American Research Journal of Humanities and Social Sciences

Volume 11, Issue 3, 112-126 Pages
Research Article| Open Access
ISSN (Online)- 2378-7031
DOI : 10.21694/2378-7031.25023

From Disclosure to Support: A Scoping Review of AI-Enabled First-Contact Responses to Sexual and Gender-Based Violence

Solbach Laskay1*, Pinsault Walsh, Chance Larsen

Maisonneuve-Rosemont Hospital Research Center, University of Montreal Affiliated Research Center, Montreal, QC, Canada

ABSTRACT

Background: Artificial intelligence (AI) has emerged as a potential resource for strengthening the initial stages of support available to individuals affected by sexual and gender-based violence (SGBV), including intimate partner and domestic violence. Proposed applications extend across helplines, healthcare services, specialist support organisations, online reporting environments, legal assistance platforms and other digital pathways. Despite growing interest, the extent of the available evidence, the developmental stage of existing applications and the quality of their evaluation are not yet clearly established. This review therefore sought to examine where and how AI is being applied, assess the maturity of the evidence, and identify important implementation, safety and research gaps.

Methods: A scoping review was undertaken and reported in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR). The Population–Concept–Context framework guided the review of AI-based technologies designed to support first-contact, immediate or early-stage responses. Searches were conducted in Scopus, Web of Science Core Collection and PubMed, with additional targeted searches of IEEE Xplore and the ACM Digital Library. Eligible records were screened using predefined criteria, relevant information was extracted through a structured charting process, and findings were summarised descriptively.

Results: The primary database searches retrieved 187 records, from which 21 sources were included. Supplementary searching generated 539 additional records or candidates; following assessment of 27 full-text publications, 6 further sources fulfilled the eligibility requirements. A total of 27 sources were therefore included. The literature encompassed AI-supported conversational assistance, emergency and specialist-service screening and triage, analysis of social media disclosures and online help-seeking, survivor-informed chatbot development, legal and support-service navigation, and speech-related technologies. Most studies concentrated on technical accuracy, system characteristics, usability, acceptability or audit-based findings. Evidence examining AI within established service workflows was absent, while outcomes relating to survivor wellbeing, successful connection with services and potential harms were infrequently evaluated.

Conclusions: Current evidence on AI-supported early responses to SGBV is diverse but remains at a relatively preliminary stage. Existing research has predominantly examined individual tools and prototypes rather than their effects across complete support pathways. Greater attention is needed to human oversight, escalation procedures, accountability, equitable access, digital safety and effective handover to appropriate professionals. The practice priorities and outcome areas identified in this review are intended to inform future pilot and implementation research rather than serve as validated practice guidelines.

KEYWORDS: Sexual and Gender-Based Violence; Sexual Violence; Intimate Partner Violence; Domestic Violence; Artificial Intelligence; Early Response; Crisis Support; Conversational Agents; Chatbots; Triage; Digital Safety; Scoping Review.

INTRODUCTION

The sexual violence and gender-based violence (GBV), encompassing intimate partner violence (IPV) and domestic violence, present substantial challenges at the point when survivors first seek assistance. Initial disclosure may occur during periods of fear, uncertainty, immediate danger, or emotional distress, creating a need for prompt, sensitive and trauma-informed support. At the same time, survivors may need to navigate complex and disconnected systems involving crisis services, healthcare providers, emergency departments (EDs), specialist organisations, legal assistance and referral networks [1,2]. Although these services provide essential entry points into care and protection, their capacity can be affected by workforce shortages, differences in staff expertise, limited service availability and gaps between initial disclosure, urgent assessment and continued support [3,4]. Consequently, the quality of the first interaction can have important implications. The ability to communicate a concern, describe immediate needs, receive a supportive response, identify appropriate services and access practical safety information may influence whether an individual remains engaged with available support or withdraws from the help-seeking process [5-9].

Artificial intelligence (AI) has attracted growing attention as a possible means of supporting these early interactions [10-17]. Its potential role in this setting should not be understood as a replacement for advocates, healthcare professionals, counsellors or legal practitioners. Instead, AI may have value as an additional layer of support within the earliest stages of a survivor's journey, particularly where services face high demand or require consistent availability. Potential applications include recognising indicators of urgent risk within written or spoken communication, categorising immediate support requirements, directing users towards relevant services, assisting with structured information collection or reporting, and providing limited, carefully designed guidance. Any such application, however, depends on clearly defined operational limits, dependable pathways for escalation and meaningful human oversight. The responsibilities associated with automated recommendations and decisions must also remain transparent and subject to evaluation [12-21].

The use of AI in violence-response settings also raises concerns that extend beyond conventional questions of technological accuracy. These environments are inherently sensitive because survivors may experience coercive control, monitoring by an abusive partner, restricted access to private devices or limited control over their digital communications [19]. Under such circumstances, strong technical performance alone cannot establish that a system is safe. An AI tool may perform well during testing yet create risks if urgent situations are not recognised in time, escalation fails, advice contributes to unsafe actions or disclosures, digital activity can be discovered by another person, or the system performs poorly for linguistic, cultural or demographic groups that were inadequately represented during development. These challenges support the use of an evidence-mapping approach rather than the statistical pooling of findings, particularly given established guidance for scoping reviews and the staged assessment of emerging AI-enabled interventions [22-25]. Wider frameworks addressing digital-health evaluation and AI governance similarly emphasise the importance of assessing technologies in relation to their evidence base, governance requirements and role within existing service systems rather than treating digital tools as independent substitutes for human care [26-28].

A scoping review is particularly suitable for examining this developing field because relevant evidence remains diverse and is spread across biomedical research, computer science, engineering, human–computer interaction, legal studies and social-service research. Existing publications vary considerably in terms of the populations addressed, the type of AI used, underlying data sources, intended position within the support pathway and methods used to assess performance. This diversity, together with the relatively early stage of many applications, limits the usefulness of a conventional meta-analysis or an effectiveness-focused systematic review at present. A broader mapping approach can instead clarify the nature of available work, identify the points at which AI is intended to contribute to survivor support and determine which areas have received meaningful evaluation.

Accordingly, this scoping review maps peer-reviewed evidence concerning AI-enabled technologies intended to assist survivors of sexual violence and GBV during first-contact and early-response stages. The review examines the settings in which these technologies have been developed or applied, their AI modalities, proposed functions, relationship to early support pathways and overall level of evaluative development. It also considers the extent to which studies have assessed technical performance, usability and acceptability, together with issues that are relevant to implementation in real-world services. By identifying areas where evidence is limited or absent, the review highlights the research and translation challenges that must be addressed before AI systems can be regarded as safe, equitable and appropriately integrated components of violence-response services [29].

In addition to mapping the evidence, this review proposes several practice-oriented considerations intended to support future research and implementation. These include evidence-maturity heuristic, safety-focused principles for escalation, an illustrative threat model reflecting the risks associated with coercive control and potential outcome domains for pilot and implementation studies. These proposals are derived from the patterns identified through the review and informed by relevant wider governance and implementation literature. They should therefore be interpreted as author-informed considerations for further testing and refinement, rather than as consensus-based recommendations or formally validated guidelines.

METHODS

Review Design and Reporting

This study used a scoping review design to examine and map peer-reviewed literature on AI-enabled technologies developed for early assistance and first-response support following sexual violence and gender-based violence (GBV), including intimate partner violence and domestic violence.

The scope covered technologies relevant to crisis services, initial response settings, legal and support pathways, and digitally mediated forms of help-seeking. The review was designed and reported in alignment with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) [22]. The overall methodological approach was also guided by established principles for conducting scoping studies [23].

The review protocol was not registered before the study commenced. No retrospective registration has been completed, and none is intended. To promote transparency and allow detailed examination of the review process, the Supplementary Materials contain the complete search strategies, PRISMA-ScR checklist, data-charting template, M0–M3 evidence-maturity conceptual crosswalk, detailed extraction tables, source-level coding framework, full-text screening decisions for supplementary searches, and the final accounting of retrieved and included records [30].

Information Sources and Search Strategy

The main literature searches were conducted on 16 June 2026 using Scopus, Web of Science Core Collection, and PubMed. Search terms were developed around three principal areas: sexual violence and GBV; settings associated with first response and victim or survivor services; and AI-related approaches and technologies. The AI component incorporated terms relating to artificial intelligence, machine learning (ML), deep learning, natural language processing (NLP), large language models (LLMs), conversational systems, chatbots, speech recognition, and sentiment analysis. Both controlled vocabulary and free-text terminology were used where appropriate. Broad and more focused versions of the search strategies were developed for individual databases, after which retrieved records were exported, combined, and deduplicated to form the master dataset. The complete search strings for each information source are available in the Supplementary Materials [31].

Because relevant research may also be published within computer science and engineering literature, additional targeted searches were undertaken on 9 July 2026 in IEEE Xplore and the ACM Digital Library. Records retrieved from IEEE Xplore were exported in a structured format and deduplicated within that source. Results from the ACM Digital Library were assessed within the platform because bulk structured export was not available. Records considered potentially relevant were retained for full-text review, and the decisions made at this stage are documented in the Supplementary Materials. Any remaining overlap between supplementary and original searches was identified during the screening process.

No language restrictions were applied during initial retrieval. However, eligibility assessment at the full-text stage was restricted to publications available in English because complete evaluation could only be undertaken in that language by the review team. Grey literature was outside the scope of the search, including service reports, operational documents, programme guidance, and hotline materials. Regional databases were also not searched. The potential consequences of these search boundaries are considered in the Limitations Section [32].

Eligibility Criteria

Eligibility criteria were established before screening using a Population–Concept–Context framework. The population included individuals who had experienced sexual violence and/or GBV. The concept required the presence of an identifiable AI component, including, but not limited to, ML, NLP, LLMs, conversational agents, and chatbot-based systems. To be eligible, the technology needed to have a stated role in survivor-facing support, screening, triage, reporting, prioritisation, legal or service navigation, or the early identification of concerns by clinicians or service personnel during the response process.

Relevant contexts included crisis hotlines, text- and chat-based support services, anti-violence response systems, acute healthcare settings, specialist services, legal support environments, and other digital first-contact pathways designed to facilitate timely assistance, referral, reporting, or professional intervention [33].

The concept of *first-response support* was defined in advance to maintain a clear boundary around the review. A record was eligible when it addressed at least one of the following functions: direct first-line assistance for survivors; decision support for clinicians, advocates, or other professionals during an initial interaction; service-adjacent triage, referral, routing, or reporting; or an enabling AI modality specifically intended for use within helplines, emergency care, specialist services, or support-navigation pathways.

Studies were excluded when they did not contain an explicit AI element, addressed awareness or prevention alone without a defined response pathway, were unrelated to survivor assistance or early-stage response, or represented secondary evidence syntheses. Protocols and conceptual publications were considered only when they described an AI-supported pathway model with a direct connection to first-response workflows. Such publications were retained as conceptually relevant evidence but interpreted separately from studies reporting empirical evaluations [34].

Deduplication and Study Selection

Duplicate records were removed using a deterministic approach. Exact matches based on digital object identifiers (DOIs) were prioritised, followed by comparisons using normalised publication titles. Screening of records from the original database searches was undertaken independently by two reviewers at both the title/abstract and full-text stages. Predefined eligibility criteria guided each stage, and differences in judgement were discussed until consensus was reached.

The supplementary IEEE Xplore and ACM Digital Library searches were assessed using the same eligibility framework. Initial screening and full-text assessment for these records were undertaken by one reviewer and subsequently checked by a second reviewer. Any unresolved questions were addressed through final consensus. Decisions made during screening were recorded using structured reason taxonomy, allowing exclusions to be accounted for systematically and distinguishing studies used as background literature from sources included as evidence in the review [35].

Data Charting and Synthesis

Information from eligible sources was collected using a structured data-charting form. Extracted variables included bibliographic details, publication type, service environment, country or geographical region, study population, data source, AI methodology, intended application, relevance to first response, evaluation design, reported outcomes, human oversight and escalation arrangements, and issues related to privacy, safety, and equity. The charting form was tested before full extraction to ensure that categories could be applied consistently across the diverse evidence base [36].

Data extraction was initially completed by one reviewer and then examined by a second reviewer to assess completeness, consistency, and the appropriateness of category assignments. Questions arising during charting were resolved through discussion and consensus. Study authors were not contacted to obtain missing information or clarification.

The findings were summarised descriptively and organised into an evidence map covering several broad areas: survivor-facing and hotline-adjacent digital assistance; first-response applications in healthcare and specialist settings; social-triage models relevant to crisis support; AI-supported legal and service navigation; and enabling approaches, including speech-based and privacy-oriented modelling. Statistical pooling was not considered appropriate because of the objectives of the scoping review and the substantial differences in study designs, interventions, outcomes, and levels of development. A formal risk-of-bias assessment was likewise not undertaken [37,38].

To provide a structured description of how extensively individual sources had been evaluated, each included publication was assigned an evidence-maturity category according to the most advanced level of evaluation reported. **M0** represented technical development, technical validation, or a protocol without reported evaluative findings. **M1** referred to evidence relating to usability, acceptability, systems auditing, or qualitative implementation and ethics considerations. **M2** indicated pilot evaluation within an operational workflow, including documented procedures for human handover or escalation. **M3** represented pathway-level assessment in which service-related outcomes and safety monitoring were reported.

The M0–M3 classification was developed for this review as a descriptive staging heuristic and should not be interpreted as a validated tool for critical appraisal. Its development was informed by staged approaches to evaluating emerging digital health interventions and AI-enabled systems, together with evidence-tier frameworks relevant to digital technologies [24,25,26]. Supplementary Table S5 presents a conceptual comparison between the M0–M3 categories and established staged-development, AI decision-support reporting, and digital-health evidence frameworks. This comparison is intended to aid interpretation only and does not imply formal equivalence, measurement validity, or direct correspondence between the frameworks [39,40].

RESULTS

Search Results and Study Selection

The initial searches conducted in Scopus, Web of Science Core Collection, and PubMed retrieved 187 records. Following deterministic removal of duplicate records, 97 unique publications remained for title and abstract screening. Of these, 74 records were excluded, leaving 23 reports for retrieval. One report could not be accessed, and 22 full-text articles were subsequently evaluated against the eligibility criteria. One of these publications was excluded from the evidence synthesis because it was a systematic review; however, it was retained for background and contextual discussion. Consequently, the original database searches contributed 21 sources to the review [41-45].

The targeted supplementary search, undertaken on 9 July 2026, produced 123 records from IEEE Xplore after within-source deduplication and 416 records or candidates from the ACM Digital Library, which were screened directly within the platform. Any remaining duplication between the primary and supplementary searches, as well as overlap between the supplementary sources, was addressed during screening. A total of 512 supplementary records or candidates were excluded because they were outside the review scope, duplicated existing records, represented secondary literature, or were considered background material only. Twenty-seven supplementary full texts were assessed in detail, of which 6 fulfilled all eligibility requirements. The final scoping review therefore included 27 sources of evidence (Figure 1).

Figure 1. PRISMA-ScR flow diagram. The flow diagram presents the combined record identification and selection process from the original database searches and the targeted supplementary searches. ACM Digital Library records were screened within the platform. Detailed search strategies, full-text eligibility decisions, and final record counts are reported in the Supplementary Materials.

Overview of Included Evidence

The 27 included sources represented a broad and diverse body of evidence (Table 1). Applications extended across survivor-facing digital support, tools associated with crisis hotlines, screening and triage systems used in healthcare and specialist settings, online and social-media approaches relevant to crisis prioritisation, legal and service-navigation technologies, and enabling approaches such as speech-based modelling. Across these different areas, AI was generally positioned as a supportive component within existing or anticipated response systems. Its proposed functions included providing information, identifying potential concerns, classifying needs, supporting prioritisation, assisting with reporting, offering legal information, and directing users towards relevant services. None of these applications were primarily presented as substitutes for professional advocacy, clinical assessment, legal expertise, or human judgement [11,12,13,14,15,16,17,18,20,29,30,31, 32,33,34,35,36,37,38,39,40,41,42,43,44,45,46].

Table 1. Short characteristics of the included sources of evidence (n = 27). Detailed source-level characteristics and extraction data are presented in Supplementary Table S4a,b.

When assessed using the M0–M3 descriptive evidence-maturity heuristic, most of the literature remained within the earliest stages of development. Sixteen sources were classified as M0, indicating technical development, retrospective model validation, protocol-stage research, or conceptual pathway proposals without testing within a real support pathway. The remaining 11 sources were assigned to M1 on the basis of usability or acceptability findings, systems auditing, preliminary feedback from users, or qualitative evidence concerning implementation. No study was classified at the M2 or M3 levels. In particular, no included source reported a pilot embedded within a functioning workflow with clearly documented handover or escalation arrangements, and none evaluated pathway-level service outcomes alongside systematic safety or adverse-event monitoring. The distribution of evidence across the maturity categories is summarised in Table 2.

The geographical distribution of the evidence was also uneven. Included studies or systems were associated with Europe, North America, East Asia, South Asia, Southeast Asia, Africa, Latin America, Hong Kong, Kenya, Bangladesh, Thailand, India, and Peru. Several chatbot audits and platform-based analyses also involved multiple countries or datasets that could not be linked to a single service jurisdiction. Nevertheless, a considerable proportion of the literature originated from relatively well-resourced digital-health and human–computer interaction environments. This pattern restricts the extent to which findings can be assumed to apply across different languages, cultural contexts, legal frameworks, service infrastructures, and conditions of digital risk.

Among the 27 included sources, 25 presented empirical, technical, qualitative, usability, audit, design, or other evaluative findings. One publication was a study protocol, while one was a conceptual perspective included because of its direct relevance to an AI-supported first-response workflow. Expanded information on study characteristics and evaluation findings is provided in Supplementary Table S4a,b.

Survivor-Facing and Hotline-Adjacent First-Line Digital Support

The largest group of included sources concerned AI-supported technologies intended for direct or near-direct interaction with survivors. These applications included chatbots, conversational agents associated with hotline functions, mobile applications, and other digital interfaces designed to facilitate early help-seeking. Several studies specifically examined survivor-facing systems. One domestic violence chatbot was developed and assessed through survivor-informed needs exploration and subsequent usability and acceptability testing, including feedback from a substantial user sample. Its intended role was primarily to provide information and guide users towards appropriate support rather than replace professional assistance [11]. A multi-country audit covering domestic violence chatbots assessed both rule-based and LLM-enabled systems, examining issues such as privacy practices, discreet access features, interface design, responses to simulated scenarios, and potential safety concerns [13]. Other studies explored a sexual violence chatbot using a combined rule-based and ML approach informed by survivor-related questions [14], as well as a chatbot designed for people affected by image-based sexual abuse that was tested for its capacity to provide immediate information and emotional support [46].

Further applications included a domestic violence mobile platform developed in Bangladesh that combined an NLP-enabled chatbot with geolocation and emergency-alert features and underwent usability testing [16]. Another platform developed in Peru incorporated a conversational assistant to support anonymous communication and direct users towards specialist appointments in situations involving gender violence or discrimination [17]. A mixed-methods investigation compared the quality of responses generated by general-purpose LLM chatbots with those of a domestic violence-specific chatbot, with particular attention to empathy, response quality, bias, and responsible handling of user information [36]. Qualitative research involving domestic violence survivors in Hong Kong also examined expectations regarding digital support platforms and AI chatbots. Participants highlighted the importance of safety, ease of access, clear step-by-step guidance, non-judgemental interaction, and the continued availability of human support alongside AI [44].

Evidence related to hotline-adjacent technologies also addressed questions about the appropriate boundaries of AI use. One document-based study examined how national sexual assault and domestic violence crisis services communicate about AI-powered chatbots and how those descriptions may influence user expectations or encourage excessive assumptions about the human-like capabilities of such systems [18]. Another study based on stakeholder perspectives identified duty of care, accountability, and clearly defined escalation mechanisms as central concerns for domestic violence support chatbots [20]. A conference publication outlined a proposed AI-enabled approach for hotline screening, prioritisation, escalation, and documentation, but did not present primary evidence demonstrating effectiveness within operational services [33]. Prototype chatbot studies from Kenya and India reported technical assessment or early-stage user feedback without evaluating service-level outcomes [37,38].

Overall, the evidence in this group concentrated on technical feasibility, user experience, acceptability, systems auditing, and design principles. There was little evidence demonstrating that AI-supported first-contact tools had successfully achieved reliable human handover, completed referral, continued survivor engagement, repeat service contact, or measurable improvements in safety outcomes.

First Response in Healthcare and Specialist Services

A separate group of studies examined AI applications in clinical and specialist first-contact environments. These systems primarily used NLP, ML, or multimodal approaches to support the identification of possible violence exposure, pre-screening, risk recognition, or triage within settings where professionals remained responsible for subsequent assessment and action. One large retrospective emergency department study used rule-based NLP to identify encounters potentially related to IPV that were not fully captured through administrative coding, suggesting a possible role for the approach in prompting additional screening or referral [29]. Another emergency care study applied rule-based NLP to narrative data from the National Electronic Injury Surveillance System to identify IPV-related visits and patterns associated with injury circumstances [42]. A further multimodal model combined structured clinical information with free-text notes to estimate IPV risk across validation cohorts, although it remained an analytical decision-support system rather than an intervention tested within an active service pathway [43].

Additional studies included a supervised ML approach designed to pre-screen IPV patients for possible traumatic brain injury [30], a triage model developed within a Sexual Assault Referral Centre to identify pre-existing mental health difficulties among children affected by sexual assault [31], and a protocol proposing AI-based modelling of psychosocial, judicial, and post-traumatic stress disorder trajectories following recent sexual assault within a forensic medical centre pathway [32].

Taken together, these applications mainly positioned AI as an additional screening, surveillance, or flagging mechanism within professional-led services. Because the intended settings involved clinicians or specialist personnel, human involvement was generally inherent to the broader context. However, explicit evaluation of escalation thresholds, staff responses to AI outputs, effects on workflow, downstream referral outcomes, and systematic safety monitoring was largely absent.

Social-Triage Approaches for Crisis Services

Five included studies used ML, deep learning, LLMs, or topic-modelling methods to analyse social-media posts, online communities, or discussion forums. Their stated or implied purpose was to assist with identifying urgent content, classifying disclosures, understanding support needs, or supporting prioritisation and professional interpretation. Several early studies based on Facebook content examined the classification of critical domestic violence posts, different forms of posting behaviour, or disclosure intentions, with the potential aim of supporting crisis attention and routing [34,39,40].

Another study employed an LLM-based approach to classify information needs expressed by survivors in online health communities. Categories included legal assistance, safety planning, and access to shelters. Performance decreased when the model was evaluated outside its original context, highlighting concerns about domain shift and the transferability of AI systems across populations and platforms [35]. A mixed-methods investigation of IPV-related forum discussions combined qualitative analysis with supervised ML and topic modelling to identify different forms of IPV and associated contextual themes [45].

These studies provided useful information on classification performance and the potential interpretation of online disclosures. However, they rarely showed that the resulting models had been incorporated into live, staffed crisis services or that their outputs led to verified improvements in response time, referral, or survivor outcomes.

Legal and Support-Routing AI

Two sources focused principally on legal information and support navigation. One system from Thailand used an AI chatbot to provide legal guidance to survivors of sexual violence by linking user-described circumstances with relevant Supreme Court decisions. The study reported model development and technical validation but did not evaluate outcomes following use within an established service pathway [12]. A later retrieval-augmented system developed in India provided simplified legal information, estimated possible statutory relief, and suggested next steps for individuals seeking assistance under domestic violence law. Its evaluation included prototype and model testing together with preliminary perspectives from survivors and practitioners, but no pathway-level or healthcare-related outcome assessment was reported [46].

These findings indicate that AI-assisted legal navigation may represent a developing area of support for survivors. Nevertheless, the available studies should not be interpreted as evidence that such technologies have been validated as effective healthcare, crisis-intervention, or comprehensive violence-response systems.

Enabling Modalities and Privacy-Oriented Modelling

One included technical study investigated the use of speech-based methods to identify a condition associated with GBV victimisation. The study also explored speaker-independent modelling as a possible way to reduce risks of personal re-identification and discussed potential future relevance to helpline and telehealth environments [41]. Although the work addressed an important concern in privacy-sensitive response settings, it remained limited to model development and did not demonstrate implementation within an operational service or assess effects across a complete support pathway.

Outcomes, Service Endpoints, and Digital Trace Safety

The outcomes reported across the included evidence were concentrated within four main areas: technical performance, usability and acceptability, systems auditing, and qualitative findings related to implementation or governance. Measures of technical performance were particularly common in studies involving healthcare screening, social-media triage, legal NLP, support-needs classification, and speech-based modelling. Reported measures included accuracy, harmonic-mean precision–recall scores, area under the receiver operating characteristic curve, precision, recall, and response time [12,14,29,30,31,32,33,34,35,37,39,40,41,42,43,45].

Usability, acceptability, interface assessment, and user feedback were more frequently reported for survivor-facing chatbots, mobile applications, and other digital support technologies [11,13,15,16,17,18,36,38,44,46]. Qualitative evidence concerning governance and implementation drew attention to issues such as appropriate system boundaries, responsibility for AI-mediated interactions, accountability, and the conditions required for safe deployment [18,20,33].

No included source provided a complete prospective evaluation of an AI system operating within a service workflow that tracked human escalation, successful completion of referrals, and safety outcomes at the pathway level. Some individual pathway-related features were described or audited, but they were not examined as part of a comprehensive implementation evaluation.

Consideration of digital trace safety became more visible within the broader evidence base identified after the supplementary search, although systematic assessment remained uncommon. Features and risks such as safe-exit options, deletion of sensitive information, no-history modes, exposure to device monitoring, unauthorised access by an abusive person, privacy-preserving data retention, multilingual and culturally appropriate safety, and procedures for responding to safety incidents were seldom evaluated as measurable outcomes. When these issues were discussed, they were generally treated as design requirements, theoretical concerns, or findings from systems audits rather than as implementation endpoints subjected to prospective testing [13,15,20,41].

DISCUSSION

Pathway-Centred Interpretation and Evidence Maturity

The findings of this scoping review show that AI use within violence-response settings cannot be understood as a single type of intervention. Instead, the evidence covers a range of related but distinct functions, including survivor-facing chatbots, mobile support and reporting applications, conversational systems associated with crisis services, clinical screening and surveillance tools, classifiers for online disclosures, legal and support-navigation technologies, and enabling approaches involving speech analysis or privacy-oriented modelling. Across these applications, the prevailing role of AI is supportive rather than substitutive. Most systems are intended to assist with activities such as structured intake, rapid prioritisation, limited information provision, needs classification, or routing towards appropriate human services, rather than replace professional judgement or direct human care [33].

A central finding is that the evidence remains concentrated before the point at which service impact can be demonstrated. Even technologies specifically designed for survivors were generally evaluated through model performance, usability, acceptability, system audits, or preliminary prototype feedback. The supplementary searches of IEEE Xplore and the ACM Digital Library expanded the range of identified chatbot and digital-support applications, particularly within computer science and human–computer interaction research. However, the overall maturity profile remained unchanged, as none of the included sources reached the M2 or M3 levels of the descriptive evidence-maturity heuristic [47-49].

This pattern is broadly consistent with wider evidence concerning AI applications related to IPV, where technical and developmental innovation has progressed more rapidly than evidence regarding sustained effectiveness, equity, privacy, safety, and implementation within health and social support systems [47]. The present findings therefore warrant a cautious interpretation. Signals of technical feasibility and model performance are increasing, but substantially less is known about outcomes that matter after the AI interaction itself. These include survivor-centred benefit, the quality and timeliness of escalation, successful connection with services, unintended harms, and the long-term integration of these tools within functioning response pathways.

A pathway-based perspective should also avoid defining an effective first response solely in terms of rapid classification and efficient transfer. The available studies do not provide sufficient evidence to draw firm conclusions about the relational quality of AI-mediated interactions. Nevertheless, research concerning survivors and crisis-support communication indicates that disclosure can be gradual, emotionally difficult, and shaped by a person's ability to decide when, how, and how much to disclose [48,49,50,51]. A technically efficient system could therefore still fail to provide meaningful support if it encourages disclosure before a survivor is ready, requires unnecessary repetition of traumatic information, or reduces the interaction to the collection of data rather than a process involving listening and responsive care.

The potential contribution of a low-threshold digital entry point should consequently be considered not only in terms of speed or automation, but also in terms of what it may preserve. Survivor autonomy, control over disclosure, and a safe transition to accountable human assistance are particularly important considerations. Future research should therefore examine whether AI-supported first-contact systems allow sufficient space for recognition, listening, witnessing, and care while supporting access to appropriate services [49,50,51,52]. These issues should be viewed as interpretive and safety considerations emerging from the review rather than as benefits already demonstrated by the included evidence [53].

Managing Heterogeneity and Defining the Scope

The diversity of the identified evidence makes a clearly defined eligibility boundary essential. Survivor-facing applications, clinician-supported identification systems, social-triage models, legal guidance tools, and speech-based technologies were included because each had an explicit connection with early support, reporting, triage, identification, or routing. Their proximity to a first-response pathway, however, does not mean that they provide equivalent forms of evidence or can be compared as if they were the same intervention [54-57].

The included literature also spans child sexual assault, adult sexual violence, IPV, domestic violence, and broader forms of GBV. These experiences arise within different clinical, legal, social, cultural, and service environments. Findings from one population or response pathway should therefore not automatically be extended to another. Similarly, a retrospective NLP model developed using emergency department records, an audit of a deployed chatbot, an AI-based legal information system, and a prototype mobile application address fundamentally different questions. Their reported outcomes and implications for practice are consequently different [58-63].

For this reason, the Results and accompanying tables distinguish among publication type, first-response relevance, evidence characteristics, and maturity category. Protocols, conceptual pathway models, and technical feasibility studies contribute to understanding the landscape of emerging innovation, but they should be interpreted as entries within an evidence map rather than as demonstrations of real-world service effectiveness.

Safety-Critical Escalation and Human Handover

Although many included systems refer to escalation or connection with human services, the operational details of these processes are rarely described or evaluated. Studies examining chatbots and crisis-related technologies suggest that users may attribute greater understanding, responsibility, or capability to an AI system than the system is actually designed to provide. Clear communication about the boundaries of the technology is therefore an important component of safety [13,18]. Evidence based on stakeholder perspectives likewise points to the need for identifiable responsibility, defined duties of care, and explicit procedures for escalation [20].

Despite these concerns, none of the included studies evaluated escalation as a fully specified and testable workflow. Important questions were generally left unanswered, including what circumstances should trigger human intervention, how quickly a response should occur, what happens when a designated service is unavailable, how failures are detected, and who remains accountable when escalation does not occur as intended [64].

In response to this gap, Table 3 presents a set of **author-informed minimum safety considerations** for AI systems intended to provide survivor-facing first-response support. These considerations are not proposed as formal guidelines. Rather, they synthesise recurring issues related to system boundaries, escalation, and human support identified within the review and related literature on domestic violence chatbots, sexual violence conversational systems, and text-based crisis communication [48,49,50]. They also reflect concerns identified in survivor-facing chatbot and humanitarian GBV settings regarding accessible support, reliable transfer to appropriate services, and the risks of allowing automated interaction to displace necessary human assistance [51,52]. More general research on conversational AI is considered only in relation to concerns about simulated empathy, anthropomorphic expectations, and untested safety claims in sensitive support contexts [53,54,55].

What Should Be Measured: Core Outcomes and Evaluation Priorities

Across the reviewed literature, technical measures dominate the assessment of AI systems. Accuracy and related model-performance indicators are valuable for understanding whether a system can complete a defined computational task. However, strong performance on documentation, classification, detection, or prediction does not by itself establish that a technology improves first-response support for survivors.

Future pilot studies should therefore evaluate service integration directly rather than infer practical value from model accuracy. Evaluation should also distinguish between outcomes experienced by survivors and outcomes relevant to service delivery. Potential survivor-centred measures include perceived safety, trust in the interaction, emotional distress, empowerment, control over disclosure, avoidance of re-traumatisation, and successful connection with personally chosen forms of support.

Service-level measures may include the accuracy of escalation decisions, completion of human handover, uptake of referrals, time taken to reach an appropriate service, effects on responder workload, rates of incorrect routing, and the identification and reporting of adverse events. Considering these outcome domains together may provide a more meaningful assessment of pathway performance than technical evaluation alone.

Equity, Adversarial Conditions, Privacy, and Governance

Equity concerns are often discussed broadly in relation to AI, but the evidence reviewed here suggests several practical mechanisms through which inequity may arise. Systems trained using limited linguistic, clinical, geographic, or platform-specific data may perform differently when introduced to populations or contexts that were not adequately represented during development. Such concerns are evident in research involving domestic violence chatbot design and LLM-based classification of survivor information needs [20,35]. In violence-response settings, performance degradation across groups should be understood not simply as a technical limitation, but also as a potential issue of ethics and survivor safety [56,57].

Research on privacy-preserving AI further supports the importance of collecting and retaining only information that is genuinely necessary and designing systems with awareness of the digital traces they create [58]. Wider work concerning fairness, inclusion, and intersectional approaches to AI also highlights the need to report performance across relevant subgroups and contexts rather than assuming that overall performance represents all potential users equally [59,60,61]. Decolonial and privacy-oriented perspectives provide an additional caution: evidence generated in one legal, social, cultural, or technological environment cannot be assumed to establish safety or acceptability in another setting [62,63].

Violence-response services must also be understood as potentially adversarial environments. Survivors may use monitored devices, shared accounts, or communication channels subject to surveillance. They may experience compelled disclosure, coercive control, or retaliation associated with visible digital activity. These risks differ substantially from those associated with many conventional healthcare chatbot applications. Audits of deployed systems have already identified limitations related to privacy and discreet access [13], while experimental work has demonstrated ways in which conversational AI may itself be exploited to facilitate, intensify, or commercialise coercive control [64].

Table 1: Author-informed minimum safety considerations for survivor-facing first-response AI.

Element	Minimum Consideration	Evidence to Report
Escalation triggers	Predefined red-flag signals that prompt handover to a trained human responder.	Trigger list, threshold logic, sensitivity analyses, and false-positive/false-negative handling.
Time-bounded handover	Rules for maximum time in automated mode and behaviour when no human is available.	Response-time targets, after-hours fallback, and documented failure modes.
Structured handoff summary	Survivor-controlled summary to reduce re-telling and preserve trauma-informed interaction.	Handoff template, consent controls, completeness and user-control measures.
Boundary setting and transparency	Clear statements of capabilities, limits and emergency guidance; avoid misleading anthropomorphism.	Disclosure text, comprehension checks, usability or audit findings.
Privacy minimisation and trace safety	Default minimisation of stored data and on-device traces.	Retention policy, data-flow summary, quick-exit/no-history options, coercive-control risk assessment.
Audit and accountability	Decision trail sufficient for incident review while respecting privacy minimisation.	Audit fields, access controls, governance owner and role allocation.
Safety incident response	Process for detecting, reporting and remediating unsafe outputs or escalation failures.	Incident taxonomy, monitoring plan, remediation timelines and post-incident review.
Equity and portability	External validation and subgroup reporting for tools that route or prioritise survivors.	External datasets, subgroup performance, language and cultural portability limits.

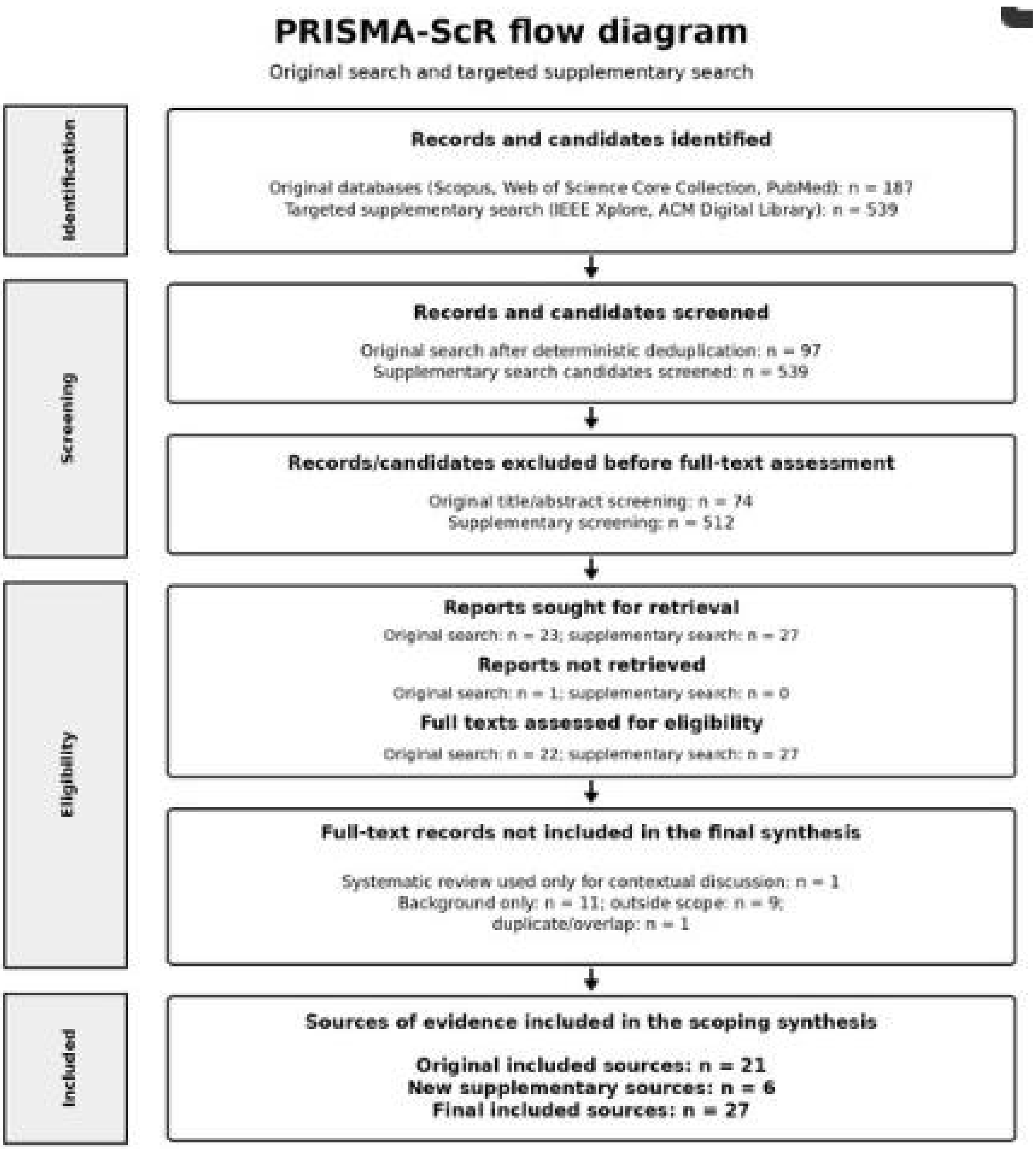

Figure 1: PRISMA-ScR flow diagram. Counts combine the original and targeted supplementary searches. ACM Digital Library results were screened in-platform; full search strategies, full-text screening decisions and final record accounting are provided in the Supplementary Materials.

For AI-supported first response, risks associated with trace exposure, compromised access, behavioural coercion, and unintended disclosure should therefore be considered fundamental design conditions rather than unusual edge cases. The illustrative threat model presented in Box 1 summarises these concerns.

Box 1. Illustrative Threat Model for AI-Enabled First Response Under Coercive Control and Restricted Digital Autonomy

Key risk conditions that should be considered as baseline possibilities rather than exceptional events

Access compromise

An abusive person may monitor a survivor's device, browser activity, shared account, call records, messages, or applications. Routine interaction traces may therefore become a direct source of danger.

Forced disclosure and retaliation

A survivor may be pressured to reveal their interaction with a digital system. Notifications, visible recommendations, referral messages, or stored conversation records may also trigger retaliation.

Model manipulation and unsafe content steering

LLM-based systems may be vulnerable to direct prompt manipulation, indirect prompt injection through copied or uploaded material, or attempts to induce the system to produce unsafe or misleading guidance.

Trace amplification through logging and analytics

Conversation histories, backend telemetry, stored metadata, and data retained for system improvement may create persistent sensitive records even where the visible user interface appears temporary or private.

Service denial and operational overload

False-positive alerts, deliberate misuse, or adversarial triggering could overwhelm human responders and reduce the availability of services for individuals experiencing genuine and immediate risk.

Governance requirements cannot remain purely conceptual when AI is incorporated into first-response pathways. Greater clarity is required regarding responsibility for risks, incident-management procedures, auditability, privacy-conscious logging, model-drift surveillance, and responses to incorrect routing, failed handover, or unsafe system output. The current evidence base is insufficient to determine how these governance arrangements should be standardised across all settings. It supports staged and carefully monitored evaluation rather than the rapid normalisation of AI within operational violence-response systems.

Strengths and Limitations

This review has several strengths. It focuses specifically on AI applications connected with first response and victim-support pathways rather than treating them as part of the much broader literature on AI in healthcare or violence prevention. The targeted searches of IEEE Xplore and the ACM Digital Library strengthened coverage of research emerging from computer science and human–computer interaction and identified additional studies involving survivor-facing chatbots, legal-support systems, and digital reporting technologies. The review also provides explicit categories for defining first-response relevance, a concise main evidence table, detailed supplementary extraction matrices, a conceptual crosswalk supporting interpretation of the M0–M3 maturity categories, and transparent reporting of supplementary full-text screening decisions.

Several limitations should also be considered. First, the review protocol was not prospectively registered, and no retrospective registration has been undertaken or is planned. Second, records were retrieved without an initial language restriction, but full-text eligibility assessment was limited to English-language publications. This may have introduced language-related selection bias. Third, the search did not include grey literature, regional databases, dedicated social work or criminology databases, or operational materials produced by crisis hotlines and anti-violence organisations. As a result, locally implemented tools, unpublished evaluations, and service-based practices—particularly from low- and middle-income settings—may be underrepresented.

Fourth, the included evidence is highly heterogeneous. It encompasses technical validation studies, usability and acceptability research, qualitative implementation studies, protocols, and conceptual pathway proposals. These forms of evidence differ substantially in purpose and evidentiary strength and should not be treated as equivalent demonstrations of effectiveness.

Fifth, the M0–M3 evidence-maturity classification was developed by the review team and applied within the same research process. It is intended to provide transparent descriptive staging rather than a validated assessment of study quality. The heuristic was not independently calibrated, no kappa statistic was calculated, and no separate inter-rater reliability evaluation was undertaken. The resulting classifications should therefore be interpreted as structured descriptions of evaluation depth rather than as an authoritative quality judgement.

Sixth, information concerning funding, commercial participation, and the developmental or product status of individual technologies was extracted where it was available but was not subjected to formal appraisal. Specific named systems, including AinoAid and other chatbot applications, are discussed descriptively and should not be interpreted as endorsed products or recommended services. Finally, the substantial concentration of evidence at the M0 level may reflect both the emerging nature of the field and publication patterns that favour technical development and feasibility studies over reports of unsuccessful implementation or negative service outcomes.

Overall, the evidence demonstrates growing activity around AI-supported intake, screening, triage, reporting, legal guidance, and service navigation within violence-response pathways. At the same time, a persistent gap remains between technological feasibility and evidence demonstrating safe, sustained, and service-integrated benefit. The author-informed considerations presented in Table 3, Table 4, Box 1, and Section 5 are intended to assist the design of future pilot and implementation studies. They are not validated framework outputs or consensus-derived recommendations. Further development should involve meaningful participation by survivors and relevant stakeholders, formal consensus-building where appropriate, and prospective testing within carefully governed real-world settings.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

CONCLUSION

This scoping review maps an emerging and highly diverse body of evidence on AI-enabled technologies for early support following sexual and gender-based violence. Current applications include conversational agents, digital support platforms, clinical screening models, social-triage systems, legal guidance tools, and privacy-oriented technologies. Across these areas, AI is primarily positioned as an aid to information provision, identification, triage, and routing rather than a replacement for professional judgement or survivor-centred care. However, the evidence remains concentrated at early stages of development, with most studies reporting technical performance, usability, acceptability, or prototype findings. Evidence of workflow integration, reliable human escalation, service uptake, survivor-centred benefit, and adverse-event monitoring is notably limited. Future research should move beyond model accuracy and examine pathway-level outcomes, including safety, equity, digital trace protection, accountability, effective handover, and sustained access to human support. Carefully governed pilot studies, developed with meaningful survivor and stakeholder involvement, are needed before AI can be considered a safe and effective component of first-response violence-support systems.

REFERENCES

1. World Health Organization. (2021). Violence against women prevalence estimates, 2018: Global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women. World Health Organization.
2. World Health Organization. (2013). Responding to intimate partner violence and sexual violence against women: WHO clinical and policy guidelines. World Health Organization.
3. World Health Organization. (2019). RESPECT women: Preventing violence against women. World Health Organization.
4. García-Moreno, C., Hegarty, K., d'Oliveira, A. F. L., Koziol-McLain, J., Colombini, M., & Feder, G. (2015). The health-systems response to violence against women. The Lancet, 385(9977), 1567–1579.
5. Campbell, R. (2008). The psychological impact of rape victims' experiences with the legal, medical, and mental health systems. American Psychologist, 63(8), 702–717.
6. Feder, G., Davies, R. A., Baird, K., Dunne, D., Eldridge, S., Griffiths, C., Gregory, A., Howell, A., Johnson, M., Ramsay, J., Rutterford, C., & Sharp, D. (2011). Identification and Referral to Improve Safety (IRIS) of women experiencing domestic violence with a primary care training and support programme: A cluster randomised controlled trial. The Lancet, 378(9805), 1788–1795.
7. O'Doherty, L., Hegarty, K., Ramsay, J., Davidson, L. L., Feder, G., & Taft, A. (2015). Screening women for intimate partner violence in healthcare settings. Cochrane Database of Systematic Reviews, 7, CD007007.
8. Hegarty, K., O'Doherty, L., Taft, A., Chondros, P., Brown, S., Valpied, J., Astbury, J., Taket, A., Gold, L., Feder, G., Gunn, J., & Brown, S. J. (2013). Screening and counselling in the primary care setting for women who have experienced intimate partner violence: A cluster randomised controlled trial. The Lancet, 382(9888), 249–258.
9. Hailes, H. P., Yu, R., Danese, A., & Fazel, S. (2019). Long-term outcomes of childhood sexual abuse: An umbrella review. The Lancet Psychiatry, 6(10), 830–839.
10. Dworkin, E. R., Menon, S. V., Bystrynski, J., & Allen, N. E. (2017). Sexual assault victimization and psychopathology: A review and meta-analysis. Clinical Psychology Review, 56, 65–81.
11. Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. International Journal of Social Research Methodology, 8(1), 19–32.
12. Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. Implementation Science, 5, 69.
13. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Lewin, S., Godfrey, C. M., Macdonald, M. T., Langlois, E. V., Soares-Weiser, K., Moriarty, J., Clifford, T., Tunçalp, Ö., & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467–473.
14. Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. JBI Evidence Synthesis, 18(10), 2119–2126.
15. Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Medical Research Methodology, 18, 143.
16. World Health Organization. (2019). WHO guideline: Recommendations on digital interventions for health system strengthening. World Health Organization.
17. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.
18. World Health Organization. (2023). Regulatory considerations on artificial intelligence for health. World Health Organization.
19. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union, L, 2024/1689.

20. National Institute for Health and Care Excellence. (2022). Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence.

21. Unsworth, H., Wolfram, V., Dillon, B., Salmon, M., Greaves, F., Liu, X., MacDonald, T., Denniston, A. K., Sounderajah, V., Ashrafian, H., Darzi, A., Ashurst, C., Holmes, C., Weller, A., & NICE Artificial Intelligence Evidence Standards Working Group. (2022). Building an evidence standards framework for artificial intelligence-enabled digital health technologies. The Lancet Digital Health, 4(4), e216–e217.

22. Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B. F., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Oakden-Rayner, L., Petersen, C., Roberts, M., Vollmer, S. J., Watkinson, P., Weber, W., Wheatstone, P., McCulloch, P., & DECIDE-AI Expert Group. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, 28(5), 924–933.

23. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26(9), 1364–1374.

24. Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nature Medicine, 26(9), 1351–1363.

25. Vasey, B., Ursprung, S., Beddoe, B., Liao, H., Nagendran, M., Chen, Y., & McCulloch, P. (2022). Association of clinician diagnostic performance with machine learning-based decision support systems: A systematic review. JAMA Network Open, 5(3), e221290.

26. Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., A'Court, C., Hinder, S., Fahy, N., Procter, R., & Shaw, S. (2017). Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. Journal of Medical Internet Research, 19(11), e367.

27. Damschroder, L. J., Reardon, C. M., Widerquist, M. A. O., & Lowery, J. (2022). The updated Consolidated Framework for Implementation Research based on user feedback. Implementation Science, 17, 75.

28. Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. Administration and Policy in Mental Health and Mental Health Services Research, 38, 65–76.

29. Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. BMJ, 374, n2061.

30. Greenhalgh, T., Abimbola, S., & the NASSS-CAT Project Team. (2019). The NASSS framework: A synthesis of multiple theories of technology implementation. Implementation Science, 14, 18.

31. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

32. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR.

33. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

34. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). Association for Computing Machinery.

35. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504.

36. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501–507.

37. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28, 689–707.

38. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399.

39. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20, 310.

40. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25, 1337–1340.

41. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

42. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.

43. Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260, 113172.

44. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2, 305–311.

45. Rieke, N., Hancox, J., Li, W., Luengo-Oroz, M., Li, M., Trotz, D., & others. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119.

46. Cohen, I. G., Mello, M. M., & Jagsi, R. (2024). Artificial intelligence, medical malpractice, and liability. JAMA, 331(2), 121–122.

47. World Health Organization. (2021). Global status report on preventing violence against children 2020. World Health Organization.

48. United Nations Office on Drugs and Crime. (2022). Gender-related killings of women and girls (femicide/ feminicide): Global estimates of gender-related killings of women and girls in the private sphere in 2021. United Nations Office on Drugs and Crime.

49. United Nations Women. (2020). Handbook for addressing violence against women in emergencies. United Nations Women.

50. United Nations Population Fund. (2015). Minimum standards for prevention and response to gender-based violence in emergencies. United Nations Population Fund.

51. World Health Organization. (2022). Addressing violence against women in health and multisectoral policies: A global review. World Health Organization.

52. Emezue, C. C. (2020). Digital or digitally delivered responses to domestic and intimate partner violence during COVID-19. JMIR Public Health and Surveillance, 6(3), e19831.

53. Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2017). Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 46.

54. Matthews, T., O'Leary, K., Turner, A., Sleeper, M., Woelfer, J. P., Shelton, M., Manthorne, C., Churchill, E., & Consolvo, S. (2017). Stories from survivors: Privacy & security practices when coping with intimate partner violence. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 2189–2201). Association for Computing Machinery.

55. Havron, S., Freed, D., Chatterjee, R., McCoy, D., Dell, N., & Ristenpart, T. (2019). Clinical computer security for victims of intimate partner violence. In Proceedings of the 28th USENIX Security Symposium (pp. 105–122). USENIX Association.

56. Slupska, J., & Tanczer, L. M. (2021). Threat modeling intimate partner violence: Tech abuse as a form of cyber violence against women. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 527–536). Association for Computing Machinery.

57. Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., Levy, K., Dell, N., McCoy, D., & Ristenpart, T. (2018). The spyware used in intimate partner violence. In Proceedings of the 2018 IEEE Symposium on Security and Privacy (pp. 441–458). IEEE.

58. Park, J. S., Sun, F., Geyer, W., & others. (2024). Generative AI and the future of human-centered support systems: Risks, safeguards, and opportunities. ACM Computing Surveys, 56, 1–29.

59. Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41, 105567.

60. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68). Association for Computing Machinery.

61. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

62. Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33, 659–684.

63. Birhane, A. (2021). The unseen Black faces of AI algorithms. Nature, 590, 327–328.

64. Bailo, P., Carsana, C., Garreffa, M., Carannante, A., Giustini, M., Fazio, C., Falzano, L., Piccinini, A., & Gaudi, S. (2026). AI-enabled first-response support after sexual and gender-based violence: A PRISMA-ScR scoping review. Healthcare, 14(14), 2174.

Citation: Laskay S, Walsh P, Larsen C. "From Disclosure to Support: A Scoping Review of AI-Enabled First-Contact Responses to Sexual and Gender-Based Violence". American Research Journal of Humanities and Social Sciences, Vol 11, no. 3, 2026, pp. 117-126.