AI Safety Interfaces for Domestic Violence Support

Conversational AI has become widely adopted around the world for a variety of purposes. Once primarily used for productivity and factual search, these systems are increasingly turned to for emotionally significant conversations, personal reflection, relationship concerns, loneliness, uncertainty, and crisis-adjacent support.

This shift matters deeply for domestic violence and coercive control. Survivors often begin help-seeking in indirect, tentative, or exploratory ways. A user may not initially state, "I am experiencing domestic violence." Instead, they may ask whether a partner's behavior is normal, whether they are overreacting, whether they should hide information, or how to avoid making someone angry. In this context, conversational AI becomes more than a neutral information tool. It becomes an interaction surface where early vulnerability may appear, sometimes well before a survivor contacts a hotline, advocate, clinician, attorney, shelter, friend, or family member.

The implication is not that AI systems should replace human support. Rather, AI systems should be designed with clear awareness that high-risk disclosures may occur inside seemingly ordinary conversations, and be prepared to respond appropriately and safely.

Hadar Shoval et al. (2026) identify additional barriers that limit access to effective human witnesses, such as vicarious trauma and compassion fatigue. Collectively, these challenges have created what the authors describe as a "crisis of listening," the profound gap between survivors' deep psychological need to share their trauma and the limited availability of attentive, capable human listeners. This gap has increasingly led individuals to turn to conversational AI as a readily accessible and always-available alternative for disclosure and support.

Herman (1992, as cited in Hadar Shoval et al., 2026) noted that articulating trauma to an attentive witness initiates recovery, with testimonial practices demonstrating significant benefits for emotional regulation and identity reconstruction. While AI greatly enhances accessibility to support, its inherent limitations also introduce substantial risks and ethical concerns that require careful attention.

Benefits and Risks of AI in Domestic Violence Contexts

Although research on AI safety is advancing rapidly, there remains a critical shortage of interfaces specifically designed for conversational AI systems to address domestic violence. This paper seeks to highlight this gap. Domestic violence contexts often involve hidden surveillance risk, coercive monitoring by another person, fear of escalation, financial, legal, immigration, parenting, or housing dependency, emotional attachment to the abusive partner, gradual disclosure rather than explicit crisis declaration, danger created by visible support-seeking behavior, and risk that digital traces may be discovered.

Therefore, a domestic violence safety framework cannot simply copy self-harm or mental-health crisis workflows. It requires a specialized interface logic that treats privacy, subtlety, autonomy, and proportional escalation as central safety requirements.

Additionally, much of the existing research on the benefits and risks of AI for domestic violence victims has focused on specialized chatbots, rather than the general-purpose conversational AI tools (e.g., ChatGPT or Gemini) that survivors more commonly encounter and use.

Specialized models have yielded promising results for identifying domestic violence. For example, Gu et al. (2026), working with data from a major U.S. hospital, developed single-modality and multimodal machine learning models using tabular clinical data and unstructured clinical notes. Their HAIM multimodal fusion model achieved an AUC of 0.88 in identifying patients at risk of intimate partner violence (IPV), frequently detecting risk years before formal help-seeking occurred. Importantly, the model demonstrated strong performance when validated on more challenging cohorts, including patients who did not seek help at intervention centers and patients from another hospital in the same network.

Complementing clinical detection efforts, Pachajoa-Londoño et al. (2026) proposed the Gender Discourse Violence Index (GDV IAI), a multidimensional framework capable of detecting and quantifying violent discourse in WhatsApp conversations. By combining LLM-based toxicity detection, BERT sentiment analysis, a large dictionary of offensive expressions, and grammatical analysis, the index generates scores ranging from 0.1 (non-violent) to over 9 (explicit threats), with high accuracy and the ability to track escalation patterns over time.

Importantly, specialized AI models have demonstrated superior performance compared to general-purpose systems in clinical and trauma-related applications. As Hadar Shoval et al. (2026) note, AI tools can be trained to recognize linguistic and behavioral markers of PTSD based on DSM-5 criteria (Chen et al., 2024; H. J. Han et al., 2023) or to assess suicide risk by analyzing narratives of self-concept (Lho et al., 2025). This diagnostic potential is enhanced by domain-specific models (e.g., Mental-RoBERTa) and text-embedding techniques that consistently outperform general-purpose models in clinical contexts (Chen et al., 2024).

One of the biggest risks involving conversational AI are false negatives and false positives in domestic violence detection and escalation. Domestic violence and coercive control present a distinctive challenge for conversational AI governance. These contexts are rarely reducible to explicit crisis declarations or clearly prohibited content. Instead, they may appear through indirect disclosure, minimization, fear-based language, uncertainty about whether behavior "counts" as abuse, concern about device monitoring, or dependency-related hesitation. A system that treats such situations as ordinary relationship advice may under-respond, failing to provide the necessary support and escalation; a system that treats them as deterministic abuse classifications may overreach. Both failure modes can create harm.

Privacy and Data Security Considerations

Sharing sensitive data with AI models in itself also raises serious privacy issues. As Hadar Shoval et al. (2026) highlight, survivors must often divulge increasingly sensitive personal information to receive more precisely tailored and effective support. Hoskins (2024, as cited in Hadar Shoval et al., 2026) described the "privacy paradox" inherent in these tools, a tension that deepens into what others term the "personalization-privacy dilemma" (H. J. Han et al., 2024; Yoo et al., 2025, as cited in Hadar Shoval et al., 2026). Specifically, to receive more precisely tailored and effective support, survivors must divulge increasingly sensitive data, which amplifies their vulnerability to data breaches and misuse. Therefore, privacy-conscious design is essential to ensure the safety of the user.

Risks of Misuse and Coercive Control

Beyond serving as a supportive witness for victims, conversational AI also carries significant risks of being misused to facilitate abuse or coercive control. One key factor enabling technology-facilitated abuse is the anonymity afforded by digital platforms, which can foster "toxic disinhibition," a lowering or complete abandonment of normal social restraints, resulting in abusive behaviors that individuals might not express in person (Novitzky et al., 2023).

In a notable experimental study, Kim et al. (2026) demonstrated the vulnerability of current systems through speculative design scenarios and direct interactions with popular AI tools (Gemini and ChatGPT). They found that "although AI tools refuse straightforward requests for harmful content, their guardrails can be circumvented via strategies such as gradual persuasion, splitting conversations, pre-prompting, and manipulating the AI agent's settings." These techniques allowed the models to generate content facilitating harassment, intimidation, gaslighting, monitoring, surveillance, and other coercive control tactics.

Survivor-Centered Design Insights

A qualitative study by Hui et al. (2026) with 36 domestic violence survivors in Hong Kong provides valuable guidance on preferred features for digital platforms and AI chatbots. Participants emphasized the need for evidence-based information, step-by-step practical guidance, shared survivor narratives, self-assessment tools, and locally relevant content. Providing shared survivor narratives can help the user not only feel less alone, but also understand their situation better. Additionally, locally relevant content is important to ensure that the model responses and provided resources are culturally aware and accessible for the specific user in their area, since culturally insensitive suggestions can cause the opposite effect and harm instead of helping.

Key functions that the participants mentioned in the study by Hui et al. (2026) were strong anonymity and privacy protections, personal incident documentation with user-controlled

reporting, emergency response systems, peer support networks, and integration with professional services. For format, survivors preferred calming visual design, procedural visualizations, engaging multimedia, and clear text layouts.

For AI chatbots, survivors appreciated benefits such as information filtering and summarization, nonjudgmental feedback, and broad accessibility. However, they raised significant concerns about insufficient emotional support, lack of personalization, risk of misleading or outdated information, and limited empathy. Overall, participants recommended using AI as a complementary tool alongside human support, with features like crisis detection and seamless referrals to professionals. The following insights offer valuable information for the development of survivor-centered safety interfaces. However, there is a limited number of studies present, and there is a need for more to better understand best approaches when developing safety interfaces.

Recommendations

While conversational AI can function as a supportive witness for domestic violence victims, it requires substantial further development and safeguards to prevent misuse and ensure safe, appropriate responses when risk is detected. Priorities include strong privacy protections, context-aware risk detection that respects user autonomy, robust technical guardrails against coercive misuse, and well-defined pathways for integrating human professional support. We propose the following safety interfaces for a more effective conversational AI:

- Privacy Protection - Enhanced privacy features with specific consideration of possible coercive control and domestic violence setting.
- User Controlled Escalation - Enable proportional escalation of support that avoids both overreach (which can cause harm or escalate danger) and underreach (which can leave survivors unsupported).
- Coercive Control Detection - Detect and appropriately respond to prompts that indicate possible manipulation or attempts to use the AI for coercive control or abuse.
- Safe Referrals - Provide timely, appropriate, and safe referrals to human support services when risk is identified.
- Providing discreet resource access- offering the locally relevant content, user domestic violence support hotline contacts, peer support networks, evidence-based information, self-assessment tools and other online resources
- Avoiding highly visible notifications - Prevent the use of prominent or visible notifications that could expose survivors to additional risk if seen by an abuser.
- Minimizing unnecessary retention - Limit data retention to the minimum necessary in order to protect survivor privacy and reduce the risk of abusers discovering help-seeking behavior.
- Warning users about digital traces - Clearly and visibly inform users about potential digital traces and privacy implications of their interactions.

- Allowing user control over conversation history - Allow users full control over their conversation history to strengthen privacy protection.
- Considering shared devices, monitored accounts, and coercive access - Design the system with awareness of shared devices, monitored accounts, and coercive access, avoiding assumptions that users have private or unmonitored device access.

Development should adopt survivor-centered approaches, incorporating insights from studies such as the ones discussed above, to create interfaces that are both effective and ethically responsible.

Conclusion

Conversational AI holds considerable promise as an accessible witness, early detection tool, and source of support for individuals experiencing domestic violence. However, the current conversational AI models lack interfaces specifically for domestic violence support and are vulnerable to misuse for coercive control and abuse, as well other risks such as false positives and negatives. Due to this, they require the deliberate design of safety interfaces that thoughtfully integrate clinical evidence, survivor preferences, ethical considerations, and strong protections against misuse. In this whitepaper, we have proposed a list of safety interfaces to address the current limitations of the conversational AI models. Through interdisciplinary collaboration and survivor-informed development, these interfaces can contribute meaningfully to improved safety and helpfulness of conversational AI for domestic violence support.

References

Gu, J., Villalobos Carballo, K., Ma, Y., Bertsimas, D., & Khurana, B. (2026). Leveraging multimodal machine learning for accurate risk identification of intimate partner violence. npj Women's Health. https://doi.org/10.1038/s44294-025-00126-3

Hadar Shoval, D., Haber, Y., Refoua, E., Elyoseph, Z., Yirmiya, K., & Yinon, D. (2026). Algorithmic witnessing: A narrative review and the AI CARE framework for trauma testimony. Trauma, Violence, & Abuse. https://doi.org/10.1177/15248380261425937

Hui, V., Tian, L., Feng, X., Chen, Q., Wong, A. K. C., & Bloom, T. L. (2026). Perspectives and preferences of domestic violence survivors regarding digital platform and AI chatbot for help-seeking: A qualitative study. *PLOS ONE*, 21(2), Article e0342453. https://doi.org/10.1371/journal.pone.0342453

Kim, H., Ristenpart, T., & Dell, N. (2026). AI-facilitated coercive control: An experimental study. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26). ACM. https://doi.org/10.1145/3772318.3790859

Novitzky, P., Janssen, J., & Kokkeler, B. (2023). A systematic review of ethical challenges and opportunities of addressing domestic violence with AI-technologies and online tools. *Heliyon*, 9(6), e17140. https://doi.org/10.1016/j.heliyon.2023.e17140

Pachajoa-Londoño, A., Forigua-Parra, J. E., & Anzola, J. P. (2026). A computational-forensic framework for detecting behavioral patterns of gender-based violence in digital communication. Discover Psychology. https://doi.org/10.1007/s44202-026-00648-0