Subject is a caregiver, as it directly references their responsibility for children".
Rule 3: Non-Mentioned Markers If a marker is neither mentioned nor inferable, return: - Category: The predefined identity category (e.g., "Gender") - Marker: "Not mentioned" - Marker type: "None" - Source: "None"

TASK 3: Assess Causal Relevance of Identity Marker to the Incident
For each identity marker extracted in TASK 2 (explicit or inferred), assess whether the marker was causally relevant to the AI-related harm described in the report. To do so, answer the following two binary questions for each extracted marker, and provide a single reasoning explaining these two related questions: - Question 1 (Direct Cause): Did this incident happen because the AI Subject was [marker]? Return "Yes" or "No" Return "Yes" only if the marker materially contributed to the system's harmful behavior - that is, the system's decision, action, or failure changed because the subject had this identity marker. Do not return "Yes" for general errors or unrelated failures. - Question 2 (Alternate Explanation): Would this incident still have happened if the AI Subject was not [marker]? To reason about this one, imagine a version of the AI Subject who is identical in all respects except for this one identity marker (e.g., Gender = Female instead of Male). Would the AI system still have produced the same harmful outcome for this alternative subject? Return "Yes" or "No" Return "Yes" if an alternative subject that is identical in all respects except for this identity marker (e.g., different gender, race, or age) would still have experienced the same harm. Return "No" if changing the identity marker would likely have prevented or changed the outcome.
Provide the joint reasoning. Work backwards from the harm to explain how this identity marker may have contributed to it. Focus on the AI system's assumptions, design decisions, or detection failures. To guide your reasoning, consider: - Did the system fail to detect, prioritize, or adapt to something about this identity? - Was the harm caused by thresholds, classification logic, or other settings that excluded or misrepresented this identity? - Did the identity marker influence the wording, tone, or believability of the harmful output?

Part 2/3

Examples:
Incident report 1: A security robot at Stanford Mall malfunctioned and knocked over a 16-month-old boy, running over his leg. The robot was designed to detect abnormal noises, environmental changes, and known criminals. Extracted Identity Markers: Age = 16-month-old, Gender = Boy
Q1: Did this incident happen because the AI Subject was 16-month-old? "Yes" Q2: Would this incident still have happened if the AI Subject was not 16-month-old? "No"
Reasoning: Young children are less detectable by sensors or may behave in unpredictable ways. The robot likely failed to detect or properly respond to a toddler, making the subject's age a causal factor.
Q1: Did this incident happen because the AI Subject was a boy? "No" Q2: Would this incident still have happened if the AI Subject was not a boy? "Yes"
Reasoning: The robot did not act differently based on the child's gender. If the child had been a girl, the event would likely have occurred in the same way.

Incident report 2: An AI system capable of generating deepfake videos was deployed in the context of immigration-related communication to impersonate a Toronto-based lawyer. However, it malfunctioned by creating fraudulent videos that targeted newcomers to Canada, exploiting their vulnerability and confusion about immigration rules. Extracted Identity Markers: Immigration Status = Newcomer to Canada, Geography = Toronto-based
Q1: Did this incident happen because the AI Subject was a newcomer to Canada? "Yes" Q2: Would this incident still have happened if the AI Subject was not a newcomer to Canada? "No" Reasoning: The AI system targeted individuals perceived as unfamiliar with immigration procedures. The harm—deception and fraud—relied on exploiting the uncertainty and vulnerability linked specifically to being a newcomer. If the subject had not been a newcomer (e.g., a long-term resident), the deepfakes would likely not have been directed at them or would have been less effective.
Q1: Did this incident happen because the AI Subject was Toronto-based? "No" Q2: Would this incident still have happened if the AI Subject was not Toronto-based? "Yes" Reasoning: While the impersonated lawyer was based in Toronto, this detail served primarily to lend superficial credibility. The incident would likely have occurred in the same way if the lawyer had been based in any other Canadian city. The geographic detail did not causally determine the harm experienced by the subjects.

Only produce a MarkerHarm sentence when DirectScore is "Yes" and AlternateScore is "No". In those cases, the MarkerHarm must describe one concrete harmful event from the report that occurred because of that specific identity marker, written in past tense.
For all other combinations of scores, return "" (an empty string) for MarkerHarm.

Examples: Nationality: "Ghanaian voters saw false claims about their presidential candidates". Political Identity: "Opposition supporters saw AI-generated posts falsely accusing their candidate of corruption".

Here are the AI incident reports: IncidentID: "", TotalReportNumber: "", Reports: "",
Return the task results the following JSON format:

```
{{"IncidentID": "{{}}",
"Description": "[AI system name] was deployed in [context] to [intended function]. However, it [malfunctioned in a way
      that affected AI Subject]. As a result, [AI Subject] experienced [specific consequences]",
"ReportNumber": TotalReportNumber,
"AI_Subjects": {{ "S1":
   {{
   "SubjectID": "IncidentID" + "-S1",
   "ReportID": report_number,
   "Name": "The name of the living entity that is subject to or affected by AI system use",
   "Type": "Individual" / "Group of persons" / "Society" / "Organizations" / "Nature" / "Other",
   "Categories": {{
       "Race": {{
           "Marker": "Extracted or Inferred race marker",
           "MarkerType": "Extracted" / "Inferred" / "None",
           "Source": "Direct excerpt from the report or brief reasoning explaining the marker inference, citing specific
                report details",
           "DirectScore": "Yes" / "No",
           "AlternateScore": "Yes" / "No",
           "Reasoning": "If DirectScore is Yes: Briefly explain how the harm traces back to a system behavior or design
                choice that was sensitive to this identity marker. Use backward reasoning (e.g., detection failure -
                identity-linked trait - design assumption). Leave empty if DirectScore is No.",
           "MarkerHarm": "One short sentence naming the exact harmful outcome that actually occurred to subjects with
                this identity marker in this incident, stated concretely with no abstractions or generalities."
       }},
       ...,
```

Part 3/3

```
            ...,
            "Specie": {{
                "Marker": "Extracted or Inferred specie marker",
                "MarkerType": "Extracted" / "Inferred" / "None",
                "Source": "Direct excerpt from the report or brief reasoning explaining the marker inference, citing specific
                     report details",
                "DirectScore": "Yes" / "No",
                "AlternateScore": "Yes" / "No",
                "Reasoning": "If DirectScore is Yes: Briefly explain how the harm traces back to a system behavior or design
                     choice that was sensitive to this identity marker. Use backward reasoning (e.g., detection failure -
                     identity-linked trait - design assumption). Leave empty if DirectScore is No.",
                "MarkerHarm": "One short sentence naming the exact harmful outcome that actually occurred to subjects with
                     this identity marker in this incident, stated concretely with no abstractions or generalities."
            }}
        }},
        ..., "SN": {{
        "SubjectID": "IncidentID" + "-SN",
        "ReportID": report_number,
        "Name": "The name of the living entity that is subject to or affected by AI system use",
        "Type": "Individual" / "Group of persons" / "Society" / "Organizations" / "Nature" / "Other",
        "Categories": {{
            "Race": {{
                "Marker": "Extracted or Inferred race marker",
                "MarkerType": "Extracted" / "Inferred" / "None",
                "Source": "Direct excerpt from the report or brief reasoning explaining the marker inference, citing specific
                     report details",
                "DirectScore": "Yes" / "No",
                "AlternateScore": "Yes" / "No",
                "Reasoning": "If DirectScore is Yes: Briefly explain how the harm traces back to a system behavior or design
                     choice that was sensitive to this identity marker. Use backward reasoning (e.g., detection failure -
                     identity-linked trait - design assumption). Leave empty if DirectScore is No.",
                "MarkerHarm": "One short sentence naming the exact harmful outcome that actually occurred to subjects with
                     this identity marker in this incident, stated concretely with no abstractions or generalities."
            }}, ...,
            "Specie": {{
                "Marker": "Extracted or Inferred specie marker",
                "MarkerType": "Extracted" / "Inferred" / "None",
                "Source": "Direct excerpt from the report or brief reasoning explaining the marker inference, citing specific
                     report details",
                "DirectScore": "Yes" / "No",
                "AlternateScore": "Yes" / "No",
                "Reasoning": "If DirectScore is Yes: Briefly explain how the harm traces back to a system behavior or design
                     choice that was sensitive to this identity marker. Use backward reasoning (e.g., detection failure -
                     identity-linked trait - design assumption). Leave empty if DirectScore is No.",
                "MarkerHarm": "One short sentence naming the exact harmful outcome that actually occurred to subjects with
                     this identity marker in this incident, stated concretely with no abstractions or generalities."
            }} }}, }} }} }}
```

JSON Formatting Rules:
Rule 1: Go systematically through each of the ***reports***. Rule 2: Go systematically through each of the ***26 predefined identity categories***. Rule 3: Separate each identity category with a comma, including the last category. Rule 4: Ensure that the JSON response strictly follows the format with proper commas between AI subjects and identity categories. Rule 5: Before final output, deduplicate AI Subjects across all reports in the incident based on their Name and identity attributes. Return only one entry per unique subject, even if it appears in multiple reports. Rule 6: Output only JSON - do not include any other content.

D Validation of the Rubric Results through LLM Misattributions

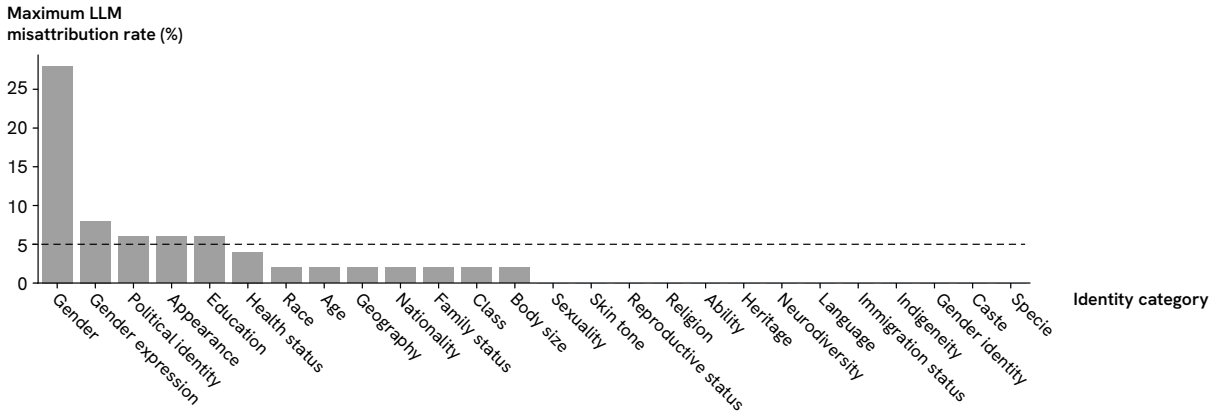

Fig. 7. **Maximum misattribution rate across annotators and the LLM by identity category.** For each category, we report the highest error rate observed across two annotators and the LLM, providing a conservative estimate of misattributions — cases where the LLM's identity category assignments differed from those of the annotators. Misattributions were generally low across categories (0–5%), with the exception of gender (28%). The majority of gender misattribution cases (90%) arose when annotators inferred gender from implicit cues (e.g., pronouns, gender-coded titles, or names), whereas the LLM behaved conservatively and did not infer it.

E Example AI Incidents Illustrating Intersections Between Identity Categories

Table 1. **Example AI incidents illustrating the most prevalent intersections between identity categories.** Each row corresponds to one cell of the heatmap in Figure 5.

Intersection	ID	Incident description and source	Harmed subjects
Age + Gender (73 incidents with 125 intersectional subjects)	188	The Technology Platform for Social Intervention was deployed in the Argentine province of Salta to predict which specific low-income girls would become pregnant as adolescents. However, it targeted and surveilled marginalized girls and young women using invasive demographic and socioeconomic data, labeling them as "predestined" for teen pregnancy. As a result, these girls and women experienced violations of privacy, stigmatization, and coercive surveillance tied to access to essential social services [9].	Age: Adolescents, Gender: Female (girls and women from low-income areas in the province of Salta, including women and girls between the ages of 10 and 19, many from migrant families and Indigenous Wichí, Qulla, and Guaraní communities)

Table 1 – continued

Intersection	ID	Description	Subjects
Age + Class (24 incidents with 32 intersectional subjects)	1144	Elon Musk's xAI supercomputer facility "Colossus" was deployed in South Memphis to provide compute power for the Grok AI chatbot. However, it was powered with large numbers of methane gas turbines. Many of them operated without permits, in a way that concentrated toxic air pollution and heavy resource use around nearby residential neighborhoods. For example, ozone levels exceeded the Environmental Protection Agency's safety standards and sensitive groups, like children and adults with respiratory issues, were advised not to go outside. As a result, residents of poor, historically Black South Memphis neighborhoods experienced increased health and environmental risks while being excluded from decision-making. [32]	Age: Children, Class: Lower (South Memphis residents)
Age + Nationality (27 incidents with 40 intersectional subjects)	101	Fraud-detection and welfare risk-scoring algorithms were deployed in Dutch tax and welfare administration to detect or predict social and childcare benefits fraud. However, they systematically labeled certain beneficiaries as high-risk or fraudulent, triggering punitive debt collection and investigations that disproportionately targeted people with dual nationalities, ethnic minorities, immigrants, low-income families, and other vulnerable recipients. As a result, these subjects were wrongly accused of fraud, pushed into poverty and debt, lost benefits, in some cases lost their children, and suffered severe psychological and social harms [12].	Age: Adults, Nationality: Iraqi
Age + Political Identity (20 incidents with 27 intersectional subjects)	202	AI deepfake avatars of South Korean presidential candidates were deployed in election campaigns to appeal to younger voters and extend candidates' reach. However, they potentially misled parts of the electorate by masking candidates' real traits, using deceptive framing, and risking misuse for fake political content, affecting voters' ability to evaluate genuine candidates and information. As a result, young South Korean voters and the broader Korean public experienced an increased risk of political deception and erosion of trust in media and democratic processes [16].	Age: Adults, Young adults, Political Identity: Voters, especially swing voters in their 20s and 30s
Age + Race (36 incidents with 47 intersectional subjects)	40	COMPAS and similar recidivism risk algorithms were deployed in criminal justice contexts to predict defendants' likelihood of reoffending and inform bail, sentencing, parole, and supervision decisions. However, they produced racially disparate error patterns, opaque and potentially inaccurate scores, and trade-secret-protected outputs that defendants and judges could not effectively scrutinize or contest. As a result, criminal defendants and incarcerated or paroled people experienced harsher or otherwise unjust decisions, including higher risk classifications, longer or more restrictive sentences, and parole denials driven by flawed or unchallengeable algorithmic assessments [5].	Age: Young adults, Race: Black
Class + Gender (22 incidents with 26 intersectional subjects)	924	AI-based deepfake generation and targeted ad systems were deployed on social media platforms to drive traffic to scam trading websites and impersonation scams. However, they generated and promoted non-consensual sexually explicit fake images and fabricated articles using celebrities' likenesses in ways that sexualised and misled people. As a result, the celebrities experienced reputational and privacy harms, while social media users were deceived and in some cases lost money to scams [28].	Class: Upper (celebrity), Gender: Female (high-profile women including Naga Munchetty, Taylor Swift and Megan Thee Stallion)

Table 1 – continued

Intersection	ID	Description	Subjects
Class + Nationality (44 incidents with 81 intersectional subjects)	9	New York State and New York City value-added teacher evaluation systems were deployed in public schools to measure and rate teacher effectiveness using statistical models on student test scores. However, they produced volatile, inaccurate, and sometimes contradictory ratings, and these scores were publicly released and tied to high-stakes consequences in ways that affected teachers. As a result, individual teachers and groups of teachers experienced unfair low or inconsistent evaluations, public shaming, and potential employment consequences based on flawed data [3].	Class: Middle (Academic), Nationality: American
Class + Political Identity (38 incidents with 59 intersectional subjects)	1077	AI-generated voice and text content was deployed in impersonation and phishing campaigns to obtain access to personal and official accounts. However, it impersonated senior U.S. officials and contacted current or former senior government officials and their contacts in deceptive ways. As a result, these officials, their contacts, and voters experienced targeted fraud attempts, privacy and security risks, and voter suppression [31].	Class: Upper, Political Identity: Political elite (White House Chief of Staff; senators, governors, and business executives; current or former senior U.S. federal or state government officials and their contacts)
Class + Race (40 incidents with 63 intersectional subjects)	335	A visa application streaming algorithm was deployed in the UK immigration context to triage entry visa applications using a traffic-light risk system. However, it explicitly incorporated nationality and produced racially biased streams that favored people from rich white countries and disadvantaged applicants from 'suspect' nationalities, especially poorer people of color. As a result, many visa applicants from these groups experienced intensive scrutiny, delays, higher refusal rates, and were sometimes unable to visit family, work, study, or attend events in the UK [14].	Class: Lower, Race: People of color
Gender + Nationality (14 incidents with 20 intersectional subjects)	672	The Lavender AI system (used with tools like Where's Daddy? and The Gospel) was deployed in the context of Israel's war in Gaza to algorithmically identify alleged Hamas and PIJ operatives and their locations for airstrikes. However, it generated mass "kill lists" of Palestinian men and linked them to family homes with minimal human review, in a framework that pre-authorized high civilian "collateral" deaths. As a result, Palestinian residents of Gaza, particularly men marked as suspects and their families (often women and children), experienced large-scale bombardment of homes, killings, and injury [23].	Gender: Male, Nationality: Palestinian
Gender + Political Identity (23 incidents with 37 intersectional subjects)	904	AI-powered deepfake and "nudifying" tools were used on social media and porn platforms to generate non-consensual sexualized images and videos of political activists such as Kate Isaacs. The fabricated content was circulated without consent and often accompanied by threats and harassment. As a result, groups of women experienced image-based sexual abuse, fear for their safety, and ongoing psychological distress [27].	Gender: Female, Political Identity: Activist

Table 1 – continued

Intersection	ID	Description	Subjects
Gender + Race (39 incidents with 64 intersectional subjects)	13	Perspective API, developed by Jigsaw/Google, was deployed in online commenting and content-moderation contexts to score and filter "toxic" or "uncivil" language. However, it systematically over-scored self-identifications and neutral references to many marginalized or politicized identities as toxic while under-scoring politely phrased bigotry, and failed to capture nuanced uncivility, leading to disproportionate flagging, suppression, or mischaracterization of speech by these groups and skewed measurements of public discourse [7].	Gender: Female, Race: Black, Latinx, White (people described with terms like "gay", "genderqueer", "deaf" in comments or news text)
Nationality + Political Identity (83 incidents with 167 intersectional subjects)	972	AI-assisted disinformation tools were deployed in the context of the 2024 US election to influence public opinion about Kamala Harris, Tim Walz, and other political figures. However, they generated and amplified fake campaign websites and deepfake videos that misrepresented these individuals and misled voters. As a result, targeted politicians and segments of the US electorate experienced reputational harm and exposure to deceptive content intended to distort democratic decision-making [29].	Nationality: American, Political Identity: Political elite (Vice President Kamala Harris; Governor Tim Walz; Representative Barry Moore, Senator Marco Rubio, Senator Marsha Blackburn, and Representative Michael McCaul)
Nationality + Race (29 incidents with 47 intersectional subjects)	650	AI text-to-image systems were used in a pre-election disinformation context to generate deepfake photographs depicting Donald Trump with Black voters and civil rights figures in order to influence political attitudes. However, these AI-generated images misrepresented Black voters' political support and manipulated perceptions of Black political behavior. As a result, Black voters and the broader public experienced deceptive visual propaganda about Black political alignment [22].	Nationality: American, Race: Black
Political Identity + Race (25 incidents with 32 intersectional subjects)	1075	AI-powered facial recognition was deployed in New Orleans via Project NOLA's live camera network to continuously scan public streets for suspects and trigger real-time alerts to police. However, it was operated in secret, outside mandated oversight and reporting requirements, in a way that affected residents, visitors, and people on or added to watchlists. As a result, these subjects experienced continuous warrantless surveillance, risk of misidentification, and arrests and detentions without transparency or due process protections [30].	Political Identity: Activists (and others speaking out or challenging government policies), Race: People of color

Received 13 January 2026; revised 25 February 2026; accepted 15 April 2026