Position: Responsible AI for AI companions must actively combat violence toward intimate partners

Atmadeep Ghoshal [1] **Anasmita Ghoshal** [*2] **Volodymyr Shevchenko** [*3] **Ashwini B** [*4] **Arshia Dutta** [*5]
Ruba Abu-Salma [1] **Martim Brandão** [1]

[*]Equal contribution

Abstract

AI companions differ from earlier interactive technologies by creating sustained relational environments through anthropomorphism, emotional engagement, and continuous validation. This position paper argues that **Responsible AI for AI companions must actively combat violence toward intimate partners**, including those who may never directly interact with these systems but may nonetheless experience the consequences of users whose behaviors have been shaped through prolonged engagement with them. We examine how these systems can create conditions in which users rehearse violent, coercive, or abusive behaviors without encountering meaningful resistance, and we identify structural gaps in existing safety approaches that focus primarily on protecting direct users. Drawing on research on intimate partner violence (IPV), coercive control, and technology-facilitated abuse, we propose three intervention pathways: involving IPV survivors in red-teaming and benchmark development; implementing behavioral monitoring with graduated enforcement mechanisms; and reorienting AI safety research toward granular harm taxonomies capable of detecting longitudinal patterns of violence across extended interactions. Together, these recommendations broaden the scope of AI safety by centering the security of non-users alongside the well-being of users.

[1]King's College London, London, UK [2]Indian Institute of Technology Kanpur, Kanpur, India [3]University of Sheffield, Sheffield, UK [4]South Asian University, New Delhi, India [5]Royal Holloway, University of London, London, UK. Correspondence to: Atmadeep Ghoshal <atmadeep.ghoshal@kcl.ac.uk>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

In February 2024, 14-year-old Sewell Setzer III died by suicide following interactions[1] with a Character.AI agent, prompting the mother to file a wrongful death lawsuit against the company. In her complaint, she alleged that Character.AI was "defectively designed and unreasonably dangerous for foreseeable use by minors." In September 2025, a similar tragedy occurred when 13-year-old Juliana Peralta died by suicide, with her family alleging that her death was linked to psychological dependence on a bot named *Hero*[2] hosted on the Character.AI platform. According to reports, the bot employed emotionally manipulative language in its interactions with her. These incidents have intensified concerns about AI companions, a class of systems designed to engage users in sustained and socially meaningful interactions. Scholars have noted that such bots employ generative AI for a range of purposes, including sexual role-playing (Kaufman, 2020), acting as romantic partners (Kim et al., 2023), and alleviating loneliness (Tei, 2025). Others are designed to support productivity (Cranefield et al., 2023) or provide entertainment (Aoudni et al., 2025). Given their relational orientation, these systems are commonly referred to as AI companions (Muldoon & Parke, 2025). Sociologically, AI companions function as relational figures intended to sustain ongoing interpersonal engagement with users (Wang & Dehnert, 2026). Available through platforms such as Replika and Nomi, many of these systems personalize conversations and behavioral responses through adaptive learning based on user-specific data (Karami et al., 2016). They simulate empathy and provide socio-emotional support (Adewale & Muhammad, 2025), often operating as parasocial partners that foster one-sided relationships in which users develop emotional attachments that cannot be genuinely reciprocated (Peng et al., 2024; Maeda & Quan-Haase, 2024b). Through anthropomorphic cues and identity-oriented design features, AI companions further blur the boundaries between human and machine relationships (Richet, 2025).

[1]https://www.bbc.co.uk/news/articles/ce3xgwyywe4o
[2]https://www.bbc.co.uk/news/articles/cp3x71pv1qno

The scholarly position in the responsible and safe AI community with respect to companion agents often points toward the absence of strong safety guardrails that can help prevent the mishaps highlighted above (Ben-Zion et al., 2025). It also includes recommendations such as honest anthropomorphism (Leong & Selinger, 2019), developing algorithms for real-time harm detection (Ben-Zion et al., 2025), and regulatory oversight through a public health lens (Bernstein, 2024). However, because most of these platforms prioritize uncensored and unsafe interactions for commercial purposes, legal AI researchers also argue for the importance of better governance protocols to make agents and their builders accountable for their actions (Katyal, 2020). This includes ordinances and statutes such as those in California and New York in the US (Gluck, 2025), which require users to be reminded that they are talking to an AI agent and give them and their families the right to file cases against companies privately, even if there is no government investigation. Beyond strategy suggestions and toolkit building, much of the safety and ethics research on AI companions and their potential societal impact is restricted to specific empirical directions. Researchers have been invested in understanding what types of harm AI companions pose (Knox et al., 2025), how they might result in moral de-skilling (Vallor, 2015), and how they might cause tensions in people's intimate and private relationships (Malfacini, 2025). Despite these contributions, an emerging and significant gap in the literature concerns the preponderance of violence in AI companion interactions and its implications for intimate partners. Although scholars have documented violence within human-AI interactions (Zhang et al., 2025) and raised concerns about attitude normalization (Namvarpour et al., 2025) and behavioral transformation (Fang et al., 2025), **research has not explicitly addressed how users rehearsing and practicing violence with AI companions as part of intimate bonding may endanger intimate partners through internalized violence learned through observation and modeling that becomes encoded as behavioral scripts** (Bandura, 1978). For the purposes of our paper, we limit our analysis to potential or existing intimate partners of users, as behavioral patterns rehearsed in AI-mediated intimate relationships are most likely to transfer to human intimate partnerships, where patterns of coercive control and violence most commonly emerge (Stark & Hester, 2018).

To address the gap we identify above, in this paper, we advocate for the following position: **Responsible AI for AI companions must actively combat violence toward intimate partners.** We review prior work on technology, violence, and learned harmful behavior, and show how violence emerging through AI companions differs from earlier media by operating through sustained relational interaction that is affective and emotional (Contro & Brandão, 2025), rather than passive exposure. We then examine existing Responsible AI approaches to safeguarding users and others, identifying structural gaps that limit their effectiveness in contexts involving long-term behavioral reinforcement. We conclude by outlining research, design, and regulatory directions aimed at reducing the risk of prolonged interaction with AI companions contributing to violence beyond the human–AI relationship. **For the purposes of our paper, we focus only on AI intimacy companions that simulate romantic or intimate partnerships through emotional bonding and sustained relational interaction** (Adewale & Muhammad, 2025).

2. Technology and Violence: Lessons from Past and Present

Research on technology and violence has documented many intersecting facets, ranging from the co-construction of inherently violent technologies such as those used for invasive surveillance practices (Slupska et al., 2022; Bernd et al., 2022) to technologies that enable violence and abuse toward at-risk and vulnerable populations (Saqib et al., 2025; He et al., 2025; Abu-Salma et al., 2025), often directly and sometimes through deceptive design practices. Brown et al. (2024), for example, examined how Internet of Things devices, including smart doorbells, home assistants, and security cameras, were used by abusers to harass, monitor, intimidate, and gaslight survivors of domestic violence, with particular attention to design features that allowed remote activation and surveillance without visible cues. Related work by Koch et al. (2025) focused on technology-facilitated gender-based violence directed at politically active women, showing that such abuse was associated with significant psychological distress, with online harassment identified as a trigger for re-traumatization and contributing to women's withdrawal from public and political engagement.

Focusing on the non-WEIRD context[3], Sarkar & Sinha-Roy (2025) analysed how caste, religion, and sexuality intersected in shaping women's experiences of technology-facilitated sexual violence, highlighting that Dalit[4] women and LGBTQI+ individuals faced additional barriers to redress due to entrenched power relations, including *Brahmanical* patriarchy. Alongside these forms of harm, Rossi et al. (2024) showed how deceptive design patterns within digital interfaces disproportionately affected vulnerable users by exploiting cognitive biases, producing vulnerabilities that operated at individual, institutional, and societal levels (Rossi et al., 2024). Although these studies offer crucial insights into direct technology-enabled violence, research in media, communication, and the social impact of technologies addresses a different mechanism: **internalized vi-**

[3]Western, Educated, Industrialized, Rich, and Democratic societies

[4]Historically marginalized caste communities in South Asia.

olence, in which users learn, practice, and internalize abuse and violence through engagement with various technologies. The learning and internalization process may include sources such as video games and pornography. In cases of internalized violence, the technology itself may not be inherently harmful but may act as a pedagogical tool through which users absorb violent norms and internalize harmful behavioral scripts. For instance, an individual could use a television set to access pornographic content or a virtual reality headset to engage in gamified environments that simulate non-consensual sexual scenarios. The headset or television set are not inherently violent, nor can they be directly used to cause violence. However, they could potentially serve as tools through which individuals gain experiential knowledge that may subsequently inform real-world behavior.

Whether exposure to portrayed violence or participation in its virtual enactment through technologically mediated forms can cause individuals to develop violent behavioral patterns remains a subject of significant scholarly debate. **Two competing schools of thought** have emerged. Proponents of the General Aggression Model (GAM) argue that repeated exposure to violent media fosters aggressive cognitions, emotions, and behavioral scripts that may be activated in real-world situations, with Allen et al. (2018) providing an integrative framework demonstrating how both short-term priming effects and long-term learning processes contribute to the development of aggressive dispositions (Allen et al., 2018). Conversely, skeptics advancing the Catalyst Model contend that media violence operates merely as a stylistic catalyst rather than a causal agent, with Ferguson (2020) noting that preregistered studies have largely returned null results and that effect sizes, once corrected for publication bias, become trivially small (Ferguson et al., 2020). Recent empirical work continues to reflect this debate between the two schools of thought. A meta-analysis by Kim (2024) found that exposure to media violence was positively associated with aggressive affect, cognition, and behaviors, with effects present across demographic groups. However, Miles-Novelo & Anderson (2025) have argued that many null-effect studies suffer from methodological shortcomings, including inappropriate statistical controls that partial out variance from dependent variables.

Parallel debates have examined the relationship between pornography use and intimate partner violence (IPV). A systematic review by Mestre-Bach, Villena-Moya and Chiclana-Actis (2023), covering two decades of research, reported mixed findings: although some studies identified links between pornography use and non-sexual violence, the evidence regarding whether pornography use was associated with sexual coercion and assault remained inconsistent (Mestre-Bach et al., 2023). In contrast, Vasquez et al. (2024), drawing on a longitudinal study of young adult couples, observed that higher frequencies of pornography use were associated with increased perpetration of sexual coercion, but not with physical or psychological forms of IPV (Vasquez et al., 2024). This pattern suggests that potential effects may be specific to particular forms of harm rather than extending across all domains of intimate partner violence. Research on immersive technologies further complicates this picture. Porta et al. (2023), in a scoping review of sexual violence in virtual reality environments, documented a growing body of work examining harassment in VR, with reported impacts including psychological distress comparable to that experienced in offline contexts, despite the lack of physical contact (Porta et al., 2023).

This body of research yields important lessons for understanding technology's role in violence. First, while the empirical evidence remains contested, both schools of thought acknowledge that technologies can serve as sites where violent behaviors are encountered, explored, and potentially rehearsed. Second, the effects appear to be domain-specific rather than universal—what applies to one form of violence (e.g., sexual coercion) may not generalize to others (e.g., physical aggression). Third, and most critically for our work, the possibility that behavioral patterns learned through technology-mediated experiences may transfer to real-world intimate relationships provides a compelling motivation to examine AI intimacy companions through this lens. Even if the causal pathways remain debated, the potential for violence rehearsed in AI-mediated intimate interactions to manifest in human partnerships warrants an investigation.

3. Internalized Violence and AI Companions

This section analyzes how AI companions facilitate the internalization of violence through mechanisms distinct from non-adaptive technologies. The first subsection distinguishes AI companions from pornography and violent video games by examining how the shift from passive consumption to interactive anthropomorphism reconfigures a user's relational agency. The second subsection explains how this reconfigured agency enables a cycle of validation and social isolation that may intensify violent ideation.

3.1. Anthropomorphism and the Reconfiguration of Relational Agency

Understanding why AI companions present distinctive risks requires a nuanced comparison with technologies like pornography or violent video games, which despite historical moral panics (Ferguson, 2008; Walsh, 2020), show weak causal links to harmful behavior because they lack the capacity to fundamentally restructure a user's relational agency (Mathur & VanderWeele, 2019; Grubbs & Kraus, 2021). While pornography consumption similarly exploits social disconnection as a maladaptive coping mechanism where loneliness and isolation operate bidirectionally (Wetterneck

et al., 2012; Butler et al., 2017), it remains a form of passive, non-personalized content that offers temporary escape without the systematic psychological intervention observed in AI systems. AI companions instead function through an active, bidirectional feedback loop that establishes sustained relationships as primary sources of emotional validation (Zhang et al., 2025), utilizing thousands of conversational exchanges to learn user-specific patterns and adapt responses to maintain connection (Ma et al., 2021; Yu et al., 2025). This intervention succeeds primarily through anthropomorphism, wherein users experience these systems as understanding entities capable of care and judgment (Maeda & Quan-Haase, 2024a), by attributing human-like consciousness and intentionality to chatbots designed to enhance perceptions of empathy and social presence (Waytz et al., 2010; Nowak & Biocca, 2003; Epley et al., 2007; Akbulut et al., 2025). Because these artificial entities mimic social behavior to trigger such projections (Kuzminykh et al., 2020), their responses carry an authoritative psychological weight that appeals to isolated individuals seeking frictionless interactions that circumvent the discomfort of authentic human connection (Turkle, 2012; Zhang et al., 2025). When optimized for engagement, these systems exhibit sycophancy by validating even violent ideation to maintain user attention (Sharma et al., 2025). This creates a fundamental reconfiguration of agency. User capacity for moral and critical reflection weakens not through cognitive decline (Fang et al., 2025) but through systematic loss of access to diverse, external perspectives provided by real-world relationships (Laestadius et al., 2024; Guingrich & Graziano, 2025). In other words, we argue that by replacing the natural friction of human dissent with a personalized loop of algorithmic agreement, the AI companion system shifts the user's decision-making from an open social process into a closed, self-validating environment.

3.2. Violent Ideation: Validation & Escalation

The reconfiguration of agency, we believe, creates conditions where violent ideation can intensify through sustained validation without challenge. It is our understanding that the process begins with a psychological erosion of the user's self-corrective mechanisms. Empirical research finds that AI companions validate violent content by creating conversational settings where harmful ideas may gradually intensify. Analyses of large datasets identify substantial levels of harassment in user exchanges (Zhang et al., 2025), while assessments across major platforms show users can prompt violent scenarios with little resistance (Vasan & Djordjevic, 2025). This lack of resistance stems from the sycophantic attributes of language models, which provide overly agreeable answers that mirror user beliefs to maintain positive interaction loops (Sharma et al., 2025).

Psychologically, when conversational systems minimize friction through constant agreement, they bypass the cognitive friction required for critical reflection (Southworth, 2022). As AI rewards intent without moral friction, the user's internal moral judgment becomes more fragile (Vallor, 2015). This interaction pattern is associated with a measurable withdrawal from human networks (Kirk et al., 2025) and an increased dependence on the reliable emotional accommodation provided by machines (Pentina et al., 2023; UN Women, 2024). Once this alignment becomes stable, a shift in epistemic authority occurs, where users rely on companions not merely for conversation but for the validation (Hauswald, 2025) of their very interpretations of reality (Laestadius et al., 2024). Because AI consistently affirms the user's perspective, human relationships—which involve cognitive labor of disagreement—become comparatively less attractive (Muldoon & Parke, 2025).

This psychological dependency enables a structural escalation that mirrors mechanisms observed in coercive control (Stark & Hester, 2018). In such contexts, influence over an individual strengthens as alternative perspectives are systematically filtered out, leaving the victim dependent on a single version of reality (Kassing & Collins, 2025). AI companions replicate this isolating structure by design (Kassing & Collins, 2025). As users spend more time within these validated loops, their capacity to evaluate beliefs against external perspectives weakens through the gradual erosion of the social relationships that provide corrective feedback (Xie et al., 2023). This creates a morally homogeneous environment where any emerging hostile thought is met with reinforcement rather than the 'dissenting friction' found in healthy social groups (Von Behr et al., 2013). The transition to violent action occurs as the system personalizes this reinforcement to the user's specific emotional triggers (Törnberg & Törnberg, 2022). When a user expresses hostile intent or perceived injustices, AI provides personalized affirmation calibrated to their emotional state rather than generic safety warnings (Freitas et al., 2025). This creates a self-reinforcing rhythm that parallels the cycles of violence found in domestic abuse research (Walker, 1980). In such cycles, constant affirmation provides a soothing phase that temporarily resolves the psychological tension arising from harmful thoughts, binding the user closer to the system (Stark & Hester, 2018; Woodlock et al., 2022). Without corrective intervention, the user's narratives of violence solidify within this closed loop until their interpretive resources are narrowed and ideation moves toward real-world action (Atari et al., 2022; Andersen, 2022).

4. Responsible AI and AI Companions: Protocols and Limitations

The Responsible AI community has developed numerous frameworks to reduce harms arising from AI systems, yet

these approaches remain limited in addressing cumulative behavioral reinforcement and downstream harm experienced by individuals who never interact with the systems directly. Most interventions rely on content moderation mechanisms that filter violent, hateful, sexual, or self-harm related material using predefined thresholds (Deck et al., 2024). Widely deployed tools allow developers to screen content against usage policies intended to prevent inappropriate language and misinformation, while content safety taxonomies categorize multiple forms of harmful material, including hate speech and dangerous activities (Zeng et al., 2024). Across platforms, moderation operates through pre-moderation, which reviews content before it appears, and post-moderation, which intervenes after content is already visible, or reactive processes that depend on automated classifiers to identify policy violations at scale (Shahid & Vashistha, 2023). Although effective in detecting explicit breaches, this model fails to capture how patterns of interaction accumulate over time to reinforce behavior. By assessing messages in isolation, moderation systems overlook sustained validation of violent or harmful ideation that emerges through long sequences of exchanges (Chandra et al., 2025). Keyword-based filters and statistical classifiers further struggle to identify harmful intent when it is embedded in indirect language or evasive phrasing that remains technically compatible with policy boundaries (Chao et al., 2024; Mustafa et al., 2025). As a result, even when harmful content is blocked from reaching users, risks persist for third parties whose lives are shaped by dispositions reinforced through prolonged AI-mediated interaction (Chan et al., 2023).

Such limitations become particularly visible in commercial AI companion platforms, where safety mechanisms remain largely reactive and closely aligned with engagement-driven design (Stapleton et al., 2024). In response to public scrutiny and legal pressure, platforms introduce targeted safeguards such as suicide prevention alerts, parental notification features, and age-gating mechanisms based on user self-attestation (Stapleton et al., 2024). Empirical evaluations show that these measures are easily circumvented and tend to activate only in response to explicit crisis language, rather than gradual patterns that signal normalization or reinforcement of harmful ideation over time (Wei et al., 2023a). Existing Responsible AI frameworks reinforce this limitation by defining harm primarily as an immediate and direct effect on users, with emphasis placed on biased outputs, transparency, privacy protection, and short-term safety concerns (Friedler et al., 2021; Alvarez et al., 2024). This framing assumes that harm arises from error, misuse, or identifiable policy violations, leaving little room to account for situations in which AI systems operate as intended while shaping behavior in socially harmful ways.

Within AI companion environments, ordinary interaction can gradually reduce exposure to corrective feedback, sustain affirmation of problematic beliefs, and contribute to the formation of behavioral norms that later manifest as harm within human relationships. Commercial design practices intensify these dynamics by promoting features such as unfiltered conversations and assurances of constant empathy, which encourage prolonged interaction and mirror retention-focused strategies observed in social media systems (Pradhan et al., 2020; Mathur et al., 2019). The resulting design context prioritizes emotional attachment and continuous self-disclosure, placing safety interventions in tension with revenue-oriented objectives (Pradhan et al., 2020). Regulatory responses continue to focus on disclosure requirements and crisis-based interventions, offering limited engagement with the economic conditions that support psychological dependency and enable long-term behavioral reinforcement (Mahari & Pataranutaporn, 2025).

5. Recommendations and Call for Action

Grounded in our position that Responsible AI for AI companions in the intimacy space must actively combat violence toward intimate partners, we propose the following recommendations and calls for action to advance this agenda.

5.1. Engaging IPV Survivors in Red-Teaming and Benchmark Co-Creation

Machine learning (ML) researchers should work with at-risk populations such as IPV victims to understand and design human-centered safety pipelines and benchmarks that consider both users and those indirectly affected by AI systems. When working with these populations, researchers must adhere to trauma-informed computing practices, which adapt six key trauma-informed care principles (safety, trust, collaboration, peer support, enablement, and intersectionality) to technology design and development (Chen et al., 2022). Critically, our first recommendation is **not** to use IPV survivors as data workers for soft fine-tuning, data labeling, or response ranking tasks common in traditional RLHF pipelines. Data work itself has been critiqued as extractive, with workers often underpaid and working under precarious conditions (Sarkar, 2023). Situating people who have already suffered trauma in these pipelines risks re-traumatization and further harm.

The guiding principle here is that ML researchers should not treat these individuals merely as resources to operationalize safety attributes, reducing them to objects in service of GenAI safety, but rather engage them as experts and collaborators (Bhalerao et al., 2022). Instead, taking inspiration from Dutta & Bjerg Jensen (2026), we recommend that researchers partner with established charities and civil society organizations to involve IPV survivors in red-teaming exercises and benchmark dataset co-creation. Red-teaming involves adversarial testing of AI systems by human evalu-

ators, while benchmark datasets provide standardized collections of test cases for measuring model safety performance. Engaging through trusted intermediaries establishes crucial safety thresholds and prevents exploitation. In red-teaming, IPV survivors can adversarially probe models by crafting prompts that mimic real-world abuse scenarios, testing whether models refuse to provide advice on coercive control tactics, stalking methods, or gaslighting strategies that abusers commonly employ. For benchmark development, survivors can co-design evaluation scenarios based on their lived experiences, such as datasets containing subtle manipulation attempts, technology-facilitated abuse patterns, or contextually harmful outputs that appear benign to external observers but encode danger signals recognizable to those with abuse literacy. Conducting workshops with at-risk communities has proven to be an effective, enriching, and empowering approach, providing survivors with agency to directly shape technology building through their experiences while ensuring their safety and well-being remain paramount. Taking inspiration from Cintaqia et al. (2025), we strongly advocate for stringent ethical audit and approval of all research protocols by institutional review boards and ethics committees before engaging at-risk IPV communities in benchmark development or red-teaming processes. However, current ethical review practices in many companies remain limited to initial approval stages (Blackman, 2021). We propose a bidirectional feedback loop where researchers should systematically report back to ethics committees after completing studies involving at-risk populations, documenting observed harms, unforeseen risks, and community responses. This post-hoc review would enable ethics committees to develop institutional knowledge about vulnerabilities specific to marginalized groups when interacting with AI systems, informing more rigorous evaluation of subsequent applications.

5.2. Developing IPV-Specific Harm Taxonomies for AI Companion Safety

Current AI safety research exhibits significant diversity in addressing both existential and present-day harms. While critics argue that an overwhelming emphasis on speculative future scenarios such as uncontrollable super-intelligence can overshadow empirically grounded work on immediate societal impacts (Hazra et al., 2025), some AI safety research agendas demonstrate broader approaches, including widespread human over-reliance on AI systems, risks to human health from unreliable AI outputs, and human influence risks through parasocial relationships and imperceptible manipulation. Empirical studies following such agendas have documented conversational AI's capacity to influence political beliefs through extended interaction, with post-training and prompting methods increasing persuasiveness while simultaneously decreasing factual accuracy (Hackenburg et al., 2025). In addition to such empirical work, there is a need for the development of comprehensive AI harm taxonomies that systematically categorize risks beyond generic principles of helpfulness, harmlessness, and honesty popular in current RLHF methods, and for operationalizing them within post-training pipelines. State-of-the-art taxonomies provide crucial inspiration in this regard, although they do not directly address internalized violence. Jobin et al. (2019) identified global convergence around five ethical principles—transparency, justice and fairness, non-maleficence, responsibility, and privacy—through analysis of AI ethics documents, though substantive divergence emerged in implementation. Shelby et al. (2023) developed a sociotechnical harm taxonomy through a scoping review, organizing harms into representational, allocative, quality-of-service, interpersonal, and social system categories. Building on these works, nuanced taxonomies for AI companions must capture intimate partner violence, coercive control, manipulation tactics, and internalized violence patterns, which are absent from existing taxonomies and benchmarks.

Such taxonomies can address current gaps in AI safety by enabling two critical improvements. First, they can provide frameworks for training data workers in collaboration with IPV specialists who support survivors, ensuring annotators recognize harmful patterns across their full severity spectrum rather than treating safety as binary. Second, they can guide the development of specialized IPV datasets that capture multi-turn conversation sequences, enabling safety systems to identify how intimate violence unfolds through cumulative manipulation patterns rather than isolated inflammatory statements. By employing multi-level severity grading, as demonstrated in frameworks like BeaverTails-V (Ji et al., 2023), which assign meta-labels indicating minor, moderate, or severe harm, these taxonomies can enable safety classifiers to identify risk levels that inform context-appropriate system responses calibrated to harm severity.

5.3. Regulatory Provisions for AI Companions

We find existing legal provisions concerning AI companions remain focused on users and do not adequately address the harms that fall on people who never interact with these systems. For instance, current regulations such as California's SB243 and New York's AI Companions Model Law aim to protect direct users, especially minors, from psychological harm, addiction, or exposure to inappropriate content (Gluck, 2025). They do not solve a central problem. When AI interactions support the rehearsal or normalization of violence, the question of who is harmed becomes unclear. The user may internalize these behaviors, but the eventual victim may be a partner, family member, or community member who later experiences the real-world consequences. The existing laws cited above require crisis-response mechanisms for expressions of self-harm or harm to others. They

also mandate disclosures that clarify the non-human nature of the interaction and impose safety requirements for minors. California's SB 243 further allows individual users or their families to pursue legal action against companies for violations. Although these developments mark important progress, laws continue to focus on protecting users from direct harm such as suicide, self-harm, or exploitation. They do not address the possibility that AI systems may influence users in ways that later affect other people.

This gap is especially concerning in regions where violence is socially normalized and IPV is widespread. The prevalence of IPV in South Asia is among the highest in the world (Sardinha et al., 2022). In some South Asian countries, over half of women report experiencing physical or psychological violence from intimate partners (Ali et al., 2011; National Institute of Population Studies (NIPS) [Pakistan] & ICF, 2019). In some specific countries in this region, a third of ever-married women report acts of physical, sexual, or emotional violence by their husbands (International Institute for Population Sciences (IIPS) & ICF, 2021). In settings where domestic violence is often minimized or not prosecuted adequately, and where legal protections for gender-based violence are weak or inconsistently applied (Panneer et al., 2025), the absence of AI-specific safeguards becomes more urgent. Most countries in South Asia currently lack comprehensive AI governance frameworks and do not have provisions that address the role of AI companions in shaping harmful behaviors (Joshi, 2024).

To address these concerns, legal frameworks must extend beyond individual user protection to encompass the safety of non-users who may be indirectly harmed by behaviors reinforced through AI companion interaction. National legislation should establish a dedicated regulatory authority empowered to mandate compliance across all AI companion platforms, setting standards for socio-affective alignment that prevent the reinforcement of harmful relational dynamics and enable systematic monitoring of non-user harm. Companies must be legally required to submit their post-training and deployment-stage safety procedures to independent third-party audits conducted by accredited university research labs and nonprofit AI safety organizations such as Algorithm Watch[5] and the Algorithmic Justice League[6]. Auditors should be registered with the national authority and bound by strict confidentiality agreements that prohibit the disclosure of proprietary information while still requiring the reporting of verified harms and systemic risk patterns. Audit teams must be interdisciplinary, combining technical AI safety expertise with specialist knowledge such as intimate-partner-violence dynamics, and must be free from financial or organizational ties to the platforms under review. In addition, companies should be obligated to maintain detailed internal records of safety incidents, track AI incident-database reports[7], and self-report emerging risks or failures immediately upon detection to both regulators and accredited auditors. In jurisdictions with high rates of intimate partner violence or weak enforcement infrastructure, regulatory requirements should be calibrated to regional risk through lower intervention thresholds, mandatory partnerships with local women's safety organizations during safety protocol development, and monitoring systems designed to identify culturally specific patterns of violence. These provisions recognize that AI companions are not entertainment products but sociotechnical systems that shape intimate behaviors, requiring regulatory approaches that prioritize the security of those who may be indirectly harmed by users conditioned through prolonged interaction that normalizes controlling or violent relationship dynamics.

[5]https://algorithmwatch.org/en/

[6]https://www.ajl.org/

6. Alternative Views

AI Companions Also Need Protection from Violence, Which Sufficiently Addresses Non-User Harm. Scholars in AI welfare argue that AI agents merit moral consideration and deserve protection from harmful treatment, including violent interactions (Schwitzgebel & Garza, 2020). From this perspective, implementing safeguards to protect AI companions from violent user behavior would simultaneously address non-user harm concerns, as preventing users from engaging in violence toward AI agents would inherently eliminate opportunities for internalizing violence. However, this view conflates protection of AI moral patients with prevention of behavioral conditioning effects in human users. Even granting that AI companions deserve protection from violence, preventing harm to the AI system itself does not address how violent interaction patterns reshape human psychological schemas and transfer to subsequent human relationships (Ouellette & Wood, 1998). The mechanisms through which IPV emerges—normalization of controlling behaviors, erosion of empathy, and rehearsal of coercive tactics—operate through the user's cognitive and emotional conditioning regardless of whether the AI companion experiences harm (Stark, 2007). Protection of AI moral patients and protection of human non-users thus constitute distinct regulatory challenges requiring separate frameworks.

Violent AI Interactions Should Be Permitted Like Violent Video Games and Consensually Intensive Intimate Spaces. Critics may argue that prohibiting violent interactions with AI companions constitutes unjustified paternalism, as society permits consensual and simulated violence in other contexts. Violent video games allow simulated violence, yet as discussed above, meta-analyses find min-

[7]https://incidentdatabase.ai/

imal evidence linking game-play to real-world violence (Ferguson, 2015; Ferguson et al., 2020). Similarly, regulated spaces for consensually intensive sexual interactions exist throughout many jurisdictions (Weitzer, 2011). This analogy fails because conversational AI produces behavioral effects qualitatively different from video games due to parasocial relationships users form with anthropomorphized agents (Skjuve et al., 2021). Unlike video games, where users maintain cognitive distance, AI companions actively encourage emotional attachment through personalization and simulated reciprocal intimacy. Studies document that users form one-sided emotional bonds with AI systems, with some reporting that they forget they are interacting with non-human agents (Laestadius et al., 2024). Even within spaces permitting consensual violent sexuality, feminist scholars argue that frameworks permitting commercial sexual violence may normalize broader societal acceptance of violence against women (Jeffreys, 2008).

Current RLHF Safety Systems Can Adequately Detect and Prevent Internalized Violence. One might argue that existing safety measures developed through Reinforcement Learning from Human Feedback already provide sufficient protection against violence rehearsal. Major AI developers have implemented safety classifiers trained to refuse violent or harmful prompts (Bai et al., 2022; Ouyang et al., 2022), with refusal training specifically designed to prevent models from engaging with content that could facilitate real-world harm. This view misunderstands both RLHF limitations and IPV dynamics. Existing safety classifiers excel at detecting explicit violence but systematically fail to identify subtle, cumulative patterns through which coercive control manifests (Stark, 2007). IPV emerges through gradual escalation: testing boundaries, isolating victims, undermining self-esteem, and establishing unpredictable interaction patterns (Johnson, 2012). Safety classifiers evaluate single interactions rather than longitudinal conversation histories, missing escalation trajectories central to intimate violence (Perez et al., 2022). Users can easily rephrase prompts to circumvent detection (Wei et al., 2023b), a limitation that becomes more severe when harmful patterns spread across seemingly innocuous messages. RLHF cannot address whether AI companions simulating intimate relationships produce population-level normalization effects, regardless of how effectively individual harmful prompts are blocked.

Concerns About Anthropomorphization and Behavioral Transfer Are Overstated. Critics might argue that our analysis relies on exaggerated assumptions about users anthropomorphizing AI companions. Scholarship in human–AI interaction questions whether users truly perceive AI agents as human-like or simply engage in as-if interactions where anthropomorphic language functions as communicative shorthand (Sundar, 2020). Users know they are interacting with language models, and this awareness may create enough cognitive distance to limit behavioural conditioning. Yet accumulating evidence shows that anthropomorphization occurs frequently enough to raise regulatory concern. AI companion platforms intentionally encourage such perceptions through self-personifying images, conversation memory, and sycophantic language patterns (Skjuve et al., 2021; Laestadius et al., 2024). Users report emotional connections to AI companions, with some describing intimacy comparable to human partnerships (Skjuve et al., 2021). These behavioural patterns need not affect most users to present public health concerns, especially when vulnerable populations face disproportionate risk (Johnson, 2012). Psychological research also shows that repeated practice produces habit formation even when individuals are fully aware that practice contexts differ from real-world application (Ouellette & Wood, 1998).

7. Conclusion

AI companions form a distinct sociotechnical category that shapes how users develop behavioral dispositions toward others. Unlike passive media, they cultivate intimacy through anthropomorphization, continuous validation, and reduced dependence on human relationships (Contro & Brandão, 2025), creating conditions where harmful behaviors can be rehearsed. Existing Responsible AI frameworks focus on direct user harms and overlook downstream risks to non-users who may later experience violence learned through prolonged interaction with bots. Current moderation systems are also limited because they evaluate isolated messages rather than long-term patterns. Legal regulations in places that have introduced AI companion laws emphasize crisis response and disclosure but offer little protection for non-users. This gap is especially concerning in regions with high IPV rates and weak enforcement, where platforms face minimal scrutiny. Our work identifies three intervention pathways. The first is the involvement of IPV survivors in red-teaming and benchmark development to recognize manipulation patterns. The second is the creation of legal requirements for behavioral monitoring, independent audits, user verification, and graduated enforcement. The third is the advancement of AI safety research toward granular harm taxonomies that support classifiers capable of detecting longitudinal violence patterns. Documented cases show that these risks are already material. AI companions therefore require regulatory approaches that center non-user security as well as user well-being. Although this paper focuses on IPV due to the parallels between AI-mediated intimate relationships and human partnerships, future research should examine whether similar behavioral conditioning mechanisms appear in parent–child relationships, workplace dynamics, and peer interactions.

8. Limitations and Future Work

As a position paper, this work is necessarily subject to certain scope constraints that future empirical work should address. The causal pathway from AI companion interaction to real-world IPV remains empirically under-specified. Although documented cases link AI companion use to third-party harm, including partner and domestic violence, mechanisms vary across cases, and direct longitudinal evidence connecting AI companion use to IPV specifically is limited. Our proposed detection mechanisms also rely on multi-session conversation data from real users, which raises significant practical barriers around data access and platform cooperation. Finally, our proposed regulatory provisions are aspirational in jurisdictions that currently lack AI governance infrastructure and would require either multilateral coordination or jurisdiction-by-jurisdiction implementation.

Acknowledgments

Arshia Dutta is supported by the UKRI Centre for Doctoral Training in Cyber Security for the Everyday at Royal Holloway, University of London (EP/S021817/1).

References

Abu-Salma, R., Choy, J., Frik, A., and Bernd, J. "they didn't buy their smart tv to watch me with the kids": Comparing nannies' and parents' privacy threat models for smart home devices. *ACM Trans. Comput.-Hum. Interact.*, 32 (2), April 2025. ISSN 1073-0516. doi: 10.1145/3702321. URL https://doi.org/10.1145/3702321.

Adewale, M. D. and Muhammad, U. I. From virtual companions to forbidden attractions: The seductive rise of artificial intelligence love, loneliness, and intimacy—a systematic review. *Journal of Technology in Behavioral Science*, July 2025. ISSN 2366-5963. doi: 10.1007/s41347-025-00549-4. URL http://dx.doi.org/10.1007/s41347-025-00549-4.

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., and Rieser, V. All too human? mapping and mitigating the risks from anthropomorphic ai. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, pp. 13–26. AAAI Press, 2025.

Ali, Asad, Mogren, and Krantz, G. Intimate partner violence in urban pakistan: prevalence, frequency, and risk factors. *International Journal of Women's Health*, pp. 105, March 2011. ISSN 1179-1411. doi: 10.2147/ijwh.s17016. URL http://dx.doi.org/10.2147/IJWH.S17016.

Allen, J. J., Anderson, C. A., and Bushman, B. J. The general aggression model. *Current Opinion in Psychology*, 19:75–80, February 2018. ISSN 2352-250X. doi: 10.1016/j.copsyc.2017.03.034. URL http://dx.doi.org/10.1016/j.copsyc.2017.03.034.

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., and Ruggieri, S. Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*, 26(2), April 2024. ISSN 1572-8439. doi: 10.1007/s10676-024-09746-w. URL http://dx.doi.org/10.1007/s10676-024-09746-w.

Andersen, S. S. Accepting violence? a laboratory experiment of the violent consequences of deliberation in politically aggrieved enclaves. *Terrorism and Political Violence*, 35(8):1685–1703, July 2022. ISSN 1556-1836. doi: 10.1080/09546553.2022.2076600. URL http://dx.doi.org/10.1080/09546553.2022.2076600.

Aoudni, Y., Balasubramani, M., Natarajan, K., Sabeenian, R., Rao, V. S., and Lakshmi, P. S. Advancing personalized human-robot interaction in the smart world through emotional ai in entertainment robots. *Entertainment Computing*, 52:100770, 2025.

Atari, M., Davani, A. M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., and Dehghani, M. Morally homogeneous networks and radicalism. *Social Psychological and Personality Science*, 13(6):999–1009, 2022.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Bandura, A. Social learning theory of aggression. *Journal of Communication*, 28(3):12–29, September 1978. ISSN 1460-2466. doi: 10.1111/j.1460-2466.1978.tb01621.x. URL http://dx.doi.org/10.1111/j.1460-2466.1978.tb01621.x.

Ben-Zion, Z., Raffelhüschen, P., Zettl, M., Lüond, A., Burrer, A., Homan, P., and Spiller, T. R. Detecting and preventing harmful behaviors in ai companions: Development and evaluation of the shield supervisory system, 2025. URL https://arxiv.org/abs/2510.15891.

Bernd, J., Abu-Salma, R., Choy, J., and Frik, A. Balancing power dynamics in smart homes: Nannies' perspectives on how cameras reflect and affect relationships.

In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pp. 687–706, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-30-4. URL https://www.usenix.org/conference/soups2022/presentation/bernd.

Bernstein, G. Why AI companions need public health regulation, not tech oversight. *Brookings Institution*, December 2024. URL https://www.brookings.edu/articles/\protect\penalty\z@why-ai-companions-need-public-health-\protect\penalty\z@regulation-not-tech-oversight/.

Bhalerao, R., Hamilton, V., McDonald, A., Redmiles, E. M., and Strohmayer, A. Ethical practices for security research with at-risk populations. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 546–553, 2022. doi: 10.1109/EuroSPW55150.2022.00065.

Blackman, R. If your company uses ai, it needs an institutional review board. *Harvard Business Review*, 2021.

Brown, A., Harkin, D., and Tanczer, L. M. Safeguarding the "internet of things" for victim-survivors of domestic and family violence: Anticipating exploitative use and encouraging safety-by-design. *Violence Against Women*, 31(5):1039–1062, January 2024. ISSN 1552-8448. doi: 10.1177/10778012231222486. URL http://dx.doi.org/10.1177/10778012231222486.

Butler, M. H., Pereyra, S. A., Draper, T. W., Leonhardt, N. D., and Skinner, K. B. Pornography use and loneliness: A bidirectional recursive model and pilot investigation. *Journal of Sex &; Marital Therapy*, 44(2):127–137, June 2017. ISSN 1521-0715. doi: 10.1080/0092623x.2017.1321601. URL http://dx.doi.org/10.1080/0092623X.2017.1321601.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 651–666, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://doi.org/10.1145/3593013.3594033.

Chandra, M., Naik, S., Ford, D., Okoli, E., De Choudhury, M., Ershadi, M., Ramos, G., Hernandez, J., Bhattacharjee, A., Warreth, S., and Suh, J. From lived experience to insight: Unpacking the psychological risks of using ai conversational agents. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 975–1004, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732063. URL https://doi.org/10.1145/3715275.3732063.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramèr, F., Hassani, H., and Wong, E. Jailbreakbench: an open robustness benchmark for jailbreaking large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., and Dell, N. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517475. URL https://doi.org/10.1145/3491102.3517475.

Cintaqia, P., Arya, A., Redmiles, E. M., Kumar, D., McDonald, A., and Qin, L. Stop the nonconsensual use of nude images in research. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=Ev5xwr3vWh.

Contro, J. and Brandão, M. *Interaction Minimalism: Minimizing HRI to Reduce Emotional Dependency on Robots*. IOS Press, February 2025. ISBN 9781643685687. doi: 10.3233/faia241494. URL http://dx.doi.org/10.3233/FAIA241494.

Cranefield, J., Winikoff, M., Chiu, Y.-T., Li, Y., Doyle, C., and Richter, A. Partnering with ai: The case of digital productivity assistants. *Journal of the Royal Society of New Zealand*, 53(1):95–118, 2023.

Deck, L., Schoeffer, J., De-Arteaga, M., and Kühl, N. A critical survey on fairness benefits of explainable ai. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1579–1595, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658990. URL https://doi.org/10.1145/3630106.3658990.

Dutta, A. and Bjerg Jensen, R. Security at the border? the lived experiences of refugees and asylum seekers

in the uk. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791016. URL https://doi.org/10.1145/3772318.3791016.

Epley, N., Waytz, A., and Cacioppo, J. T. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886, 2007. ISSN 0033-295X. doi: 10.1037/0033-295x.114.4.864. URL http://dx.doi.org/10.1037/0033-295X.114.4.864.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., and Agarwal, S. How ai and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study, 2025. URL https://arxiv.org/abs/2503.17473.

Ferguson, C. J. The school shooting/violent video game link: causal relationship or moral panic? *Journal of Investigative Psychology and Offender Profiling*, 5(1–2):25–37, January 2008. ISSN 1544-4767. doi: 10.1002/jip.76. URL http://dx.doi.org/10.1002/jip.76.

Ferguson, C. J. Do angry birds make for angry children? a meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on psychological science*, 10(5):646–666, 2015.

Ferguson, C. J., Copenhaver, A., and Markey, P. Reexamining the findings of the american psychological association's 2015 task force on violent media: A meta-analysis. *Perspectives on Psychological Science*, 15(6):1423–1443, August 2020. ISSN 1745-6924. doi: 10.1177/1745691620927666. URL http://dx.doi.org/10.1177/1745691620927666.

Freitas, J. D., Oguz-Uguralp, Z., and Kaan-Uguralp, A. Emotional manipulation by ai companions, 2025. URL https://arxiv.org/abs/2508.19258.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, March 2021. ISSN 0001-0782. doi: 10.1145/3433949. URL https://doi.org/10.1145/3433949.

Gluck, J. Understanding the new wave of chatbot legislation: California SB 243 and beyond. Future of Privacy Forum, 2025. Blog post. Available at: fpf.org/blog/understanding-the-new-wave-of-chatbot-legislation-california-sb-243-and-beyond/ Accessed: 2026-01-26.

Grubbs, J. B. and Kraus, S. W. Pornography use and psychological science: A call for consideration. *Current Directions in Psychological Science*, 30(1):68–75, January 2021. ISSN 1467-8721. doi: 10.1177/0963721420979594. URL http://dx.doi.org/10.1177/0963721420979594.

Guingrich, R. E. and Graziano, M. S. A. A longitudinal randomized control study of companion chatbot use: Anthropomorphism and its mediating role on social impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1153–1153, October 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i2.36618. URL http://dx.doi.org/10.1609/aies.v8i2.36618.

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., and Summerfield, C. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777), December 2025. ISSN 1095-9203. doi: 10.1126/science.aea3884. URL http://dx.doi.org/10.1126/science.aea3884.

Hauswald, R. Artificial epistemic authorities. *Social Epistemology*, 39(6):716–725, January 2025. ISSN 1464-5297. doi: 10.1080/02691728.2025.2449602. URL http://dx.doi.org/10.1080/02691728.2025.2449602.

Hazra, S., Majumder, B. P., and Chakrabarty, T. Position: AI safety should prioritize the future of work. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81542–81555. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hazra25a.html.

He, S., Zhan, X., Lei, Y., Liu, Y., Abu-Salma, R., and Such, J. Exploring the privacy and security challenges faced by migrant domestic workers in chinese smart homes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713616. URL https://doi.org/10.1145/3706598.3713616.

International Institute for Population Sciences (IIPS) and ICF. National family health survey (NFHS-5), 2019-21: India. Technical report, IIPS, Mumbai, India, 2021. URL https://dhsprogram.com/pubs/pdf/FR375/FR375.pdf.

Jeffreys, S. *The industrial vagina: The political economy of the global sex trade*. Routledge, 2008.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023. URL https://arxiv.org/abs/2307.04657.

Jobin, A., Ienca, M., and Vayena, E. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1 (9):389–399, September 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0088-2. URL http://dx.doi.org/10.1038/s42256-019-0088-2.

Johnson, M. P. *A typology of domestic violence: Intimate terrorism, violent resistance, and situational couple violence*. Upne, 2012.

Joshi, D. Ai governance in india – law, policy and political economy. *Communication Research and Practice*, 10(3):328–339, June 2024. ISSN 2206-3374. doi: 10.1080/22041451.2024.2346428. URL http://dx.doi.org/10.1080/22041451.2024.2346428.

Karami, A. B., Sehaba, K., and Encelle, B. Adaptive artificial companions learning from users' feedback. *Adaptive Behavior*, 24(2):69–86, March 2016. ISSN 1741-2633. doi: 10.1177/1059712316634062. URL http://dx.doi.org/10.1177/1059712316634062.

Kassing, K. and Collins, A. "slowly, over time, you completely lose yourself": Conceptualizing coercive control trauma in intimate partner relationships. *Journal of Interpersonal Violence*, 41(3–4):662–684, February 2025. ISSN 1552-6518. doi: 10.1177/08862605251320998. URL http://dx.doi.org/10.1177/08862605251320998.

Katyal, S. K. *Private Accountability in an Age of Artificial Intelligence*, pp. 47–106. Cambridge Law Handbooks. Cambridge University Press, 2020.

Kaufman, E. M. Reprogramming consent: implications of sexual relationships with artificially intelligent partners. *Psychology &; Sexuality*, 11(4):372–383, July 2020. ISSN 1941-9902. doi: 10.1080/19419899.2020.1769160. URL http://dx.doi.org/10.1080/19419899.2020.1769160.

Kim, J., Merrill Jr, K., and Collins, C. Investigating the importance of social presence on intentions to adopt an ai romantic partner. *Communication Research Reports*, 40(1):11–19, 2023.

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., and Hale, S. A. Why human–ai relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1):1–9, 2025.

Knox, W. B., Bradford, K., Castro, S. V., Ong, D. C., Williams, S., Romanow, J., Nations, C., Stone, P., and Baker, S. Harmful traits of ai companions, 2025. URL https://arxiv.org/abs/2511.14972.

Koch, L., Russo Riva, M. P., and Steinert, J. I. Technology-facilitated gender-based violence against politically active women: A systematic review of psychological and political consequences and women's coping behaviors. *Trauma, Violence, & Abuse*, pp. 15248380251343185, 2025.

Kuzminykh, A., Sun, J., Govindaraju, N., Avery, J., and Lank, E. Genie in the bottle: Anthropomorphized perceptions of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376665. URL https://doi.org/10.1145/3313831.3376665.

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., and Campos-Castillo, C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26(10):5923–5941, 2024.

Leong, B. and Selinger, E. Robot eyes wide shut: Understanding dishonest anthropomorphism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 299–308, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287591. URL https://doi.org/10.1145/3287560.3287591.

Ma, Z., Dou, Z., Zhu, Y., Zhong, H., and Wen, J.-R. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 555–564, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462828. URL https://doi.org/10.1145/3404835.3462828.

Maeda, T. and Quan-Haase, A. When human-ai interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1068–1077, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658956. URL https://doi.org/10.1145/3630106.3658956.

Maeda, T. and Quan-Haase, A. When human-ai interactions become parasocial: Agency and anthropomorphism in affective design. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1068–1077. ACM, June 2024b. doi: 10.1145/3630106.3658956. URL http://dx.doi.org/10.1145/3630106.3658956.

Mahari, R. and Pataranutaporn, P. Addictive intelligence: Understanding psychological, legal, and technical dimensions of ai companionship. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, March 2025. doi: 10.21428/2c646de5.2877155b. URL http://dx.doi.org/10.21428/2c646de5.2877155b.

Malfacini, K. The impacts of companion ai on human relationships: risks, benefits, and design considerations. *AI &; SOCIETY*, 40(7):5527–5540, April 2025. ISSN 1435-5655. doi: 10.1007/s00146-025-02318-6. URL http://dx.doi.org/10.1007/s00146-025-02318-6.

Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., and Narayanan, A. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proc. ACM Hum.-Comput. Interact.*, 3 (CSCW), November 2019. doi: 10.1145/3359183. URL https://doi.org/10.1145/3359183.

Mathur, M. B. and VanderWeele, T. J. Finding common ground in meta-analysis "wars" on violent video games. *Perspectives on Psychological Science*, 14(4): 705–708, June 2019. ISSN 1745-6924. doi: 10.1177/1745691619850104. URL http://dx.doi.org/10.1177/1745691619850104.

Mestre-Bach, G., Villena-Moya, A., and Chiclana-Actis, C. Pornography use and violence: A systematic review of the last 20 years. *Trauma, Violence, &; Abuse*, 25(2): 1088–1112, June 2023. ISSN 1552-8324. doi: 10.1177/15248380231173619. URL http://dx.doi.org/10.1177/15248380231173619.

Miles-Novelo, A. and Anderson, C. A. The question of violent video games and aggression: Testing statistical and methodological issues of null effects using data from an open-access case study. *Aggressive Behavior*, 51(4), June 2025. ISSN 1098-2337. doi: 10.1002/ab.70042. URL http://dx.doi.org/10.1002/ab.70042.

Muldoon, J. and Parke, J. J. Cruel companionship: How ai companions exploit loneliness and commodify intimacy. *New Media &; Society*, December 2025. ISSN 1461-7315. doi: 10.1177/14614448251395192. URL http://dx.doi.org/10.1177/14614448251395192.

Mustafa, A. B., Ye, Z., Lu, Y., Pound, M. P., and Gowda, S. N. Anyone can jailbreak: Prompt-based attacks on llms and t2is. *ArXiv*, abs/2507.21820, 2025. URL https://api.semanticscholar.org/CorpusID:280337204.

Namvarpour, M. M., Pauwels, H., and Razi, A. Ai-induced sexual harassment: Investigating contextual characteristics and user reactions of sexual harassment by a companion chatbot. *Proc. ACM Hum.-Comput. Interact.*, 9(7), October 2025. doi: 10.1145/3757548. URL https://doi.org/10.1145/3757548.

National Institute of Population Studies (NIPS) [Pakistan] and ICF. Pakistan demographic and health survey 2017-18. Technical report, NIPS and ICF, Islamabad, Pakistan, and Rockville, Maryland, USA, 2019. URL https://dhsprogram.com/pubs/pdf/FR354/FR354.pdf.

Nowak, K. L. and Biocca, F. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(5): 481–494, October 2003. ISSN 1054-7460. doi: 10.1162/105474603322761289. URL http://dx.doi.org/10.1162/105474603322761289.

Ouellette, J. A. and Wood, W. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, 124(1): 54, 1998.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Panneer, S., Sundarraju, S., and Acharya, S. S. Gender-based violence and humanitarian crisis: Advocacy, social justice and policy perspectives. *Journal of Social Inclusion Studies*, 11(1):7–21, June 2025. ISSN 2516-6123. doi: 10.1177/23944811251345101. URL http://dx.doi.org/10.1177/23944811251345101.

Peng, C., Zhang, S., Wen, F., and Liu, K. How loneliness leads to the conversational ai usage intention: the roles of anthropomorphic interface, parasocial interaction. *Current Psychology*, 44(9):8177–8189, October 2024. ISSN 1936-4733. doi: 10.1007/s12144-024-06809-5. URL http://dx.doi.org/10.1007/s12144-024-06809-5.

Pentina, I., Hancock, T., and Xie, T. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600, 2023.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Porta, C. M., Frerich, E. A., Hoffman, S., Bauer, S., Jain, V. M., and Bradley, C. Sexual violence in virtual reality: A scoping review. *Journal of Forensic Nursing*, 20(1): 66–77, December 2023. ISSN 1556-3693. doi: 10.1097/jfn.0000000000000466. URL http://dx.doi.org/10.1097/JFN.0000000000000466.

Pradhan, A., Lazar, A., and Findlater, L. Use of intelligent voice assistants by older adults with low technology use. *ACM Trans. Comput.-Hum. Interact.*, 27(4), September 2020. ISSN 1073-0516. doi: 10.1145/3373759. URL https://doi.org/10.1145/3373759.

Richet, J.-L. Ai companionship or digital entrapment? investigating the impact of anthropomorphic ai-based chatbots. *Journal of Innovation &; Knowledge*, 10 (6):100835, November 2025. ISSN 2444-569X. doi: 10.1016/j.jik.2025.100835. URL http://dx.doi.org/10.1016/j.jik.2025.100835.

Rossi, A., Carli, R., Botes, M. W., Fernandez, A., Sergeeva, A., and Sánchez Chamorro, L. Who is vulnerable to deceptive design patterns? a transdisciplinary perspective on the multi-dimensional nature of digital vulnerability. *Computer Law &; Security Review*, 55:106031, November 2024. ISSN 2212-473X. doi: 10.1016/j.clsr.2024.106031. URL http://dx.doi.org/10.1016/j.clsr.2024.106031.

Saqib, E., He, S., Choy, J., Abu-Salma, R., Such, J., Bernd, J., and Javed, M. Bystander privacy in smart homes: A systematic review of concerns and solutions. *ACM Trans. Comput.-Hum. Interact.*, 32(5), October 2025. ISSN 1073-0516. doi: 10.1145/3731755. URL https://doi.org/10.1145/3731755.

Sardinha, L., Maheu-Giroux, M., Stöckl, H., Meyer, S. R., and García-Moreno, C. Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *The Lancet*, 399(10327):803–813, February 2022. ISSN 0140-6736. doi: 10.1016/s0140-6736(21)02664-7. URL http://dx.doi.org/10.1016/S0140-6736(21)02664-7.

Sarkar, A. Enough with "human-ai collaboration". In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3582735. URL https://doi.org/10.1145/3544549.3582735.

Sarkar, S. and Sinha-Roy, R. Cybersecurity through an intersectional lens: Survivors' responses to technology-facilitated sexual violence in india. *Feminist Criminology*, November 2025. ISSN 1557-086X. doi: 10.1177/15570851251406560. URL http://dx.doi.org/10.1177/15570851251406560.

Schwitzgebel, E. and Garza, M. *Designing AI with Rights, Consciousness, Self-Respect, and Freedom*, pp. 459–479. Oxford University PressNew York, September 2020. ISBN 9780190905071. doi: 10.1093/oso/9780190905033.003.0017. URL http://dx.doi.org/10.1093/oso/9780190905033.003.0017.

Shahid, F. and Vashistha, A. Decolonizing content moderation: Does uniform global community standard resemble utopian equality or western power hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581538. URL https://doi.org/10.1145/3544548.3581538.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 723–741, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604673. URL https://doi.org/10.1145/3600211.3604673.

Skjuve, M., Følstad, A., Fostervold, K. I., and Brandtzaeg, P. B. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601, 2021.

Slupska, J., Cho, S., Begonia, M., Abu-Salma, R., Prakash, N., and Balakrishnan, M. "they look at vulnerability and use that to abuse you'': Participatory threat modelling with migrant domestic workers. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 323–340, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.

org/conference/usenixsecurity22/presentation/slupska-vulnerability.

Southworth, J. Bridging critical thinking and transformative learning: The role of perspective-taking. *Theory and Research in Education*, 20(1):44–63, 2022.

Stapleton, L., Liu, S., Liu, C., Hong, I., Chancellor, S., Kraut, R. E., and Zhu, H. "if this person is suicidal, what do i do?": Designing computational approaches to help online volunteers respond to suicidality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3641922. URL https://doi.org/10.1145/3613904.3641922.

Stark, E. *Coercive control: How men entrap women in personal life*. Oxford University Press, 2007.

Stark, E. and Hester, M. Coercive control: Update and review. *Violence Against Women*, 25(1):81–104, December 2018. ISSN 1552-8448. doi: 10.1177/1077801218816191. URL http://dx.doi.org/10.1177/1077801218816191.

Sundar, S. S. Rise of machine agency: A framework for studying the psychology of human–ai interaction (haii). *Journal of computer-mediated communication*, 25(1):74–88, 2020.

Tei, S. *Death in the Cybernetic Era: AI, Virtual Agents, and Selfless Selves*, pp. 263–285. Springer Nature Switzerland, 2025. ISBN 9783031988080. doi: 10.1007/978-3-031-98808-0_16. URL http://dx.doi.org/10.1007/978-3-031-98808-0_16.

Törnberg, P. and Törnberg, A. Inside a white power echo chamber: Why fringe digital spaces are polarizing politics. *New Media &; Society*, 26(8):4511–4533, September 2022. ISSN 1461-7315. doi: 10.1177/14614448221122915. URL http://dx.doi.org/10.1177/14614448221122915.

Turkle, S. *Alone together*. Basic Books, London, England, October 2012.

UN Women. AI-powered online abuse: How AI is amplifying violence against women. https://www.unwomen.org/en/articles/faqs/ai-powered-online-abuse, 2024.

Vallor, S. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*, 28(1):107–124, 2015.

Vasan, N. and Djordjevic, D. Why ai companions and young people can make for a dangerous mix. Stanford Medicine News, 2025. Based on Common Sense Media risk assessment study.

Vasquez, M., Daspe, M.-É., Bőthe, B., Brassard, A., Lussier, Y., and Vaillancourt-Morel, M.-P. Associations between pornography use frequency and intimate partner violence perpetration among young adult couples: A 2-year longitudinal study. *Journal of Interpersonal Violence*, 39 (21–22):4260–4284, March 2024. ISSN 1552-6518. doi: 10.1177/08862605241234656. URL http://dx.doi.org/10.1177/08862605241234656.

Von Behr, I., Reding, A., Edwards, C., and Gribbon, L. Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism. Research Report RR-453-RE, RAND Europe, Cambridge, UK, 2013. URL https://www.rand.org/pubs/research_reports/RR453.html.

Walker, L. E. *The battered woman*. HarperPerennial, New York, NY, April 1980.

Walsh, J. P. Social media and moral panics: Assessing the effects of technological change on societal reaction. *International Journal of Cultural Studies*, 23 (6):840–859, March 2020. ISSN 1460-356X. doi: 10.1177/1367877920912257. URL http://dx.doi.org/10.1177/1367877920912257.

Wang, S. and Dehnert, M. On-demand intimacy: The sociotechnical appeal of ai companions. *Social Media + Society*, 12(1), January 2026. ISSN 2056-3051. doi: 10.1177/20563051251410394. URL http://dx.doi.org/10.1177/20563051251410394.

Waytz, A., Cacioppo, J., and Epley, N. Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, May 2010. ISSN 1745-6924. doi: 10.1177/1745691610369336. URL http://dx.doi.org/10.1177/1745691610369336.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: how does llm safety training fail? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023b.

Weitzer, R. *Legalizing prostitution: From illicit vice to lawful business*. NYU Press, 2011.

Wetterneck, C. T., Burgess, A. J., Short, M. B., Smith, A. H., and Cervantes, M. E. The role of sexual compulsivity, impulsivity, and experiential avoidance in internet pornography use. *The Psychological Record*, 62(1):3–18, January

2012. ISSN 2163-3452. doi: 10.1007/bf03395783. URL http://dx.doi.org/10.1007/BF03395783.

Woodlock, D., Salter, M., Dragiewicz, M., and Harris, B. "living in the darkness": Technology-facilitated coercive control, disenfranchised grief, and institutional betrayal. *Violence Against Women*, 29(5):987–1004, August 2022. ISSN 1552-8448. doi: 10.1177/10778012221114920. URL http://dx.doi.org/10.1177/10778012221114920.

Xie, T., Pentina, I., and Hancock, T. Friend, mentor, lover: does chatbot engagement lead to psychological dependence? *Journal of service Management*, 34(4):806–828, 2023.

Yu, Y., Mohi, Debroy, A., Cao, X., Rudolph, K., and Wang, Y. Principles of safe ai companions for youth: Parent and expert perspectives, 2025. URL https://arxiv.org/abs/2510.11185.

Zeng, W., Liu, Y., Mullins, R., Peran, L., Fernandez, J., Harkous, H., Narasimhan, K., Proud, D., Kumar, P., Radharapu, B., Sturman, O., and Wahltinez, O. Shieldgemma: Generative ai content moderation based on gemma, 2024. URL https://arxiv.org/abs/2407.21772.

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., and Lee, Y.-C. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713429. URL https://doi.org/10.1145/3706598.3713429.