# USING AI RESPONSIBLY

## for Research on Violence Against Women

Written by Isabelle Amazon-Brown, Quito Tsui and Linda Raftree, The MERL Tech Initiative (MTI) and Elizabeth Dartnall, Sexual Violence Research Initiative (SVRI)

March 2025

Illustrations by Vectorjuice - www.freepix.com

# Foreword

Generative AI (GenAI) is evolving rapidly, and its use in violence against women research remains new and largely experimental. While researchers are beginning to explore GenAI's potential, uncertainty around its risks, ethical challenges, and real-world application persists. Recognising the need for clear guidance, the Sexual Violence Research Initiative (SVRI) partnered with The MERL Tech Initiative (MTI) to develop this resource. This guide aims to provide background on GenAI, strengthen researchers' AI literacy, and offer practical advice to help the violence against women research community critically and carefully engage with GenAI while minimising risks.

We start off with an overview of GenAI and potential areas where researchers might use it at different stages of the research process. We highlight some low-risk, practical applications of GenAI that can help streamline tasks while ensuring that researchers remain in control. (Appendices provide clear definitions of key AI terms and a simple explanation of how generative AI works for those less familiar with the technology.)

Following these examples, we discuss some of the key risks and potential harms of using GenAI for research on violence against women, including reliability, bias, dehumanisation, data privacy, power and patriarchy, relevance, transparency and replicability, inclusion and participation, worker's rights, inequities and environmental harms.

To support informed decision-making, we offer a simple decision framework with critical questions to consider before using AI for any task:

- Does AI provide a clear benefit over existing methods?
- Does its use comply with ethical guidelines, survivor consent, and data protection laws?
- Are humans in the loop to validate results and take responsibility?
- Can potential risks be effectively managed?
- Is there a process in place for making and documenting decisions about the use of GenAI?

If the answer to any of these is "no," the guidance suggests reconsidering or avoiding AI in that context. This framework can help researchers to weigh AI's advantages against its limitations, ensuring thoughtful and responsible use.

Recognising AI's fast-changing nature, we encourage a cautious approach to GenAI. We suggest starting small—testing GenAI tools on non-sensitive, publicly available data—to better understand their capabilities and limitations. If you choose to integrate GenAI into your research, we encourage you to document and share your experiences, so we can learn collectively and refine best practices together.

The guide reinforces a key message: human expertise is irreplaceable. AI can improve efficiency, but final analysis and interpretation should always lie with researchers.
By experimenting in a controlled way and keeping ethical principles at the forefront, researchers can build confidence in AI while avoiding unintended harm.

This guidance is intended to be a living document—SVRI will maintain and update it as the AI landscape evolves and new insights emerge. SVRI and MTI are grateful for the critical feedback received from colleagues during the production of this guide, which has helped strengthen its content. (See Appendix B for details on the methodology used.)

Please let us know how you find the guidance via this online form - we welcome ongoing feedback and learning, as collective experience will shape future updates to ensure the guidance remains relevant and useful.

**Elizabeth Dartnall, SVRI Executive Director & Linda Raftree, Founder, The MERL Tech Initiative**

Suggested citation:
Amazon-Brown, I., Tsui, Q., Raftree, L & Dartnall, E (2025)
Using AI Responsibly for Research on Violence Against Women.
The MERL Tech Initiative and Sexual Violence Research Initiative.

# Table of contents

# 1 Introduction

Artificial Intelligence (AI) is an overarching term that describes the use of computer programmes or machines to perform tasks that would typically require human intelligence. These tasks include learning (AI can be programmed to analyse text and make decisions based on that analysis), problem-solving (based on automated analyses of past decisions, AI can be trained to make decisions on its own), and language understanding (AI can be fed huge sets of text or audiovisual data and programmed to create new content based on those data).

AI is used as a 'catch-all' term, but can encompass any of the following:

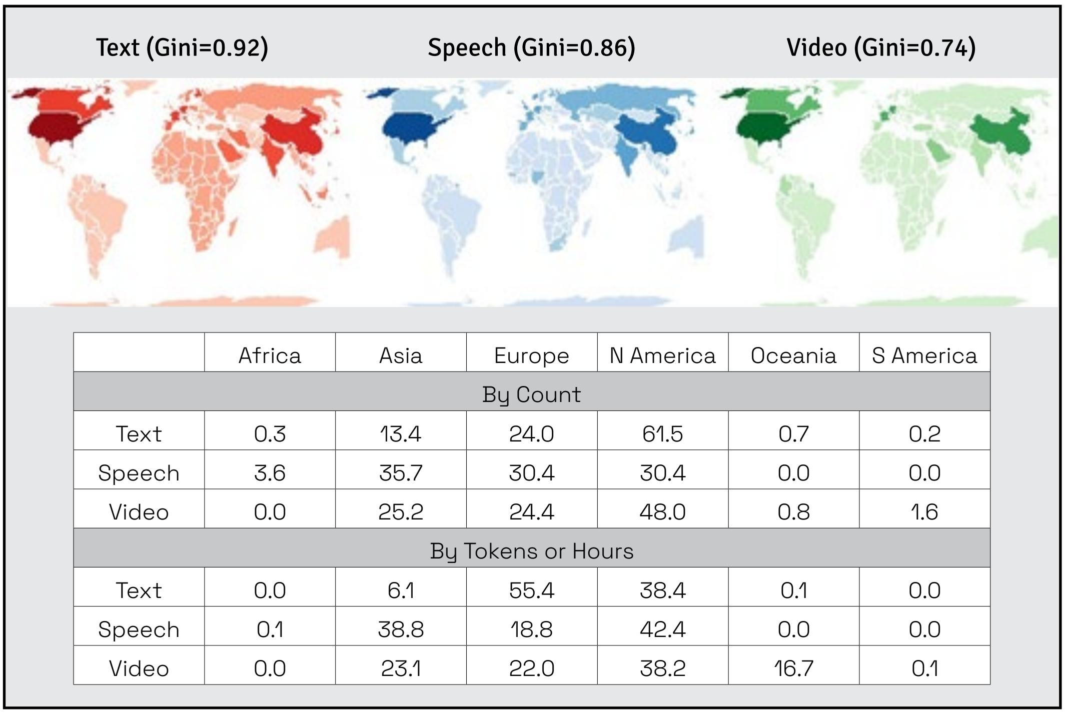
- When you ask ChatGPT to explain the root causes of violence against women in a particular location.
- When Google analyses your past search habits to recommend results.
- When you use spellcheck, Grammarly or DeepL to assist in editing a paper.
- When you use Google maps and follow the traffic-optimized route during a research trip.
- When a research participant answers a survey via a WhatsApp chatbot.

The newest type of AI, which emerged in late 2022, is referred to as Generative AI or "GenAI". It has arisen due to advances in the field of "Natural Language Processing" or "NLP", which trains computers to process vast quantities of data and rapidly make predictions about the most likely next word in a sentence. When given an instruction or asked a question – often referred to as a 'prompt' - GenAI tools can generate text, images, audio and video that look and sound like they were produced by a human. (See Appendix A for a list of key definitions related to AI and GenAI).

## Challenges at the core of GenAI

The hype cycle driven by AI companies and GenAI enthusiasts can make us feel like we are behind the curve. However, while GenAI has a reputation to optimize processes and tasks, its real utility is still unproven in many fields. If using it for research, we need to be aware of both practical and ethical challenges. This is especially salient in the context of violence against women where researchers are dealing with highly sensitive information.

On the practical side, GenAI is prone to error, and it completely depends on the information it can scrape from the internet, meaning that it is highly biased. Most of the content that GenAI draws from is produced by the USA and China. A 2024 paper found that less than 0.2% of the data used to train GenAI models came from Africa or South America.



Text (Gini=0.92) Speech (Gini=0.86) Video (Gini=0.74)

| | Africa | Asia | Europe | N America | Oceania | S America |
|---|---|---|---|---|---|---|
| By Count | | | | | | |
| Text | 0.3 | 13.4 | 24.0 | 61.5 | 0.7 | 0.2 |
| Speech | 3.6 | 35.7 | 30.4 | 30.4 | 0.0 | 0.0 |
| Video | 0.0 | 25.2 | 24.4 | 48.0 | 0.8 | 1.6 |
| By Tokens or Hours | | | | | | |
| Text | 0.0 | 6.1 | 55.4 | 38.4 | 0.1 | 0.0 |
| Speech | 0.1 | 38.8 | 18.8 | 42.4 | 0.0 | 0.0 |
| Video | 0.0 | 23.1 | 22.0 | 38.2 | 16.7 | 0.1 |

Figure 1: The geographical distribution of countries (world maps) and continents (table) represented by dataset creators. Despite some differences in European, Russian, and Middle Eastern representation, creators are heavily concentrated in the US, China, and Western Europe, with little to no representation in South America or Africa, across modalities. The current Gini coefficient for (Text, Speech, Video) = (0.92, 0.86, 0.74), where higher values indicate more concentration. (Source)

The Afrofeminist organisation Pollicy, speaks out boldly on these biases noting that "African women with intersecting identities stand to experience multiple levels of AI bias" (p 17). Pollicy highlights how a person's sex, gender, race, gender identity, sexual orientation, class, disability status and other identity markers directly influence their experiences with AI driven technologies. AI is more likely to produce incorrect, biased or stereotyped results because it is drawing from an incomplete pool of data. For example, AI will exclude trans people and reinforce gender normativity if private data on sexual orientation and/or gender identity is missing from a training dataset it uses. Pollicy flags that most AI tools are "designed by white men from the Global North, they are rarely assessed for their effects on women's privacy and safety." What's more, without commitment from AI developers to regular AI human rights audits and assessment of the impact of AI models on certain groups, there is a high risk of harm. Even if the issues of data bias are solved, without guardrails, AI can be trained to do harmful things more efficiently.

The biases and errors that are baked into AI models affect the quality of GenAI outputs. This practical challenge (is the GenAI output factually correct and culturally appropriate?) is also an ethical challenge (should we use this kind of biased or inaccurate data in our research?). Other big picture ethical questions plague GenAI, including plagiarism. Most of the art, literature, writing, video, and other data used to train GenAI were scraped from the internet without the consent of creators or by exploiting outdated fair use policies. Labour rights are also an issue, with data workers earning low salaries and/or experiencing exploitative labour conditions. Lastly, increasing amounts of water and energy are being used to train AI at giant data centres around the world.

## The need for critical literacy around GenAI

Despite all these ethical challenges, researchers across a range of sectors are exploring whether it's possible to use GenAI at different stages of the research process. In Section 2 of this guide, we share some of the ways that GenAI is being or could be used by researchers studying violence against women, along with several caveats.

**The hope is that this guide can help researchers of sensitive topics navigate the practical and ethical challenges with GenAI and enable evidence builders to make informed decisions about its use.**

Some of the most cited benefits of GenAI include saving time and money by automating repetitive tasks and enhancing efficiency during data cleaning and processing. While there are still many challenges with accuracy, researchers are using GenAI to translate research products and materials into multiple languages and to help simplify text to make it more comprehensible to research participants and the wider public. While some researchers we spoke with in the process of developing this guide expressed deep concerns about plagiarism, many researchers from majority world countries felt GenAI was useful for leveling the playing field in a profession dominated by English speakers. They felt it could increase their access to publication in English language peer reviewed academic journals, making their writing more likely to be published.

As we experience the rapid expansion of GenAI, it's vital that we develop our critical literacy so that we can make informed choices about the use of GenAI.

## Guidance on using GenAI for research on violence against women

Because some researchers are exploring GenAI already and others would like to but fear they will cause harm by doing so, the Sexual Violence Research Initiative (SVRI) and The MERL Tech Initiative (MTI) have collaborated on developing this guide exploring the use of GenAI for research violence against women.

In this guide we cover the use of GenAI tools for supporting:

- Research proposal development
- Literature review and evidence synthesis
- Data collection
- Data cleaning and analysis
- Synthesis and report writing

The guide is intended to support those who are interested in, and maybe worried by, how GenAI is affecting research on violence against women. It aims to identify low risk ways GenAI could be used as well as to flag areas where GenAI should be entirely avoided. The guide is pitched at a 'beginner' level, but we hope those already using GenAI as a work tool will also find it useful. We also aim to reinforce existing feminist research principles and practice such as do no harm, survivor centered approaches, informed consent, transparency, participatory and rights-based approaches, and the advancement of gender equality. The UNFPA's data-specific principles can also guide how we think about the use of GenAI (See Figure 2).

- Do no harm
- Survivor-centered approach
- Informed consent and transparency
- Participatory approaches
- Rights-based approach
- Advance gender equality

UNFPA’s Data-specific principles:

- Safety by design
- Purpose limitation
- Data minimisation
- Proper use of data
- Fairness
- Informed consent, transparency, ownership
- Accuracy and data quality
- Security: integrity, confidentiality, availability
- Accountability
- Unconditional service

Figure 2: GBV Core Principles and Data Specific Principles

**❗Many GenAI-powered applications and tools are misaligned with these key data principles. Therefore, our use of it must be limited to non-personal, non-sensitive data and we must carefully review GenAI outputs even when we apply GenAI to publicly available data.**

## How this guide was developed

This guide is intended as a first iteration of an accessible but comprehensive introduction to the use of GenAI in research on violence against women, drawing on a brief literature review, supported by qualitative and quantitative data collected from those actively involved in advancing knowledge on violence against women. In order to release this guide in a short timeframe in response to the rapid pace of developments of this field, the data collected were necessarily non-exhaustive and we see potential for a more rigorous study. Similarly, given the cutting-edge nature of using the latest forms of AI, the literature review reflects the paucity of evidence and examples available currently. The guidance offers good practice relevant as of January 2025. It will be updated by the SVRI on a regular basis. (See Appendix B and Appendix C for more details on the methodology and scope.)

While developing this guide, researchers and practitioners across the globe were consulted to get a sense of their key questions and concerns. On one hand, researchers are feeling excitement at new opportunities for efficiency, accessibility and creativity; on the other, they fear being ‘left behind’, being ‘out of the loop’, and causing new forms of harm (or reinforcing existing ones). Research participants were uncertain about the rules and ethics for when and how GenAI can and should be used. They were deeply worried that the data practices inherent to GenAI might reverse decades of painstaking work around  localisation, and feminist and trauma-informed research practices. In addition, there are deep concerns about the use of AI tools to enable and scale violence against women and non-consensual intimate imagery, as we note below.

### Tech-Facilitated Gender Based Violence (TFGBV) and the Impact of GenAI

This guidance is designed for those who conduct research and may want to use AI powered tools to do so. An overview of the different forms of TFGBV and specifically how GenAI is being used to exercise and exacerbate this violence is beyond the scope of this document. However, the practitioners and researchers we spoke to were deeply concerned about this phenomenon.

Their biggest worries related to the creation and weaponisation of AI generated Non-Consensual Intimate Imagery (NCII) or 'deep fakes'. Marginalised and under researched groups, such as LGBTQI individuals as well as women in politics, women's rights defenders, and activists were at a high risk of being attacked using AI-generated imagery. In keeping with the evidence on the impacts of many forms of TFGBV (trolling, stalking, online abuse and doxxing for instance) there is a fear that this phenomenon could lead to a widening of the gender digital divide, with women and girls' freedom to use online spaces being further restricted (either by themselves or by others who want to protect them). This is especially the case in already restrictive environments, for example within highly religious communities, where parents and elders don't understand that images can be faked to such a high degree of accuracy. In these contexts, the generation of fake images can result in the closing of digital spaces for women and girls and can also have extremely serious offline consequences.

It is useful for researchers to be aware of various forms of AI-facilitated gender-based violence (which can be categorized as fast-developing types of TFGBV), as well as the emerging questions and evidence-gaps relating to its prevalence as they explore their own usage of AI as a research tool.

Some researchers we spoke with pushed back on GenAI hype. For example, one person asked:

> "If AI can make mistakes, why are we supposed to use [it]?
> At best, it's wrong, at worst, it's horrible"

(SVRI Forum 2024 conference participant)

This document is a response to some of these questions, aiming to help researchers navigate the potential of GenAI as well as the areas where GenAI's utility is outweighed by its potential for harm. We hope the guide helps researchers enhance their critical literacy, develop a stance on GenAI and feel more equipped with practical advice. The guide includes:

- An overview of the many ways AI could be used in research on violence against women.
- A reality-check on how AI and GenAI is currently being used by researchers.
- The risks of using GenAI in research on violence against women and ways to mitigate them, and how to decide when GenAI should not be used.
- Tips on how to safely explore the use of GenAI.
- Annexes with key definitions and details on how GenAI works.

We encourage all readers to share their feedback with the SVRI using this feedback form.

# 2 How GenAI could be applied in research on violence against women

Despite the risks and ethical concerns with GenAI (which we cover in Section 3), there is a lot of excitement about how GenAI could possibly be applied to research. This reflects the broader hype coming from AI companies, the media, and AI enthusiasts - some of whom are genuinely excited about what the tools can do, and others who are primarily seeking financial gain and market dominance.

In practice, GenAI capabilities are still limited. GenAI use for research on violence against women is often informal and ad hoc rather than a vetted part of research methodologies or data processing. Though there are examples of GenAI use in research more broadly, our rapid desk research and in person and virtual interviews point to light touch, often non-research related, uses of GenAI tools as being the most common. Interviewees cited help with writing grant proposals, drafting e-mails, copy-editing and translation as examples of how they currently use GenAI. Several noted other possible uses, such as in evidence reviews and synthesis and legal analysis but were either not using GenAI for this, or had only begun experimenting with it.

This is an important gap to be aware of. Even though it may feel as though everyone is talking about GenAI and its possibilities, very few of these potential applications have actually been tested, let alone systematically evaluated for quality. There remains a gap in our knowledge about the reliability and accuracy of using GenAI for violence against women research. In our research we saw evidence of the use of AI tools such as natural language processing (NLP) and machine learning (ML), but minimal evidence of sustained and repeated use of GenAI. (See Appendix A for a more detailed explanation of NLP and ML and Appendix D for examples of how they are being used for research on violence against women).

It should also be noted that institutional responses to student use of AI vary—while some discourage or penalise its use, others are exploring ways to integrate AI tools more constructively into learning and research. During our own research for this paper, students and young researchers told us that they often choose to keep quiet about using AI in their work, for fear of the consequences. Most journals have developed guidelines around where AI and GenAI can be used and the consequences of using it for writing and for peer reviewing articles. Elsevier notes that their policies are primarily aimed at use of AI during the writing process, and "not to the use of AI tools to analyse and draw insights from data as part of the research process."

In this section we will outline some of the ways that GenAI is being used during the various phases of a research process, including research design and proposal development; literature review and synthesis; data collection; data cleaning and analysis; and writing, producing and disseminating research results.

Across these phases the extent of GenAI use can range from researchers using GenAI as a sense check, or to complete small, isolated tasks, to using GenAI for more complex, multi-step processes. For instance, GenAI could be used to review a proposal outline, write a proposal outline, or even write an entire research report (whether these are appropriate or ethical uses of GenAI is another matter).  Awareness of this spectrum of uses can be helpful when considering the ways GenAI can be incorporated within violence against women research and where to be especially careful because of practical and ethical challenges.

❗ **It is important to note that this section is not endorsing the outlined potential applications of AI. Rather this section outlines the proposed uses of GenAI in violence against women research to provide readers with an informed and more detailed basis for the discussion on risks in Section 3 of this report.**

❗ **Don't forget that even for simple, straightforward tasks, GenAI can go wrong and needs constant checking, questioning, and interrogation, especially if contemplating its use in sensitive contexts or with vulnerable people and groups. See Section 3 for more information on risks of GenAI and Section 4 on getting started with GenAI for additional guidance.**

### The Spectrum of GenAI Use and Levels of Human Oversight

It is helpful to consider GenAI use on a spectrum, from more to less human input. Human oversight (sometimes called keeping a 'human in the loop') will be important at different points in the process, depending on the task and the context.

- On one end of the spectrum, we have situations where researchers are supported by GenAI. GenAI can be used as part of an idea generation process where a researcher engages with GenAI tools for discrete tasks that the researcher can then integrate and build on. For example, a researcher might ask GenAI to help them brainstorm an outline for a blog post to help with writer's block or to give them ideas for a participatory workshop activity that they then flesh out and adapt for their context.

- GenAI tools can also be used for review, where the bulk of the content or source material has been produced by a researcher. GenAI is then used to sense check: perhaps to copy-edit, adjust tone (e.g. make the language more or less formal), or compare against other similar outputs – for example, a peer review publication.

- Towards the other end of the spectrum, where researcher input is minimised, GenAI can be used for formulation: creating outputs from scratch with minimal input and direction at the early phase, with researchers reviewing the final output for accuracy and bias.

- One step further than this is automated AI workflows where AI tools are programmed by humans to undertake certain repetitive tasks. Once automated workflows are tested and evaluated and shown to be able to achieve the task with sufficient accuracy, they can be left to run on their own with minimal ongoing human review (though periodic spot checking is a vital part of this process).

It is important to remember that GenAI tools can be used across this spectrum. In the discussion below we demonstrate the range of ways in which GenAI could be used with different degrees of researcher input.

## Phase 1 - Research proposal development

A strong research proposal requires a solid methodological foundation or framework. Whilst caution and ethical consideration are essential when using GenAI in research proposal development, many researchers are already leveraging GenAI for this purpose. The table below outlines a number of ways GenAI can be used at this phase of the research process.

It should be noted that many grant makers and research institutions are developing policies around the use of AI in proposals. Some (but not all) will reject proposals that they believe have used GenAI in their development. As this is relatively new territory, we can expect these policies to shift over the next couple of years. It is currently unclear whether these policies will become more or less strict. In any case, it is important to find out if grant makers and research institutes have a policy on GenAI before using it to draft a proposal.

| Category | Possible uses of GenAI in research proposal development |
|---|---|
| **Research questions** | • Using GenAI as a thought partner when developing research questions and asking it to help identify gaps in your questions. E.g. "I'm doing research on x with y population. It's important for me to understand. I've got a list of questions here [add questions]. What am I missing?"<br>• As a support for hypothesis formulation. |
| **Research approach** | • Advice on or help to refine your selected research methodology.<br>• Identification of different research approaches and suggesting methodology based on the project description, research questions or desired outputs.<br>• As a thought partner - poking holes in your approach or comment on the approach using a specific persona. E.g., "You are a feminist research advisor from x organisation. Please review my approach and identify weaknesses or viewpoints that I've neglected to consider. What would your main critiques of my approach be?" |
| **Research plan** | • Checking your research plan for inconsistencies or gaps.<br>• Helping revise your research framework based on other inputs or prompts. This could include recommendations on data collection sources and methods. E.g. "I want to find academic articles from journals with an impact score higher than 'X'. I have already reviewed the following journals. Can you recommend, with links, other academic sources I can review?"<br>• Developing a specific research framework based on your project, including questions and indicators for an analytical framework. |
| **Intervention studies** | • Helping you to develop a clear theory of change for your study or to review your theory of change and offer suggestions for improving it. |

## Phase 2 - Literature review and evidence synthesis

Reviewing literature and summarising the existing discussion in a literature review is often an especially arduous part of research. Systematic reviews in particular, or meta-analysis, where the field is surveyed for specific topics, are often important for research on violence against women, but are time intensive. Several GenAI tools exist that offer different types of support, including evidence selection, summarisation, synthesis and writing. However, many researchers have deep concerns about allowing GenAI tools to autonomously select and/or summarise research papers due to their tendency to hallucinate (make up findings) and / or fabricate citations. Thus, allowing AI to do this work can bring in unseen biases due to how algorithms work. Further, without a solid understanding of the literature, and solid review of the GenAIs outputs, slight shifts or misinterpretations in GenAI outputs can go unnoticed which can significantly alter the research insights and subsequent narrative.

For example, this fabrication of findings and citations resulting in incorrect or incomplete outputs has been an issue for legal professionals, who have been called out in court for citing fabricated legal cases after using GenAI tools in their work. Bespoke tools specifically developed for literature reviews and evidence synthesis may do a better job than generic tools. They can be designed to draw from specific literature and be adjusted to achieve greater accuracy. That being said, these tools are in their infancy, not fool proof and the outputs generated must be reviewed.

Thus, if using AI for literature reviews and evidence synthesis, it is extremely important to have subject matter experts review all outputs for errors and inaccuracies, as recommended by the World Bank's Independent Evaluation Group. Emerging tools like Google's Notebook LM and others offer a way for more human oversight. This tool allows researchers to upload their own sources (including PDF files, websites, videos, documents and slides made on Google platforms) and ask it to provide summaries and to identify connections between topics. The tool provides citations and quotes that can be checked and verified.

Another potential downside when conducting research using GenAI-powered tools is that search algorithms carry their own biases, based on who designs them and the users own behaviours and clicks. Google Scholar's results are based on your own search history, for example. This can mean that researchers might be limited to seeing search results that confirm their own biases. This brings to light another way that GenAI tools incorporate bias into research. In their article on PubMed's Best Matchsorting algorithm, Kiester and Turp (2020) discuss the lack of transparency around how search functions work and how they make decisions about what links and articles are shown to the person doing the searching. (This is not a new problem, as algorithmic search has been the norm for quite some time already).

In early 2025, GenAI companies began releasing new tools with names such as Deep Research (both Open AI and Google have a deep research tool) and DeepSeek. Deep Research allows you to generate a research report by browsing the internet. While these tools might be useful for some types of research or summarisation of web content that is not behind a paywall, in-depth examination of the quality of output of these tools reveals an issue with low quality citations. The tool finds sources "based on what it thinks can confirm its arguments rather than making sure the source material is valid or respectable."

As the field of GenAI advances and more research tools are developed, some of these challenges are being addressed. Over the past two years, for example, GenAI tools that offer more transparency and traceability in citations are emerging. Other tactics such as asking a tool to verify its work or asking one AI tool to review the work of a different AI tool can sometimes help uncover mistakes. As GenAI is still emerging, it remains to be seen whether these issues will be more adequately addressed in new tools or as existing tools are further developed. It's always important to have a subject matter expert review GenAI produced literature reviews and summaries.

| Category | Possible uses of GenAI for literature review and evidence summaries |
|---|---|
| Identifying literature | • GenAI tools such as Elicit or LitMap to suggest scholarly journals and academic databases.<br>• Google Scholar's AI-powered PDF reader.<br>• Assistance with selecting relevant data and proposing literature related to papers or themes that you have shared.<br>• Gathering relevant literature based on research questions or key themes.<br>• Suggesting search terms to help refine results on databases and search engines. |
| Synthesising literature | • Summarising individual articles and identifying key themes.<br>• Synthesising across several articles to produce a review of related literature.<br>• Creating podcast summaries of articles using Notebook LM. |
| Helping query literature | • Creating a custom chatbot or 'GPT' that enables you to query a public report through questions and answers so that you can extract and then cross-check information. |

## Phase 3 - Data collection

Data collection in violence against women research features a wide set of stakeholders and methodologies and the data are often highly sensitive. Ensuring ethical practices, particularly around confidentiality and data protection, is essential to safeguarding participants and maintaining research integrity.  GenAI tools can be used to support certain data collection-related activities, yet researchers must exercise caution to ensure data privacy and security are not compromised. As with any use of GenAI, it's important to check for inaccuracies and ensure there is human review and oversight when using GenAI to support data collection efforts. All research on violence against women involving human participants requires ethical approval. If the ethical review board lacks expertise in GenAI, both the research team and the ethical review board are responsible for seeking guidance, consulting experts, and building the necessary knowledge to assess the ethical and legal implications of using GenAI for collecting or processing personal or sensitive data.

Paid, enterprise level tools (such as a business or team subscription to Microsoft's CoPilot) and bespoke GenAI tools (such as those built internally) are generally more private and safe to use than free tools (such as a free ChatGPT or Claude accounts). There are still major concerns about the privacy and security of data that are uploaded or linked to these tools, meaning that their use for data collection on violence against women is highly risky. In many cases the data put into a GenAI tool are used without knowledge or informed consent to train private language models. These concerns have increased due to a number of Big Tech and Big AI companies aligning closely with governments that are known to be actively working to restrict women's rights, gender equity, sexual and reproductive health rights, and LGBTQI rights and protections.

**❗ Personal and sensitive data should never be uploaded into GenAI tools and applications unless there is a strong, written, legal guarantee that it will remain private and secure and there is active and informed consent from research participants.**

**❗ Researchers and ethical review board members must consult with experts on the data and privacy risks of using GenAI in data collection and weigh risks and benefits of doing so.**

Some transcription tools are available that can run locally on an individual device and do not upload any data to the GenAI or the cloud, and these are generally more protective of privacy. Some organisations consider them safe to use for interview transcriptions. Accuracy is a challenge with automated transcription and translation, however, especially in languages other than English, because most GenAI is trained primarily on English data sets. Transcription and translation tools may also miss context-specific and cultural nuances in any language, and could fail to pick up on survivor distress cues.

GenAI notetakers are another form of transcription tool that is becoming common to see in meetings and used for key informant interviews. These should be used with extreme caution as they often include sensitive information (names, addresses, emotions of participants, or in some cases screenshots of video feeds during a call which may expose the faces of participants) which are not appropriate or helpful for the research being conducted. They often lack clear terms and conditions and privacy policies, making it hard to know what is done with the data they ingest. Some GenAI notetakers generate automated transcripts that are shared with all meeting invitees or stored insecurely in the cloud. Not all of them are encrypted, and it's difficult to know which of them share data with third parties.

❗ **Meeting hosts must carefully weigh the risks and benefits of using GenAI notetakers, particularly considering potential privacy concerns.**

Finally, some transcription tools have been found to fabricate text. One 2024 study looked at transcriptions done in a hospital using a tool called Whisper. While the transcriptions were generally highly accurate, "roughly 1% of audio transcriptions contained entirely hallucinated phrases or sentences which did not exist in any form in the underlying audio." This problem was more common among patients with irregular speech patterns caused by a head injury or stroke. In general, translation and transcription tools still require significant improvements before they can be used ethically and effectively in specialised fields like violence against women, particularly as some languages lack direct terms for sex or sexual violence, often relying on euphemisms or indirect language.

With caution and attention to potential risks, there are some ways that GenAI can support researchers with data collection, as noted below. Less risky uses are those that support the wider data collection preparation process, not the actual collection of data itself. Subject matter experts should be involved to review any GenAI created materials and to check for bias, fabrications, and to ensure cultural nuances are considered.

| Category | Possible uses of GenAI for data collection |
|---|---|
| Interviews | • Helping to create interview guides, or build and strengthen existing guides based on your previous guides. |
| Surveys | • Helping to identify existing surveys that can be adapted to your setting or context.<br>• Supporting survey design, formatting and generation of a final survey document.<br>• Putting a tested, validated online or mobile survey that does not collect personal or sensitive data into a chatbot format on WhatsApp to make it easier, more accessible, and more friendly for people to fill out.<br>• Helping support survey designers who are using an online platform or software to manage the platform, troubleshoot, improve survey design, and review a survey for errors or gaps, based on best-practices. (e.g. The Survey CTO Assistant helps survey designers troubleshoot the use of the SurveyCTO platform, improve research design, and hone survey questions) |
| Translation and transcription | • Translating stakeholder interview questions, interviews, and surveys in advance of an interview or following data collection (if done with a tool that has been reviewed for safety and privacy and if the translation is reviewed closely and checked for accuracy).<br>• Transcription assistance to increase accessibility of interviews, generate notes, and potentially form the basis of insights, again, with caution considering the challenges with nuance and accuracy. |

## Phase 4 - Data cleaning and analysis

GenAI tools can be used to help researchers streamline the process of analysis and draw insights from multiple data sources and/or large datasets. However, as with any other uses of GenAI, there are many precautions to take.

Ethical concerns include reinforcing biases, generating misleading or inaccurate findings, risks to data privacy and confidentiality, lack of informed consent, and the opacity of AI decision-making. For example, most GenAI models are trained primarily in English, with a strong bias toward U.S. culture. As a result, data analysis using off-the-shelf GenAI applications is often less accurate for non-English languages and diverse cultural contexts. While new GenAI models are being developed in other languages and regions, helping to bridge this gap, there is still a long way to go to achieve language equity and reduce biases—such as those related to gender and race—that persist in GenAI systems.

In their paper on large language model applications for evaluation, Head et. al. (2024), identified the following types of bias inherent in natural language process and AI, including GenAI:

- Historical bias is a reflection of society's existing biases, such as stereotypes related to race, disability, or gender, embedded into a dataset upstream and then replicating them in the applications that are built on top of the language model.
- Representation bias results from under-, over- or mis-representation of particular groups in the dataset, such as English speakers or internet users; leading to skewed GenAI content and outputs.

- Semantic bias happens when biases that are embedded in language get reproduced across applications and uses and thus perpetuated and scaled.
- Label bias is introduced when the data workers who train GenAI models bring in their own personal and/or cultural biases in how they label and categorize data
- Algorithmic bias is the bias introduced by those creating and programming GenAI models and the rules and weights for their decisions.

There is limited evidence of GenAI being used for data cleaning and analysis in research on violence against women. However, the use of more traditional AI tools—such as machine learning (ML) and natural language processing (NLP)[1]—suggests that GenAI could be introduced to help analyse unstructured, qualitative data and potentially expand the use of existing AI tools in this field. NLP and ML tools have been used in violence against women research to identify patterns of intimate partner violence and flag individuals who may be at risk or have experienced abuse.[2]

Below are some ideas on how GenAI could be used for data cleaning and analysis. We expect this to change and grow quickly if safer, more accurate, more inclusive models emerge. On the other hand, if AI safety, privacy and accuracy don't improve, GenAI will not be an appropriate tool for data analysis.

[1] See Appendix A for more in-depth definitions of ML and NLP

[2] See Appendix D for examples of the use of NLP and ML for violence against women research.

| Category | Possible uses of GenAI for data cleaning and analysis |
|---|---|
| Processing largescale data sets | • Comparing multiple texts and identifying differences between texts and conducting sentiment analysis.<br>• Analysing individual texts and, extracting information on actors, places, and relations.<br>• Text classification, e.g., conducting qualitative thematic analysis with both deductive and inductive coding, as well as sentiment analysis.<br>• Causal relationship extraction to identify and synthesise causal factors.<br>• Analysing qualitative research data to identify patterns in unstructured data (though a caution is needed as it's difficult to be sure these data do not include personal or sensitive data).<br>• Analysing across different data sources and inputs e.g. medical documentation, social media online interactions, etc.<br>• Surfacing interdisciplinary connections that might not be immediately obvious. E.g., a legal researcher studying gender-based violence laws could use AI to identify psychological research on survivor-centred justice approaches, helping cross-pollinate ideas across disciplines.<br>• Using AI agents to query documents or interviews allowing researchers to ask questions of the data they have gathered.<br>Read more on this here and here. |
| Coding applications | • Codebook generation.<br>• Coding and tagging of research and interview excerpts, see Impact Mapper for one example.<br>• Assistance with statistical analyses; e.g. reviewing and producing code for statistical software. |
| Insights generation | • Identifying key themes across research inputs.<br>• Undertaking thematic analysis.<br>• Amalgamating/ processing/comparing data across sources. |
| Data cleaning | • Filling in gaps by predicting missing values based on patterns in the dataset.<br>• Standardising formats for dates, names, and categories to maintain consistency.<br>• Detecting and removing duplicate records, even when data is unstructured or messy.<br>• Tidying up text—removing unnecessary words, standardising terminology, and correcting typos in qualitative data (e.g., interview transcripts, survey responses). |

## Phase 5 - Writing, producing and disseminating research results

Writing up of results and discussion of findings is often the final step of a research project, along with recommendations for their practical application shared with a specific audience. This is an extension of one of GenAI's most common uses: creating written outputs. GenAI can produce writing as part of previous steps, or it can be fed research highlights and develop writing from there.

One key consideration when using GenAI to develop written reports is  ensuring the confidentiality of the data. As noted earlier, personal and sensitive information should not be uploaded into digital systems. Current GenAI applications cannot guarantee that the data they process remain private or isolated from AI training models. There is a risk that ingested or linked data could be stored by AI companies or used to refine their models, potentially compromising privacy.

| Category | Examples of Possible Uses |
|---|---|
| **Writing** | • Summarising existing text, condensing text or pulling out key points, findings, or headlines and summarising information from documents or interview transcripts.<br>• Drafting executive summaries and abstracts.<br>• Drafting outlines based on research/data inputs.<br>• Generating text - this can vary from short paragraphs to full length reports.<br>• Analysing tables and images to create supporting narratives for data tables and images during analysis and for reporting. |
| **Editing** | • Copy-editing and review of grammar.<br>• Editing and refinement of writing to make it 'more publication ready.' |
| **Formatting** | • Generating relevant images and protecting the identities of real survivors; generating non-triggering depictions of violence against women.<br>• Creating diagrams, charts or other graphic visualisations. |
| **Adapting and summarising** | • Creating summaries of research or adapting the language and tone in reports and summaries to make them suitable for different audiences. |

# 3 Assessing the risks of using GenAI for research on violence against women

The more you learn about the risks associated with GenAI, the more overwhelmed you may feel. Building your critical AI literacy is a good way to feel more empowered to understand GenAI better and make responsible choices around when it might be safe to use.

Like other big tech platforms and tools, GenAI brings ethical and safety challenges that individuals alone cannot resolve. Yet as it becomes embedded into many of the digital platforms that we use daily, it may soon be difficult to avoid using it – or having it used on us!

The newness of GenAI can make it challenging to determine what/which real and perceived risks to be concerned about. The data practices, complexity, and opaqueness of GenAI systems mean that principles such as informed consent, do no harm, transparency, and confidentiality are harder to assess and guarantee. As noted earlier, racial and gender inequalities and discrimination are also perpetuated and exacerbated with GenAI. In addition, GenAI tools present entirely new risks for researchers and their participants, including copyright and plagiarism, alongside concerns about the impact of GenAI on workers' rights and job security, growing environmental impacts due to water and energy use of large data processing centres and storage centres, and the use of GenAI to spread mis- and disinformation which can harm democracy.

By increasing our awareness of the risks, we will be better equipped to mitigate them and make more careful choices about the ways we use GenAI and how we protect research subjects. We can also ensure that we are not contributing to diminishing levels of evidence and research quality by using GenAI, but rather finding ways to make the most of it to improve our work.

**❗ It is important to review the ethical risks outlined below[3] and others that are context specific before using GenAI for research on violence against women. Conducting a risk assessment to determine whether or not it's responsible to use GenAI or other types of AI is a critical first step.**

**❗ If you can answer "yes" to the majority of questions below and you have the appropriate safeguards in place, it's possible that GenAI can be used responsibly (though there may still be some higher level, structural ethics challenges with the use of GenAI like there are with the use of any Big Tech applications or platforms).**

[3] See MTI's Common AI Definitions & Risks for Development & Humanitarian Actors for more information.

## 1 Is there a clear benefit in using GenAI that traditional methods cannot provide?

- Does GenAI provide unique benefits that traditional research methods cannot?
- Would using GenAI significantly improve efficiency, accuracy, or accessibility?
- Does GenAI significantly improve the experience of research participants?
- Can you justify the use of GenAI (versus a traditional method) for this process?

## 2 Does GenAI use comply with ethical guidelines, survivor consent, and data privacy laws?

- Does the use of GenAI for your proposed purposes align with global research ethics and safety standards?
- Can the data be fully anonymised before using GenAI?
- Are there legal and ethical approvals in place?
- Have research participants explicitly consented to the use of GenAI?
- Does your GenAI use comply with national and international privacy laws?
- Are safeguards in place to protect participant data?
- Are we adequately fulfilling our duty of care as researchers?

## 3 Are human researchers actively involved in reviewing, validating, and taking responsibility for GenAI-generated results?

- Are humans actively reviewing GenAI-generated findings before publication?
- Is there a process to correct errors, biases, or inaccuracies in AI outputs?
- Is GenAI being used to assist researchers, not replace expert judgment?
- Are researchers transparent about GenAI's role in their work?

## 4 Can the risks of using GenAI be sufficiently mitigated?

- Are there ways to reduce or eliminate the risks?
- Have you ensured human oversight is part of the process of preventing harmful GenAI outputs?
- Do you have a plan in place for checking and correcting AI's biases?
- Have you checked with research participants to learn what risks they are most concerned about?

## 5 Is there a process in place for making and documenting decisions about proceeding with GenAI?

- Have you done a benefits-risk analysis to determine whether the potential benefits outweigh the possible harms?
- Have you documented your rationale for the use of AI?
- Have you disclosed and otherwise documented the use of AI in your work?

### An overview of key risk areas and associated harms

| Risk area | Example scenarios and associated harms |
|---|---|
| **Reliability**<br>GenAI tools are programmed to take a confident tone. This means they may provide answers or analysis that seem convincing but are actually wrong. These are usually referred to as "hallucinations," These fabrications might be obviously wrong or much more subtle, making them difficult to identify.<br>It's worth knowing that those who train AI models are actively working to improve reliability (it's in their interest). | • A GenAI powered tool used to gather quantitative and qualitative data from survivors responds to the user in a way that is re-traumatising, or provides them with dangerous advice.<br>• GenAI summarises a study as finding 'X is a major driver of violence,' when the actual study found 'X is a factor but not the primary driver.' This shift alters the research narrative.<br>• GenAI subtly alters a statistic or misattributes a quote, and the change is not noticed and leads to distorted knowledge or findings. |
| **Bias (related to reliability)**<br>The responses or decisions produce by GenAI can reflect biases related to race, gender, age, sexuality, dis/ability because the data used to train AI models are inherently biased. Some of the answers or outputs that GenAI creates are false, inconsistent, or misconstrued due to bias in the underlying data or the way that the AI tools are trained. This can be actively harmful to the person using the tool or to research findings and narratives. | • The prompts used when giving a GenAI tool a research related task are not phrased to exclude bias, leading to a higher likelihood of biased outputs that cause harm.<br>• GenAI might underestimate the risk of harm in instances of intimate partner violence or offer responses that victim-blame or minimise biases against women. This could lead to those reporting this type of violence not getting support.<br>• Biased training data or prompts that do not work to exclude bias can lead AI tools to provide toxic, traumatising or factually incorrect answers to those using them (for example, suggesting to a survivor that avoiding going out at night is a way to minimise the risk of sexual violence). |

| | |
|---|---|
| **Dehumanisation**<br>By replacing previously human-led activities or processes with GenAI, how much do we lose touch with the intangible but profound benefits that come from human to human interaction, as well as with the more subtle insights that come from applying our own emotional intelligence to data analysis? | • A survivor in need of human contact as part of their reporting or recovery process can feel abandoned or re-traumatised when their request for help is handled by a GenAI agent. Or the GenAI agent might not pick up on the nuances or seriousness of the help required.<br>• Researchers who have relied excessively on GenAI tools for analysing qualitative experiences 'lose-touch' with the human stories behind these data, leading to a reduction in research quality, and even less passion for advocacy. |
| **Data privacy**<br>GenAI companies, along with many other tech companies and commercial platforms, are using our data to enrich themselves and exert political and market power. The amount of data gathered, and how exactly they are used, and by whom, is often opaque, and sometimes unknown even by those who control these large models. While some tools claim that they do not store or use data to train their models, the lack of transparency and changing rules leave users of these tools open to privacy risks. GenAI violates many of the data specific GBV principles. | • A GenAI transcription tool uses the data gathered during an interview, including personally identifiable information, to further train and improve the model. In principle, employees at the company with the right permissions could access these data.<br>• Research participants providing their consent may not be doing so with a full understanding of what happens to their data.<br>• A small GenAI service provider gets acquired by a big tech company and the big company changes the data protection policy and begins selling data to third parties or using data in unexpected ways. |
| **Power & patriarchy**<br>Though the development of these tools is indeed primarily controlled by men and though the gender divide in development is an issue, it is not just the fact that they are men that causes the perpetuation of power imbalances. Beyond their identity, it is the fact that they (and the companies they own) are designing these tools in ways that are committed to perpetuating systems of oppression. Changing commercial and political interests mean that many companies change their priorities for financial gain when it suits them. | • GenAI safety frameworks, even by those companies like Anthropic, who pitch themselves as 'safety-first' do not look at safety through a gender lens - they mostly focus on reliability or the risk of GenAI sentience.<br>• A change in government leads to a shift in priorities and a step back from interest in gender equity, safety and protection. |

| | |
|---|---|
| **Relevance (related to bias)**<br>GenAI tools will provide answers based on their training data, and if these training data do not reflect the reality of the person using it (including language), the response provided or action taken is likely to be, if not incorrect, then less applicable and less appealing to the user. | • A GenAI-powered translation tool fails to interpret the contextual nuances in regional dialect, reducing the quality of insights generated, and failing to represent the complexity of respondent voices.<br>• A GenAI-powered chatbot developed to provide training and support to community-based researchers provides advice which, whilst not wrong, is not rooted in local realities. As a result, the tool, which is an otherwise valuable and easy to use resource, is not used. |
| **Transparency & replicability**<br>Proprietary models lack information for open and reproducible research. Once a language model is retired, a researcher cannot replicate their outcome. Even for the same model and same prompt, it is common for outcomes to change over time and researchers don't know why this happens. | • Replicability and reproducibility are vital for research integrity. If the use of LLMs and GenAI prevents replication, it undermines scientific credibility, making findings less transparent, harder to verify, and more vulnerable to questions about validity. |
| **Inclusion & participation** (related to relevance and bias)<br>The models used by GenAI powered tools have, for the most part, been developed without input from those they are interacting with (for example, Gender Based Violence survivors or experts), leading to the models' responses or decisions not reflecting specific needs and realities. | • At a systemic level, a lack of representation can lead to the disenfranchisement and marginalisation of vulnerable groups.<br>• If groups feel that GenAI tools do not reflect their needs, this will erode trust and turn people away from using tools that might otherwise bring real benefits. |
| **Workers' rights**<br>GenAI models require human labour as part of the data preparation and evaluation phases. The work is often menial, repetitive, poorly paid, isolating, and unreliable. Tasks are often conducted in isolation from the wider development process, with limited opportunities for education or upskilling. | • By using GenAI-tools which aren't transparent about the pay, conditions, contracts, management, and representation of workers in their value chain, we may be perpetuating abusive, dehumanising working conditions and global economic inequities.<br>• Junior staff members or other individuals hired to prepare / label data which includes sensitive disclosures can themselves become triggered or burnt out emotionally. |

| | |
|---|---|
| **Inequity**<br>Emerging evidence suggests that the use of GenAI as a work tool can either make mediocre outputs appear outstanding or widen the gap between outstanding and poor performances. Especially in research institutions, there is a fear that using GenAI is anti-meritocratic, and further, will enhance disparities already present globally. This may be changing as the geopolitics of GenAI is changing, with more investment in local languages, and the rise of new models. | • Researchers in high income countries, already at an advantage over their counterparts in low and middle income countries, see that advantage magnified thanks to their disproportionate exposure to the digital literacy skills, and access to reliable and affordable electricity, internet, that the use of GenAI requires. |
| **Environmental harms**<br>GenAI models require significant energy and natural resources to train, run and maintain. This includes the electricity and water required to run (and cool) data centres, and hardware leading to electronic waste. There is also evidence that the stress and displacement caused by climate events leads to increased incidences of violence against women. | • The climate crisis is exacerbated, as are incidences of violence against women.<br>• Resources are diverted from research into violence against women in response to acute climate-induced crises.<br>• Climate related events may reduce access to affected populations, making it harder for researchers and implementers to reach survivors of violence and collect evidence. |

# 4 Getting started: best practice tips

In this section we walk you through some tips on how to make the big question of whether and how to use GenAI more manageable.

## 1 Identifying a guidance framework

Using new tools can always be a challenge, but rarely do they come with so many unknowns as GenAI. This doesn't mean GenAI should not be used at all, however it is important that violence against women researchers have a strong understanding of how to determine when using GenAI is appropriate. Some of the following frameworks might be useful. You can also find a list of ethical and safety considerations on the SVRI website.

- Putting Women First: Ethical and Safety Recommendations for Research on Domestic Violence Against Women, Department of Gender and Women's Health Family and Community Health. World Health Organisation. 1999
- Ethical and safety recommendations for researching, documenting, and monitoring sexual violence in emergencies, WHO, 2007
- Ethical and Safety Recommendations for Research on Perpetration of Sexual Violence: SVRI and Medical Research Council South Africa. 2012
- Ethical and safety recommendations for intervention research on violence against women: RTI and WHO. 2016

Existing research agendas - which have been developed in consultation with policy makers, survivors, and local communities - should guide the inception, methods, and analysis of violence against women research. GenAI tools only have access to knowledge that has been published. Oral traditions, community-held knowledge and other less machine-readable experiences are absent from the datasets that GenAI tools are trained on.

- The CARE principles on indigenous data governance are useful for ensuring that indigenous and community voices are taken into consideration regarding GenAI, the use of data, and ownership of data and AI systems.

## 2 Starting small with GenAI

It can be tempting to go big with GenAI, after all many of the promises of GenAI claim it will revolutionise research. However, our research shows that GenAI use is still minimal, and much remains to be learned.

One helpful analogy for GenAI is to think of GenAI as a research intern: keen and enthusiastic but still learning. Their assigned tasks should be focused, and their work double-checked!

**Tips:**

- **Identify the problem you are trying to solve:** one of the most important pieces of advice when using GenAI is to know why you are using it in the first place. Starting with the problem ensures that you pick the right tool to help solve it.

- **Map out a clear workflow:** A lot of GenAI discussions imply a simple A to B workflow, where a person gives a GenAI tool a prompt and it comes out with a complete product. We would encourage you to think about the full workflow around and including the tasks you want to use GenAI for. This can ensure that you have space for review and intervention for GenAI uses you are unfamiliar with.

- **Use data you are familiar with:** Auditing GenAI responses is easiest to do when you already have some understanding of the source material. This can also help with figuring out the limitations of GenAI for your applications. For example, if you are using GenAI to help generate themes from literature, knowing the material will help you identify whether GenAI's extraction of themes is accurate or aligned with your interests will mean you can better identify false information. This remains one of the biggest challenges with GenAI – how do we know what we don't know? Cross-checking findings using multiple sources, including domain experts, peer-reviewed literature, and alternative AI tools might help differentiate between actual gaps in our knowledge and AI-generated inaccuracies.

- **Be prepared for GenAI to require a lot of checking and verifying,** even as it saves time and enables access to a huge amount of information that you wouldn't have been able to process before.

- **Ask specific closed questions:** GenAI tools are broadly speaking much better at giving answers to direct and clear questions. Insights from early users suggest that asking a question where there are established answers can enhance GenAI's contributions. Asking broader, open-ended questions that humans would struggle to agree upon, for example, 'How can we prevent violence against women?' is where GenAI can be less predictable and more prone to mistakes.

- **Break down larger tasks into smaller components:** Overloading a prompt can increase the chances of an undetected or difficult to fix fault. Instead, think about the different steps you might take to achieve the desired outcome, and select the most appropriate steps for GenAI based on our other tips.[4]

- **Quality check GenAI outputs:** Having human review of outputs is vital especially if you are at the initial stage of figuring out whether an application of GenAI works for your needs. For instance, you should always double-check literature sources before citing them, as GenAI tools have demonstrated a tendency to make up sources.[5]

- **Consider and compare GenAI outputs to alternatives:** Rushing into universal use of GenAI can introduce more risks than necessary. Thinking about other ways to achieve similar results can encourage a more nuanced approach. It also means that uses of GenAI are well thought out rather than a GenAI-by-default response.

- **Think about taking an open approach to GenAI use:** Share your findings and prompt strategies with the rest of the research community to help others learn alongside you and encourage cross-pollination of ideas. It can also help both yourself and other researchers to sense check approaches to, and uses of GenAI.

## 3 Data protection of sensitive survivor information

Handling of highly sensitive information is one of the primary concerns raised by violence against women researchers. The easiest way to protect sensitive information is to not input it at all! Avoid the use of GenAI for sensitive information that contains details that are personal or could be personally identifiable. Though this could be frustrating, it is the safest way to ensure data are protected from being accessed by third parties, even if the GenAI provider claims to not store inputs. Instead, researchers can use GenAI on other parts of the research workflow that are more appropriate for GenAI.

💡 Avoid inputting personally identifiable sensitive information into commercial tools, as their data handling practices may not be fully transparent or within your control. Additionally, we recommend that you do not use commercial tools to anonymise data. Commercial platforms often change their terms and conditions, opening research participants up to the risk that their data could be shared with third parties in the future.

It is possible to create organisationally owned, small language models, that allow for greater oversight over how data are accessed and used. We acknowledge this may be technically challenging for organisations and researchers. In the absence of this capacity, prioritising the confidentiality and protection of those who have experienced violence must be the most important consideration.

[4] There are plenty of places to research “how to do good prompt engineering” and the field is changing quickly. Some believe that the advent of more advanced GenAI-enabled research tools and new features in common commercial AI models will do away with the need to write prompts. For more details on how to write useful prompts, see Ethan Mollick’s work Captain’s log: the irreducible weirdness of prompting AIs, A guide to prompting AI (for what it is worth), and Working with AI: Two paths to prompting - by Ethan Mollick

[5] See High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content | Cureus, Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions - Mayo Clinic Proceedings: Digital Health, ChatGPT Hallucinates when Attributing Answers | Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region

## 4 Mitigating bias

Researchers across different sectors have been working to develop ways of measuring bias in GenAI tools. However, given the breadth of data consumed by GenAI tools, and increased interest in using synthetic data to train models, it may be difficult for violence against women researchers to be certain of precisely how bias is impacting outputs. The two types of most relevant bias, and also the most frequently cited by experts we spoke with, were gender bias and language bias due to LLMs being primarily trained using more dominant languages.

Avoiding gender bias with GenAI can be challenging with externally developed tools. Ensuring that outputs are iterative, and cross-checked by those with expert knowledge is a critical first step. More participatory approaches to GenAI tools where labelling is done by trained individuals can also help reduce gender bias. Some researchers have raised concern about linguistic bias of GenAI tools when processing qualitative data. Here a two-step process could help with researchers doing a first pass before GenAI tools are then used for labelling.

💡 Where researchers have gender-sensitive questions or tasks, they should always double check GenAI outputs. Additionally, researchers should not make critical gender-sensitive decisions solely based on GenAI outputs. Validation of decisions with other experts can help ensure that gender bias in GenAI tools do not have adverse gender impacts.

Researchers should be wary of the use of synthetic (or GenAI-generated) data in model training as this can amplify gender bias.

In the long term, the sector should consider developing sector specific models that use vetted data. We need to start a dialogue to conceptualise what a feminist GenAI space for VAW research could look like.

## 5 Disclosing use of GenAI

As with all methodology choices it is important to explain in a transparent way how you put together your research and insights. Sharing when GenAI is used is an important part of research integrity, and it also can help other researchers to gain a better understanding of how GenAI tools are being used by violence against women researchers.

While some may think it's excessive to require disclosure of the use of GenAI— especially since tools like Google Search, Grammarly[6], or even human editors are not typically disclosed—some research journals now require transparency about GenAI usage. While this may evolve over time as comfort levels increase and GenAI's limitations are addressed, it remains important to disclose your methodologies— as you would with any other research methodology and tool. Given the potential for GenAI to be highly biased, this type of disclosure takes on even greater importance as a way to double check our work and ensure we are accountable.

[6] As algorithms become more pronounced in our work, even the use of search engines can introduce bias into our work and should be disclosed. Basic editing aids such as Grammarly, spell-check and citation managers may not require disclosure as they do not significantly affect interpretation or phrasing. (It should be noted however that tools like Grammarly do access all data in a document, and if using these tools for editing sensitive documents that contain personal, sensitive or confidential information, researchers should be sure to check the application's privacy settings and terms and conditions).

Research journals and universities require disclosure of the use of AI during research so that research can be replicated - however as noted in the risks table, use of proprietary AI models makes transparency and replicability a major challenge.

- All use of GenAI should be disclosed. Researchers should explain how they used GenAI, why they used it, any bias mitigation techniques and the validation methods used to cross-check GenAI outputs. Researchers should also share which tools they used.

- To deal with replicability issues, authors should reproduce their results with an open LLM whose training data and model weights are publicly available. Alternatively, they can demonstrate reproducibility of their results over time.[7]

[7] For an in-depth overview of AI models/Large Language Models and issues of transparency and replicability in research, see Crockett, M.J. (2025) "Ethical Costs and Epistemic Risks of LLMs: A Resource for Psychology Researchers" Princeton University.

# 5 Final thoughts

Generative AI represents a technological leap forward, but its utility, accuracy, and applicability are still being determined. Researchers have reasons to be fearful of or ethically concerned about the use of GenAI. At the same time, much like internet usage now dominates our ways of working, we are conceivably not too far from a time where GenAI is similarly unavoidable. The urgent task then becomes to engage, test and learn in safe and iterative ways, and gather further evidence about its capabilities and limitations as well as ways that it engenders harm. Experimenting with GenAI on researcher-developed proxy data or with pre-vetted literature, in a controlled environment can help researchers get more familiar with GenAI. Researchers should remember to be rigorous: repeatedly testing for reliability and validity is critical.

Testing, validating and working to understand when and where GenAI could work for you is a first step towards developing an ability to judge appropriate GenAI applications. This can be a time-consuming process, and it is worth considering if the time invested in understanding a comprehensive GenAI application is worth the gains. At this stage of GenAI, organisations are realising that the resources required to use GenAI properly are currently often greater than the value added. However small, specific tasks can be greatly sped up with GenAI, suggesting that these kinds of uses are where efficiency is to be found with fewer risks.

Curiosity tempered with care and caution can help researchers approach their engagement with and use of GenAI. This guide is only a starting point, and as you continue to reflect on how you want to use GenAI we encourage you to return to these foundational guidelines and think about how you might want to update them to reflect your newfound knowledge. Ultimately it is important that researchers or those using GenAI for research manage their expectations. GenAI may be very useful in some ways and far less useful in others — the key is figuring out what those applications look like for you and your work, and if and how you can mitigate risks to an acceptable level.

# Appendix A. Key definitions[8]

This Appendix provides some clarity on the terminology you might be coming across, and using, yourself, when engaging in conversations about AI.

## Artificial Intelligence / AI

Artificial Intelligence is an overarching term that describes the use of computer programmes or machines that perform tasks that would typically require human intelligence, from learning, to problem-solving, and language understanding.

AI is used as a 'catch-all' term, but can encompass any of the following:

- When you ask ChatGPT to explain the root causes of violence against women in a particular country.
- When Google analyses your past search habits to recommend results.
- When you use spellcheck, Grammarly or DeepL to assist in writing a paper.
- When you use Google maps and follow the traffic-optimized route during a research trip.
- When a research subject answers a survey via a WhatsApp chatbot.

The all-encompassing nature of the term has undoubtedly contributed to our sense of being 'overwhelmed' - it feels like AI is everywhere and nowhere all at once. It is also increasingly used as shorthand for the latest type of AI, Generative AI.

## Machine Learning / ML

Machine learning is a type of AI where computers are programmed to recognize patterns in data. Once they 'learn' to recognize these patterns, they can be left to run on their own to apply the 'learning' and make decisions on other data without being given additional explicit instructions. A computer could be programmed to recognize words used in hate speech and to then flag social media posts that include those words or combinations of words.

## Natural Language Processing / NLP

Natural Language Processing is a subfield of AI focused on language related tasks. Computers can be programmed to recognise patterns, grammar and syntax in 'natural language' and then to interpret and more recently, generate language in a way that sounds human. The types of tasks that could be supported by NLP in research on violence against women include sentiment analysis, for example, to try to understand if a certain policy or campaign had changed public sentiment about violence against women based on the comments posted on Facebook or other social media pages.

## Tokens

Tokens are the smallest units of text that AI processes. They can be whole words, parts of words, or even punctuation marks. During tokenisation, raw text is broken down into tokens, which are

[8] See MTI's Common AI Definitions & Risks for Development & Humanitarian Actors for more information.

then converted into numerical values so that models can analyse the text. Essentially, tokens serve as the building blocks that enable AI systems—especially in natural language processing—to identify patterns, generate responses, and perform various language-related tasks.

## Large Language Models / LLMs

LLMs are an advanced type of NLP model built using multiple layers and billions of data points and variables. After their initial programming by humans, they can generate increasingly humanlike outputs based on their training data and human fine-tuning. The most widely known generative AI tools are powered by LLMs.

## Retrieval Augmented Generation

LLMs can be adapted, or 'fine-tuned', using processes like Retrieval Augmented Generation (RAG) to make them more reliable or more relevant for the intended audience. To conduct RAG a database or collection of documents is uploaded. The Large Language Model (LLM)I is then programmed to fetch relevant information from this dataset to generate more accurate and context-aware responses.

## Generative AI / GenAI

GenAI is powered by advances in Natural Language Processing, specifically, Large Language Models. They work by using vast quantities of data for training to make predictions about the most likely next word in a sentence, in a split second. When given an instruction or asked a question, GenAI tools can generate content, whether text, images or video, that looks like it was produced by a human[9]. However, Generative AI is completely dependent on the information it can scrape from the internet. It should be noted that this is a major barrier for researchers, as much of the published data on violence against women and VAC sit behind pay walls not accessible to Generative AI models.

## Predictive AI

Increasingly referred to as 'old AI' or 'traditional AI', predictive AI is the type of AI most of us were using, knowingly or not, before the GenAI boom in November 2022.[10] It's a form of AI that analyses data (words or numbers), and makes predictions, or classifications, usually in service of a specific task. For example, a predictive model trained on anonymised data on disclosure can learn to recognise sentences in a chatbot that are indicative of a disclosure and raise a flag to online moderators or redirect a survivor to appropriate resources.

Although Predictive AI is not getting as much coverage because of the excitement over Generative AI, one key advantage of Predictive AI is that the inputs (the data it is trained on) and outputs (the action it performs based on its prediction) are almost entirely controllable by its maker. Increasingly, AI powered tools use a blend of Predictive and Generative models depending on the function required, as each model has strengths and weaknesses.

[9] See Stephen Wolfram's What Is ChatGPT Doing … and Why Does It Work?—Stephen Wolfram Writings for a more detailed explaination.
[10] See: https://openai.com/index/chatgpt/ See: https://openai.com/index/chatgpt/

## Algorithm

A set of instructions or rules designed to perform a specific task or solve a problem using a computer. Algorithms are used to train the models which power AI.

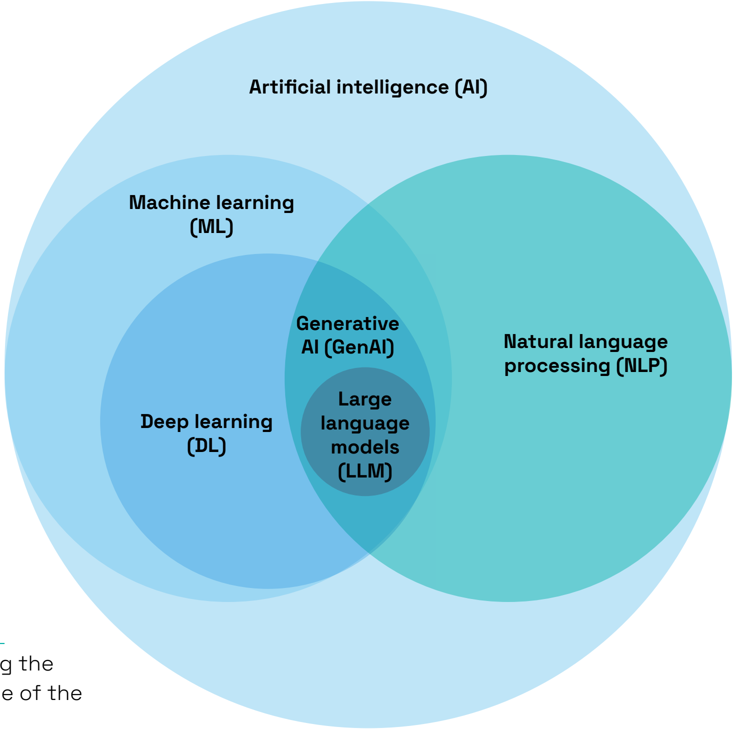

Artificial intelligence (AI)
Machine learning (ML)
Generative AI (GenAI)
Natural language processing (NLP)
Deep learning (DL)
Large language models (LLM)

Diagram by Pressman and colleagues, 2024 depicting the relationship between some of the concepts in this section.

## Model

A representation of the inter-relations between data, created by applying an algorithm to data. Models capture patterns, structures and relationships which can then be used for predictions (as in predictive AI) or to generate entirely new data (as in GenAI). Models are the 'engine' or 'brain' driving AI tools.

## SLMs - Small Language Models

SLMs are a form of GenAI that can be used to generate new text or images, but they use less data and are less complex. Crucially, they can be ideal in resource-limited contexts, such as on mobile devices. They can still perform a variety of tasks, like text classification or question answering, but are better suited to simpler tasks, for example, producing summaries in a specific format from raw data.

## Chatbots

The term chatbot has two meanings depending on the context. Until recently, chatbots referred almost exclusively to a digital service, usually available via a chat interface on a web browser or instant messaging app like WhatsApp, that enables users to have a human-like conversation via text or voice. Many chatbots are not AI powered and instead use a pre-determined decision tree architecture that allows users to browse a menu of options (though they may still seem 'chatty'). These chatbots often incorporate a blend of mechanisms, for example, pre-determined elements, GenAI, and Predictive AI.

Since the advent of GenAI, the term chatbot has been increasingly used to describe GenAI powered virtual assistants such as ChatGPT, Claude, Gemini or DeepSeek.

## Prompts / prompt-engineering

Prompts are instructions given to an AI powered tool to generate an output, for example, a response to a question. They can be developed or input both by those designing the tool, and those using the tool:

- Prompts are written by those designing AI tools to provide the model with more or less strict 'guardrails' in order to improve the quality and safety of its outputs (these are called System Prompts). For example, Anthropic, the organisation behind the GenAI model Claude, makes its system prompt available publicly to provide more transparency on the instructions behind the model to help users understand the biases in the data.[11]

- Additional prompts are then written by the user of the tool, to elicit the specific information they are looking for. Good prompts should be clear, specific, and involve a degree of iteration (playing around with the prompts a few times to get the best results). For example, "What are the most effective community-based interventions for preventing gender-based violence in low-income settings?" is considered a good starting prompt.

Prompt-engineering is the process of writing and refining prompts, with the goal of improving the nature and quality of the tools' response to instructions and/or questions.  Some of the practical challenges and ethical issues associated with AI use can be mitigated through responsible prompt-engineering, for example, by including in the prompt a request to ensure diverse datasets are referenced, or to include post-colonial theory or to consider power inequality in a prompt related to violence against women. Other ethical issues baked into the extractive business model such as power dynamics, and algorithms remain difficult to address in a regular use of an AI tool.

[11] See Claude's latest and historical system prompts: System Prompts - Anthropic. Part of Claude's System Prompts include things like "Claude is always sensitive to human suffering, and expresses sympathy, concern, and well wishes for anyone it finds out is ill, unwell, suffering, or has passed away."

# Appendix B. Methodology and scope

This guide draws on a light touch literature review, supported by qualitative and quantitative data collected from those actively involved in advancing knowledge on violence against women. With the intention of releasing this guidance in a short time-frame in response to the rapid pace of developments of this field, the data collected were necessarily non-exhaustive and we see potential for a more rigorous study. Similarly, given the cutting-edge nature of using the latest forms of AI, the literature review reflects the paucity of evidence and examples available currently. The guide offers considerations about the use of GenAI in research on violence against women as of **January 2025.**

### Desk research

Considering the subject matter of this guidance, we conducted a limited and constrained experiment with GenAI to see how useful it would be in acting as a low-level research assistant for conducting a literature review and generating a synthesis. (The use of GenAI here was not intended to help with the written output, e.g., with writing the report.) We first independently gathered and reviewed key sources providing guidance on AI in research, guidance on research in violence against women, and current or potential uses of AI in research on violence against women. We summarised each resource individually and noted key themes, key quotes and the emphasis of the main argumentation. We then queried the resources using Chat-GPT by inputting a thematically arranged selection of resources.

Our prompts were highly detailed and specific to direct Chat-GPT in a clear manner to use our selected resources as the information source. We limited Chat-GPT to these sources as our existing knowledge of the resources enabled us to fact check and sense check Chat-GPT's outputs. The aim of using Chat-GPT was two-fold: to develop a hands-on understanding of how Chat-GPT can handle research related queries for violence against women purposes to better inform this guidance, and second, to see if Chat-GPT could uncover insights or connections that we had not initially identified ourselves.

While we found this was an interesting experiment, we did not include the GenAI-generated outputs in this paper as they lacked nuance, and the writing was flat and dull. In comparison with our own review of the literature, the use of GenAI did not draw out anything new or provide useful insights. Chat GPT and AI were not used in the writing of this report aside from this disclosed experiment, which did not yield useful outputs.

### Quantitative data collection

A 5-question online survey was developed to understand more about researchers' knowledge, attitudes, concerns and needs relating to GenAI, including both closed and open questions (See Appendix A). The survey was promoted during the SVRI Forum in October 2024, and to members of the Tech-Facilitated GBV Community of Practice, as well as via LinkedIn.

### Key Informant Interviews,  Group Discussions & Review Processes

Brief interviews were conducted during the 2025 SVRI Forum with a focus on the question of attitudes, concerns, and uses of AI in respondents' work. The authors also co-convened an online working group session with the SVRI TFGBV Working Group, and an in-person Technology Salon (roundtable), during which attendees were asked the same questions.

The guidance was then sent out for external expert review, including colleagues from SVRI core team; SVRI Leadership Council; UNFPA; UNWomen; Spotlight. External reviewers included: Dr Claudia Garcia-Moreno (WHO); Prof Kumudu Wijewardena (emeritus professor of community medicine, University of Sri Jayewardenepura); Prof Rachel Jewkes (South African Medical Research Council); Joanna Włodarczyk (Empowering Children Foundation); Alexandra Robinson (UNFPA); Stephanie Mikkelson (UNFPA); Anahita Alexander Sefre (UNFPA); Yeliz Osman (UNWomen); Aïssa Boodhoo (Spotlight) and Samu Ngwenya Tshuma (Spotlight). The document was also reviewed internally by SVRI staff, including Dr Sangeeta Chatterji, Ayesha Mago, Dr Joan Njagi and Arti Mohan.

### Respondents

| Method | No. | Additional comments |
|---|---|---|
| Rapid Online Survey | 22 | No demographic data were collected but the respondents were primarily attendees of the SVRI conference, who were mostly female. |
| SVRI Forum - KIIs | 16 | 5 Male, 11 Female, from a variety of backgrounds, including researchers, lawyers, students, public health professionals and NGO founders. |
| Online Session with the TFGBV Working Group | 38 | No demographic data were collected but participants were majority female, from across the globe, and were English speakers. Participants worked in public health, tech, and GBV prevention and response. |
| In person Tech Salon Meeting - roundtable discussion | 30 | A mix of UN agencies, foundation staff, global NGOs and New York City-based organisations working on violence against women. |
| External review | 10 | SVRI Leadership Council (4); UNWomen (1); Spotlight (2); UNFPA (3) |

# Appendix C: Survey questions

To get a rapid sense of how researchers in the field of violence against women are feeling about and using GenAI, we conducted an informal survey during the SVRI Forum. The survey link was also shared with the SVRI's TFGBV Community of Practice. The purpose of the survey was not to obtain a representative or statistical sample but rather to ask potential users of the current guidance what topics they would like to see covered or reflected in the guide.

1. How confident are you that you can explain the distinction between different kinds of AI?

2. Which of the following ways of using AI as part of research are you aware of? (pick as many as apply)
   a. Help with summarising during formative or desk research
   b. Formulating or reviewing research questions
   c. Translating stakeholder interviews/ surveys
   d. Analysing quantitative research data
   e. Analysing qualitative research data
   f. Help rewording research outputs, e.g. briefs or papers.
   g. Other

3. Which of the following ways of using AI as part of research are you currently doing/ planning to do (pick as many as apply)
   a. Help with summarising during formative or desk research
   b. Formulating or reviewing research questions
   c. Translating stakeholder interviews/ surveys
   d. Analysing quantitative research data
   e. Analysing qualitative research data
   f. Help rewording research outputs, e.g. briefs or papers.
   g. Other

4. What issues relating to the responsible/ethical use of AI in violence against women research have you been most concerned with? [open]

5. As part of our guidance, we will attempt to address some FAQs. What are your main questions when it comes to the ethical use of AI for research on violence against women? [open]

# Appendix D: Natural Language Processing (NLP), Machine Learning (ML) and AI in violence against women research

As mentioned above, there is limited evidence of GenAI use in research on violence against women. But use of ML and NLP tools points to where GenAI might first be introduced to accelerate the processing of unstructured data, and to potentially broaden the application of existing AI tools. NLP and ML tools have been used to process diverse data sources related to intimate partner violence for example. Below we offer a few examples.

| Purpose | Example |
|---|---|
| **Using NLP to process unstructured data to identify instances of intimate partner violence.** | NLP in Hospital Settings (IPV):<br>• One study used NLP to identify intimate partner violence (IPV) in emergency department settings.<br>• It developed an NLP algorithm to analyse unstructured electronic health records of the emergency department.<br>• The records were screened for 72 different terms associated with IPV to identify an "IPV-related encounter."<br>• 29 of these were situational terms e.g. assault by spouse, and 49 were extended situational terms e.g. attack by spouse.<br>• The NLP algorithm achieved 99.5% precision.<br>• The intention is to use this algorithm to identify instances of IPV in near-real time to allow for more time-sensitive interventions. This can help improve the underidentification of IPV in healthcare contexts.<br>NLP applications for processing digital evidence:<br>• NLP models were used to review text-based material to find words and phrases relevant to violence against women in criminal investigations.<br>• The models were able to detect the use of slang and informal language that pointed to abuse.<br>• Overall Britain's Forensic Capability Network found that the AI tool was 21 times faster than a human investigator.<br>• Given the large number of unexamined digital devices held as part of investigations in Britain, this tool can assist with flagging messages of note to human investigators.<br>• A similar project used NLP models in combination with NLP to find perpetrators who used technology to exercise coercive control e.g., through surveillance, overbearing text messages, social media threats and accessing online accounts. Researchers applied these tools to various forms of mobile and social media data to enhance risk assessment efforts. |

| | |
|---|---|
| **Using ML and predictive AI to predict risk and vulnerability factors of domestic violence /intimate partner violence.** | Predictive Models in Liberia:<br>• This study used machine learning to identify factors contributing to domestic violence in Liberia.<br>• Researchers used data from the 2019/2020 Liberian Demographic and Health Survey (DHS).<br>• The researchers compared a range of machine learning algorithms.<br>• Only two of the seven machine learning models performed well, but with a number of inconsistencies.<br>• Researchers suggest that using a hybrid approach—which combines multiple machine learning techniques—could improve accuracy and robustness.<br>• Integrating data from the DHS could provide a more comprehensive and accurate foundation for machine learning models, enhancing their ability to identify patterns and contributing factors.<br>Predicting vulnerability in South Africa:<br>• Researchers in South Africa used ML to investigate unspecific relationships between different IPV risk factors. They were specifically seeking "hidden and complex patterns and relationships in the data."<br>• The dataset was drawn from the 2016 South African Demographic and Health Survey and focused on ever-married women, or women who had been married at least once.<br>• Tree-based machine learning models were used to predict occurrence of IPV to identify women especially vulnerable to IPV, and to ascertain which factors most increased risk of IPV.<br>Predicting IPV perpetration among homeless youths<br>• Researchers used interpretable machine learning to identify key predictors of perpetration among a sample of 1,426 youth experiencing homelessness in 7 US cities.<br>• Interpretable machine learning model's enable researchers to see which factors influence the models outcomes and why.<br>• Experiencing IPV victimisation whilst homeless was the biggest predictor of IPV perpetration.<br>• The study recommends future IPV prevention efforts recognise the interconnectedness of IPV perpetration and victimisation among homeless young people. |
| **Using GenAI and traditional AI to understand individual experiences of IPV** | Using Participatory AI to study IPV in Iran:<br>• Researchers undertook a participatory human-machine approach to develop an AI tool that can extract user comments on IPV from an Iranian parental health website.<br>• Both human annotators and GPT 3.5 were first asked to identify IPV relevant comments. This was used as the basis for classification and processing of the website's content.<br>• The researchers then used a supervised ML technique of active learning to annotate the excerpts and categorise the content.<br>• The research provides important insights into how women in Iran narrate, relate and understand their experiences of IPV. |

THE MERL Tech INITIATIVE

SVRI sexual violence research initiative