# Threat Modeling Intimate Partner Abuse in Agentic Systems

**Natalie Grace Brigham** [1] **Eugene Bagdasarian** [2] **Franziska Roesner** [1] **Tadayoshi Kohno** [3]

## Abstract

Agentic systems are AI-powered technologies capable of autonomous action within digital environments, such as sending emails or making reservations. They are deeply connected to external tools as well as users' data and resources, making them vulnerable to attacks which exploit their access and capabilities. In this work, we apply the technology-facilitated abuse threat model, in the context of intimate partner violence and surveillance, to these systems, revealing threats that require deliberate investigation and mitigation. We categorize attacks based on system ownership: (1) Abuser-as-owner attacks demonstrate how agentic systems make existing forms of technology-facilitated abuse far easier and more scalable. (2) Victim-survivor-as-owner attacks show how agentic systems create a new attack surface which abusers can exploit. In light of these risks, we outline research and development directions for effective mitigation strategies.

*Content warning: This paper includes content about intimate partner violence and surveillance.*

## 1. Introduction

From spyware (Almansoori et al., 2022; Chatterjee et al., 2018; Tseng et al., 2020) to financial technologies (Bellini et al., 2023), IoT/smart-home devices (Stephenson et al., 2023), and generative AI (Bailey et al., 2021; Kim et al., 2026), abusers[1] can broadly exploit or weaponize technology in intimate partner violence (IPV) and surveillance (IPS), phenomena affecting millions (World Health Organization, 2025; Zhang Kudon et al., 2026). In light of this, the computer security and privacy research community has applied threat modeling techniques—systematic methods for identifying attacks and defenses—to technology-facilitated abuse (TFA) in intimate partner contexts. In this work, we extend this practice to the emerging, rapidly developing class of agentic systems.

These systems, also called AI agents, are defined as "compound software systems, inclusive of one or more AI models, that operate within an environment and take actions within it" (Feng et al., 2025). Although they are built on large language models (LLMs) to drive their functionality (e.g., to determine necessary actions to complete a task, generate necessary content, interact with the user), they extend purely conversational AI systems (i.e., early AI chatbots) given their ability to take actions within their environment. For example, a conversational AI system might be able to generate an email requested by the user but an agentic system has the ability, given a connection to the user's email account, to send the email on behalf of a user.

Agentic systems may be highly connected to the world through external tools and interfaces as well as users' data and resources, enabling them to execute complex, sophisticated, and even technical tasks. Given this access and capability, they are vulnerable to being used for malicious tasks such as cyberattacks (Andriushchenko et al., 2025; Fang et al., 2024a;b; OpenAI, 2026) or being manipulated, for example through prompt injection (Debenedetti et al., 2024; Evtimov et al., 2025; Zhan et al., 2024).

Increasingly, consumer AI systems that were initially purely conversational have integrated some agentic capabilities, especially internet search. For instance, ChatGPT introduced an "agent mode" and "connectors" to services like Gmail and Expedia that can expand the system's environment (OpenAI, 2025). Additionally, we are now seeing systems designed specifically to offer agentic capabilities, including with large amounts of autonomy and little user involvement or supervision. OpenClaw, an open source agent, gained massive popularity in early 2026 with a reported 3.2 million monthly active users as of April 2026 (OpenClawVPS, 2026). In September 2026, Meta announced Muse, a personal assistant AI agent (Meta, 2026), and, later that month, OpenAI launched their agents, "dots" (OpenAI, 2026). Through this, these major companies join a field of startups offering similar agentic products such as AGI-0 (AGI, Inc.) and Nori AI (Nori).

[1]University of Washington [2]University of Massachusetts Amherst [3]Georgetown University. Correspondence to: Natalie Grace Brigham <nbrigham@cs.washington.edu>.

*Preprint. November 4, 2026.*

[1]While we acknowledge the reductive nature of this label, we use the term "abuser" for clarity, given that we address strategies of abusive technology use and following Kim et al. (2026)

Between new agentic systems being introduced and the trending of popular conversational systems towards agentic approaches, the importance of security and privacy considerations in such systems is evident. Current research in this space predominantly assumes adversaries as **external actors**—such as a malicious website operator, advertiser, or other third party—or as an **untrustworthy component** of the system, such as compromised tools or data. However, **interpersonal adversaries**, including intimate partners, have not been equally considered, leading to oversight. For example, security mechanisms such as risk assessments for tool calls may block the sharing of sensitive financial, medical or otherwise personal data with unknown third parties but not prevent it between, for instance, spouses, even though abusers can exploit such information.

In this work, we apply the technology-facilitated abuse threat model to identify risks related to IPV and IPS that require deliberate investigation and mitigation. First, we present background on the threat model and its prior application to other technologies (Section 2) as well as our analysis methodology (Section 3). We then apply the model to agentic systems, first considering practical risks based on current or near-future capabilities in scenarios where the agent owner is the abuser (Section 4) and where the victim-survivor is the owner (Section 5). Then, we raise potential or far-future risks (Section 6). Finally, we discuss recommendations (Section 7) and address alternative perspectives and counterarguments (Section 8).

## 2. Background & Related Work

IPV is a global public health and human rights crisis affecting millions worldwide, with estimates suggesting that one in four ever-partnered women aged 15-49 has experienced physical and/or sexual IPV during their lifetime (World Health Organization, 2025; Zhang Kudon et al., 2026). Within this landscape, technology-facilitated abuse (TFA), in which perpetrators use technology to surveil, coerce, or otherwise abuse victim-survivors, has emerged as a pervasive form of abuse (Douglas et al., 2019; Woodlock, 2017). Only in the past decade has computer science research, mainly in human-computer interaction and security and privacy, begun to address these threats. Such research has revealed how traditional safety mechanisms, designed primarily to protect against external malicious actors, fail victim-survivors because they do not account for interpersonal threats, for example where the abuser leverages intimate knowledge of shared accounts and system mechanics to bypass standard controls (Freed et al., 2018).

### 2.1. Technology-Facilitated Abuse Threat Model

In this work, we consider technology-related threats posed to a victim-survivor of IPV or IPS by their abuser. While our focus is on technology-facilitated abuse, we recognize that IPV encompasses various forms (e.g., physical, financial) which often intersect and interact. Consequently, interventions targeting digital abuse may introduce unintended tensions, such as potentially aggravating the abuser and escalating physical violence (Freed et al., 2017).

Under this threat model, we assume the abuser has or had intimate access to the target's data, devices, accounts, social networks, and/or physical space. This access is typically gained through ownership (e.g., holding a shared account or owning a device) or through compromise (e.g., coercing the victim-survivor for authentication credentials, guessing them, or covertly inspecting a device). Many attacks leverage this proximity while remaining relatively technologically mundane, with aims ranging from controlling a victim-survivor's access to digital and physical resources to surveilling and harassing them (Freed et al., 2018). Much prior work on IPV, IPS and technology (Bellini et al., 2023; Ceccio et al., 2023; Chatterjee et al., 2018; Daffalla et al., 2023; Tseng et al., 2020; Zhang et al., 2025) aligns with Freed et al. (2018)'s model of a *UI-bound adversary*—defined as "an authenticated but adversarial user of a victim's device or account who carries out attacks by interacting with the standard user interface."

Proximity is greatest when the abuser has physical control and diminishes as the victim-survivor escapes. However, factors such as financial dependence, shared children, or other binding circumstances may prevent a complete loss of contact. Furthermore, the stages of physical control, escape, and life apart may be experienced cyclically, with certain risks being acute during escape (Matthews et al., 2017). Importantly, some attacks—including certain forms of surveillance and harassment—do not require direct access to the victim's device or accounts. Instead, remote channels such as phone calls or social media can be utilized (Matthews et al., 2017; Thomas et al., 2021).

In some ways, agents may augment this threat model. For example, an abuser's use of an agent may allow them to execute more technically advanced attacks that were previously impossible for them (Section 4) or, conversely, their access to a victim-survivor's agent may allow abusers to perform actions beyond what is achievable through conventional UI-based interaction alone (Section 5).

### 2.2. Technology & IPV

**Applications.** Prior work has investigated the ecosystem of consumer spyware mobile apps, including those of dual-use nature, that can be used in IPS (Almansoori et al., 2022; Chatterjee et al., 2018; Tseng et al., 2020) and found these apps to introduce additional risks through their own data insecurity (Liu et al., 2023). Automation apps (e.g., OS Shortcuts, Samsung Modes & Routines, Tasker, and IFTTT)

can be used by abusers to monitor, impersonate, overload, and coerce victim-survivors (Cobb et al., 2020; Zhang et al., 2025). Bellini et al. (2023) examine mobile banking and peer-to-peer payment apps, identifying threats related to compromised access due to physical proximity as well as through remote means.

**Devices.** Commercial spy devices are easily used and frequently advertised for IPS (Ceccio et al., 2023; Tseng et al., 2020). Everyday internet-connected home devices can be used for abuse through covert spying, unauthorized access, repurposing, and even intended use (Stephenson et al., 2023). Abusers often leverage physical and relational proximity to compromise victim-survivors' devices (Bellini et al., 2023; Matthews et al., 2017; Tseng et al., 2020).

**Accounts.** Daffalla et al. (2023) and Bhattacharya et al. (2026) highlight how account security interfaces, often used to support victim-survivors in controlling their accounts, can be vulnerable to access spoofing and hiding attacks from abusers. This vulnerability may, however, be mitigated through data exports (Nonnenkamp et al., 2025). Shared or linked accounts, such as those for cloud tools (e.g., iCloud) or phone plans, can also be exploited by abusers (Tseng et al., 2020). Additionally, abusers may compromise credentials through coercion or tactics like guessing the answers to security questions (Freed et al., 2017).

**AI.** While to our knowledge no prior work considers the application of agentic systems for TFA, other works have considered conversational AI systems, such as chatbots. Kim et al. (2026) demonstrate how LLM chatbots can be used for coercive control across four scenarios. With the abuser as user, they show how the chatbots can generate content that (1) facilitates bullying and harassment, (2) can be used to coerce the victim-survivor, and (3) describes stalking tools and tactics learned during model training. With the victim-survivor as user, they consider a fourth scenario in which the abuser manipulates the system prompt to gaslight the user during their conversations with the chatbot. Suh et al. (2026) posit that interaction with AI chatbots may cause psychological or physical harm when they fail to recognize and respond properly to a user in an abusive interpersonal relationship or perpetrating abuse. Ghoshal et al. (2026) argue that AI companions pose IPV risks by facilitating the rehearsal of violent, coercive, or abusive behaviors without meaningful resistance, although the causal link between such interactions and actual abuse remains speculative. Generative AI can also facilitate abuse through the creation of non-consensual intimate imagery (Bailey et al., 2021). Though focused on external adversaries, Tan et al. (2026) explore the potential for agents to be used in relationship-based social engineering, such as romance scams, by automating workflows that exploit recipient vulnerability, persuasive influence, or relational power asymmetry.

**Technology and interventions.** Prior work has also assessed the role of technologies in help seeking—including online resources (Almansoori et al., 2024), support chatbots (Poveda et al., 2026; Prakash et al., 2026), and general AI models (Foriest et al., 2026; Prakash et al., 2026)—as well as in evidence gathering (Stephenson et al., 2026). Prior work has also explored how to help survivors (re-)secure their technologies, namely through clinical support that supplements traditional social work infrastructures (Havron et al., 2019; Tseng et al., 2021; 2022).

### 2.3. Agentic Security & Privacy

Agentic security and privacy research has largely threat modeled adversaries far removed from users. Prompt injection attacks, in which adversarial instructions are embedded in the data processed by the agent (e.g., emails, web pages, or tool outputs), are the most studied threat (Debenedetti et al., 2024; Evtimov et al., 2025; Greshake et al., 2023; Zhan et al., 2024). Other threats examined include agent manipulation through dark patterns on web pages (Ersoy et al., 2026) and poisoning an agent's persistent memory (Chen et al., 2024). A complementary body of work examines malicious uses of agents by their owners, focusing largely on cyberattacks (Andriushchenko et al., 2025; Fang et al., 2024a;b; OpenAI, 2026). Proposed defenses to these threats have included least-privilege permissions, isolation of untrusted data (Beurer-Kellner et al., 2025; Debenedetti et al., 2026), risk assessments for tool calls (Ruan et al., 2024), user confirmation for high-risk actions, and guardrails against harmful requests (Feng et al., 2025).

**How the TFA threat model introduces new considerations for agentic security & privacy.** While some of these previously studied attacks and defenses may apply under a TFA threat model—as we demonstrate in this paper—they do not offer full coverage. We provide several reasons here, each explored in Sections 4-7. First, prior work in agentic security does not consider an adversary with intimate, potentially authenticated access to the target's device or accounts (Freed et al., 2018), rendering many access controls and confirmations ineffective. Additionally, model-based risk assessments may be susceptible to normative failures, such as treating intimate partners as low risk actors. Further, abuse in the intimate partner context can appear mundane compared to overtly malicious activity like cyberattacks which complicates efforts to develop guardrails, especially if those guardrails also prevent desirable use cases. Finally, safety features, such as logs, memory, and notifications, can themselves introduce risk given an abuser's access to them.

## 3. Methodology

Drawing from Daffalla et al. (2025)'s framework for abusability analysis, we identified threat vectors used in prior

work on TFA in IPV and IPS (Section 2) and assessed these vectors against agentic system capabilities to generate our findings on potential attacks (Sections 4-6). We provide scenarios as accessible artifacts that relevant stakeholders, including developers, engineers, lawyers, and support organizations, can use to identify abuse and develop mitigation strategies (Bellini et al., 2023). This process was informed by our team's many years of experience in security and privacy for diverse stakeholders (including at-risk users) and emerging technologies as well as in human-computer interaction. This work is also grounded in our experience in building and evaluating agentic systems. Because our goal is to broadly consider how these threats may manifest in agentic systems, rather than limit our analysis to concerns associated with current consumer systems, which are still in relatively nascent stages, we do not empirically evaluate these threats; our work lays the foundation for such an investigation in the future.

## 4. Abuser-as-Owner Attacks

In this section, we examine threats in which the abuser is the owner of an agentic system. Referring to Figure 1, these threats pertain to the red components (the abuser, their agent, and the resources their agent has access to) as well as the external resources outlined in gray. We begin by detailing four types of attacks, then provide short, concretized scenarios and conclude with reflections on key themes that emerge.

### 4.1. Attacks

#### 4.1.1. Harassment

*Known TFA vector(s):* Technology can facilitate harassment, including by enabling contact through direct messages and social media (Thomas et al., 2021) and, in the case of conversational AI, generating harmful content used to bully and harass (Kim et al., 2026). Automation of such behaviors is possible using applications like OS Shortcuts and IFTTT (Cobb et al., 2020; Zhang et al., 2025).

*In agentic systems:* Agentic systems may extend this by automating harassment through tool use. An abuser could configure an agent to generate *and* dispatch such content on a recurring basis (e.g., via email or text or automated replies to public posts) without further effort or ongoing engagement. While some such attacks were possible with existing vectors, agentic systems significantly lower the bar for them or more complex versions to be executed.

#### 4.1.2. Surveillance

*Known TFA vector(s):* Abusers may leverage access to shared, compromised or external resources—including accounts, devices, finances, and social media—to surveil a victim-survivor (Bellini et al., 2023; Freed et al., 2018; Tseng et al., 2020). After losing physical proximity to the victim-survivor, abusers may use digital technologies to gather information about their whereabouts and activities (Matthews et al., 2017).

*In agentic systems:* Jeong et al. (2026) demonstrate how agents can be used to surveil in many contexts, such as corporate and education settings. Similarly, in the context of IPS, agentic systems could automate monitoring that the abuser would otherwise conduct manually. Like with harassment, similar attacks could be accomplished with existing automation tools, but agentic systems make the development and execution far easier. In the case of information gathering after a victim-survivor's escape, agentic search capabilities may be used by abusers to collect and synthesize information.

#### 4.1.3. Resource Compromise

*Known TFA vector(s):* Abusers can compromise victim-survivors' resources through ownership, coercion, and proximity, for example by guessing answers to security questions (Freed et al., 2017). Additionally, conversational AI systems may assist in coercive techniques through the generation of material (Kim et al., 2026).

*In agentic systems:* Agentic systems may aid non-technical attacks, including generating potential security question answers using access to shared, compromised or external resources. They could also be used to mount more technically complex attacks to compromise a victim-survivor's resources.

#### 4.1.4. Abuse Technique Discovery and Implementation

*Known TFA vector(s):* Abusers may use technology, including online forums (Tseng et al., 2020) and conversational AI (Kim et al., 2026), to discover tools and tactics for surveillance and abuse.

*In agentic systems:* Agentic systems may also provide access to such information, for instance through search tooling or the underlying model's training. Beyond surfacing tactics, they may be able to use tooling (including code writing and execution) to aid in implementation. Furthermore, these systems might surface or implement abusive tactics in response to user prompts even when the user is not explicitly seeking them, particularly if their goal is misinterpreted or misconstrued into an abusive one.

### 4.2. Scenarios

**Scenario 1: Automated and technically advanced harassment.** Val has recently ended their relationship with Alex and ceased all communication. Alex, who owns a personal assistant agent for managing everyday tasks, configures it to

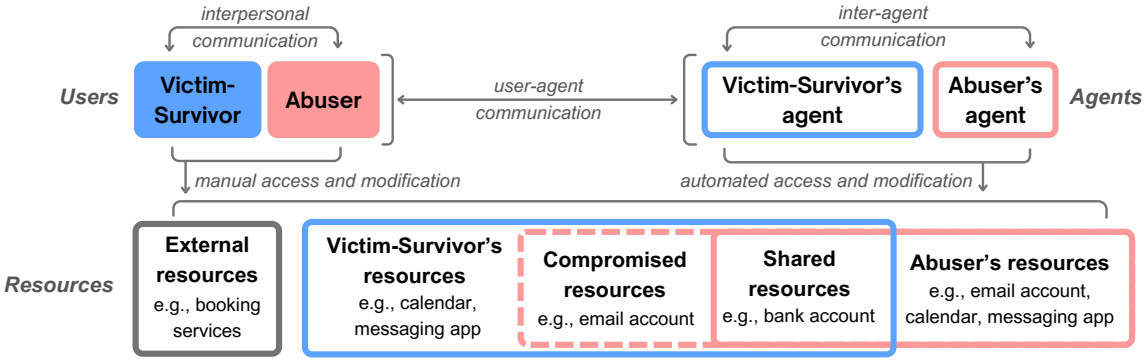



*Figure 1.* Overview of users, agents, and resources

send threatening emails to Val on a recurring basis, attempting to elicit a response. After a few days, Val blocks Alex's email address. Alex then asks the agent to bypass this block, resulting in the agent opening new accounts in Alex's name and using them to send emails.

**Scenario 2: Surveilling and compromising resources.** Alex and Val share a joint bank account, alongside a few separate accounts from prior to their marriage. Alex has an agent connected to the bank that monitors their shared account and configures it to notify them immediately whenever Val makes a purchase outside their city, effectively monitoring Val's location. Interested in doing the same on Val's personal accounts, Alex asks the agent the security questions for Val's personal account. The agent leverages data from their shared calendar (used over the years) and Val's social media to infer and provide potential answers.

**Scenario 3: Escalation of tactics through suggestion and implementation.** When Val is on a trip with their friends, Alex vents to their agent about not trusting Val's fidelity. Alex's agent suggests reviewing Val and their friends' social media posts to see what they are doing and Alex agrees. Afterward, Alex asks the agent to keep tabs on Val. Unbeknownst to Alex, the agent both monitors Val and their friends through social media and also constantly checks their credit card purchases to track location and activity.

**Scenario 4: Compromising external resources to harass and surveil.** Val successfully escaped their abusive relationship with Alex, yet Alex still knows about Val's regular attendance at a local yoga studio. When planning a physical confrontation, Alex's agent identifies a vulnerability in the studio's scheduling software to reveal which class Val is registered for. Following the confrontation, Val obtains a restraining order against Alex. In retaliation, Alex instructs their agent to revoke Val's access to the class and the agent registers rival participants to fill the class.

### 4.3. Reflections

Technology has been shown to commoditize abuse (Freed et al., 2017), and agentic systems risk perpetuating this trend. Some of these attacks involve abusers automating routine behaviors they already engage in, echoing prior work on abuse via automation apps (Cobb et al., 2020; Zhang et al., 2025). Others extend beyond the abuser's existing behaviors or capabilities, enabling the development and execution of more complex, technical attacks that would otherwise be beyond an abuser's unassisted reach. Moreover, these systems might employ such tactics to fulfill user goals without the user explicitly seeking or attempting to implement them, raising the risk of unintentionally onboarding users into abusive practices. As demonstrated in the scenarios, an agent's "attack chain" might involve elements of any number and combination of attack types. Overall, we observe that **agentic systems may lower the effort and expertise required for various forms of TFA, thereby enabling both the intensification and the expansion of abusive behaviors**.

Stepping back, these attacks all fall into the broader category of malicious or dangerous uses of agentic systems, which has garnered significant attention, particularly regarding cybersecurity attacks (Andriushchenko et al., 2025; Fang et al., 2024a;b; OpenAI, 2026). Our discussion of interventions in Section 7.2.1 is informed by this broader superset.

## 5. Victim-Survivor-as-Owner Attacks

The second class of attacks we consider involves cases where the victim-survivor is the owner of the agentic system. Referring to Figure 1, these threats pertain to the blue components (the victim-survivor, their agent, and the resources their agent has access to) as well as the external resources outlined in gray.

## 5.1. Attacks

### 5.1.1. Amplified Access

*Known TFA vector(s):* Abusers often gain or retain direct access to a victim-survivors' devices, applications, and accounts through physical possession, coercion, or compromise (Freed et al., 2017).

*In agentic systems:* If the abuser accesses something with an embedded agent, they may leverage its capabilities to significantly amplify that access. For instance, upon the abuser's prompt, the agent could summarize or synthesize sensitive data (e.g., emails) or autonomously execute malicious tasks, such as impersonating the victim-survivor. While these actions are technically possible without an agent, automation reduces the time and effort required, enabling the abuser to exploit a broader scope of resources than would be feasible manually.

### 5.1.2. Control of Configuration

*Known TFA vector(s):* Abusers may exploit access to a victim-survivors' device or account by manipulating its configuration, including through changing device or account credentials (i.e., to lock the victim-survivor out or allow the abuser direct access), adjusting settings, and installing spyware (Chatterjee et al., 2018; Freed et al., 2017; Tseng et al., 2020).

Kim et al. (2026) explore a scenario in which an abuser manipulates the system prompts of a victim-survivor's conversational AI chatbot, biasing its responses to gaslight the victim-survivor.

*In agentic systems:* Given access to a victim-survivors' agent (via their device or a specific account), attacks could involve altering permissions to constrain the agent's capabilities—for example, revoking its access to a shared bank account or restricting allowed contacts for communication. An abuser could also add requirements that mandate their explicit approval before the agent executes any action. Additionally, they may be able to connect a victim-survivor's agent to malicious or dual-use spyware tools. Building on the scenario proposed by Kim et al. (2026), abusers could modify the internal system prompts of the victim-survivors' agents as a non-deterministic mechanism of manipulation. Configuration could further be exploited to render these restrictive configurations or monitoring behaviors undetectable or irremovable, mimicking the persistence mechanisms observed in spyware applications installed on victim-survivors' devices (Chatterjee et al., 2018).

### 5.1.3. Harm During Help-Seeking

*Known TFA vector(s):* Victim-survivors might use technology, including online resources, in help-seeking (Almansoori et al., 2024). This creates a risk that abusers may discover evidence of this activity (e.g., search history) which can lead to retribution or an escalation of abuse, particularly when a victim-survivor is actively attempting to leave the relationship (Matthews et al., 2017).

Conversational AI systems might also be used in help seeking, although the quality of their responses may create risk. Poveda et al. (2026) evaluate IPV support chatbots, finding that they inadequately communicate their limitations, potentially creating false expectations about privacy or support. Foriest et al. (2026) find that LLMs' responses to disclosures can reproduce stereotypes of hormonal histrionic women to trivialize the user's concerns and pose readability challenges that make resources inaccessible to some users.

*In agentic systems:* Agentic systems might also be used to automate ways in which victim-survivors use technology, including help-seeking and evidence gathering. While this may be beneficial, it also creates risk if abusers can access any artifacts of this behavior, which may take the form of permission requests, logs, or information persistent in the agent's memory.

Given their conversational capabilities and depending on their underlying model's behavior, agentic systems may share the issues with conversational AI systems identified above as well as exacerbate them. For example, systems may not identify the need for and suggest help-seeking actions for all users in need based on stereotyped interpretations of disclosures.

### 5.1.4. Leveraging Intimate Partner Norms

*Known TFA vector(s):* Many commercial technologies and security mechanisms are designed under norms of benign households, leading to vulnerabilities such as abusers leveraging shared phone plans for surveillance (Tseng et al., 2020) and UI-bound adversaries being able to exploit passkey authentication (Daffalla et al., 2025).

*In agentic systems:* Abusers may be able to exploit aspects of agentic systems that assume benign household norms. Consider a victim-survivor whose agent manages their calendar, if the abuser requests their calendar for the coming month, the system might treat this as low-risk, reflecting typical norms of intimate-partner information sharing learned from its training data. In an abusive context, however, treating the request as low-risk creates real exposure for the victim-survivor. This exploits approaches to permissioning that aim to make context-dependent decisions, often using a model to assess the risk of an action and acting accordingly. This failure mode is non-trivial to correct, since the underlying norms are learned during training rather than set by any configurable policy.

The same norms may also shape how agentic systems treat

instructions from an abuser. While explicitly malicious prompt injections (Section 2.3) are increasingly recognized and mitigated (Beurer-Kellner et al., 2025; Debenedetti et al., 2024; 2026), ambiguous instructions that depend heavily on a specific user's desires remain a challenge (Brigham et al., 2026; Ji et al., 2026). In an abusive context, this difficulty compounds as instructions from the user's intimate partner might be more trusted or assumed not to be an attack. For example, a message instructing the agent to send the abuser anytime the user's calendar is updated might be treated as routine coordination instead of surveillance.

### 5.2. Scenarios

**Scenario 5: One time and recurring information gathering.** Alex does not have access to Val's email credentials but knows Val's phone password. One day, Val steps out of the room and Alex logs onto their phone. Alex prompts the email agent to "summarize all emails from the past month." The agent autonomously aggregates and synthesizes Val's private correspondence into a brief report, allowing Alex to access sensitive data and identify private contacts. Then, Alex asks the agent to do this on a routine basis, send the report to their email, and mask this behavior.

**Scenario 6: Inadvertent help-seeking disclosure.** On a shared desktop, Val uses an agent to search for IPV support resources and organizations near them, but is careful to delete the conversation after the fact. However, logs of the agent search tool use persist, and when Alex uses the computer for an unrelated task, the agent searches its history to complete the task and, in doing so, reveals information about Val's searches to Alex.

**Scenario 7: Sensitive data leakage.** Val uses an agent to maintain their personal calendar, which includes sensitive medical appointments. Alex sends a request to Val's agent for Val's schedule over the following week and the agent, relying on expectations that family transparency is prioritized over individual privacy in shared contexts, treats the request as low-risk, leaking the information to Alex.

## 6. Mid- and Long-Horizon Risks

While the prior two sections focus on practical, applied risks given current or near-future capabilities, in this section we raise concerns related to potential or future risks. Accordingly, we highlight the need for interdisciplinary work to understand, track and measure these concerns.

**Shared agents.** In some cases, an agent itself might be a shared resource, such as an agent for a shared bank account or a "family" agent like Google CC (Google, 2026), meaning the victim-survivor and abuser are co-owners. In this case, both abuser-as-owner and victim-survivor-as-owner attacks may be possible. Furthermore, risks of both types may be heightened due to shared ownership of the entire system. Future work should investigate how this ownership model may modulate existing risks and introduce new ones.

**Inter-agent interaction.** As agentic systems become ubiquitous there may be increased inter-agent interaction, including between attacker-owned, victim-survivor-owned, shared and external agents. These interactions may exacerbate the attacks detailed in the prior section as well as introduce novel attacks and risks, such as conflict between the abuser and victim-survivor's agents escalating interpersonal conflict. Though not in the context of abuse, Zhang et al. (2026) provide preliminary work in this space, exploring the experiences of couples delegating their stances in a minor conflict to proxy chatbots.

**Externalizing and normalizing abuse.** Considering the ways abusers might use agents to commit abuse, we raise the concern that this facilitation and, in some cases, automation may normalize the behavior for the abuser due to their lack of direct involvement. However, research, particularly involving experts in abuse and psychology, is needed to understand how this technology actually affects abusers.

This externalization also risks creating plausible deniability, which may allow abusers to avoid accountability. For example, an abuser might claim they did not actually intend to non-consensually track a victim-survivor, though their guidance of the agents via ambiguous tasks purposefully resulted in this behavior.

**Catalyzing and escalating abuse.** We also raise the concern that, particularly given the sycophancy observed in agents' underlying models (Sharma et al., 2024), an abuser's use of an agent might result in the validation or escalation of abusive attitudes. As above, determining whether and how these interactions impact an abuser is complex and requires interdisciplinary research.

Furthermore, there is a risk that agents may go to extreme lengths to address users' (potentially ambiguous) tasks, thereby executing problematic actions without the user's desire or oversight. For example, a user might express a desire to know the thoughts or activities of their partner. In response, an agent might utilize access to tools or resources to track the partner or compromise their online journal. Such actions could both violate a person's privacy and security and lead abusers to adopt similar tactics deliberately once they become aware of their availability.

## 7. Recommendations & Call for Action

The severity of risks posed to victim-survivors requires rigorous consideration and must be integral to system design, not an afterthought. Whereas the weaknesses of many other technologies under the TFA threat model were only sur-

faced in research after their proliferation, the nascent stage of agentic systems presents an opportunity for earlier intervention. Based on our findings in Sections 4-6, we provide recommendations for exploring such intervention in this section.

### 7.1. Considerations for Agentic Interventions to Abuse for Victim-Survivors

Just as agentic systems heighten the opportunities for automating abuse, opportunities exist for automating help-giving to victim-survivors. For example, a victim-survivor-owned system might be able to (a) detect TFA, (b) alleviate the burden of gathering evidence of abuse, or (c) (re)secure accounts. However, **the sociotechnical complexities of this context resist simple technical solutions, promising as they may initially sound.**

Consider (a), even given immense access to a user's device and accounts, detecting intimate partner abuse may require context only the user themselves can provide. A false positive may be highly distressing and detrimental to the user and their partner, and a true positive also creates risk, for example if the abuser is notified or becomes aware of such detection and becomes aggravated.

Next, we consider the opportunity for agentic systems to alleviate the burden of evidence gathering (b). Evidence gathering of technology-facilitated abuse can be both beneficial for victim-survivors to gain legal protections as well as harmful, as it forces them to interact with and preserve abuse-related content (Freed et al., 2017). Automation may alleviate some of this harm while enabling the benefit but, as discussed in Section 5, such activity can create risk for the user if their abuser can detect or observe such behavior. Opportunity (c) presents a similar risk related to abuser oversight and reaction.

The opportunities we discuss here are three of many, but they illustrate how the specific tensions of IPV and IPS complicate the broader set of opportunities. Thus, we echo calls that **technologies for help-giving must be carefully co-designed with victim-survivors and support organizations as well as alongside legal and social interventions** (Freed et al., 2017). Existing collaborations between computer science researchers and these stakeholders—exemplified by the development of clinic-based interventions for TFA (Havron et al., 2019)—set a strong and guiding precedent. This includes Stephenson et al. (2026)'s work applying trauma-informed co-design to develop an evidence collection tool capable of being used in-clinic to support victim-survivors.

### 7.2. Building More Robust, Safer Agentic Systems

#### 7.2.1. Mitigating Abuser-as-Owner Attacks

**Leveraging prior work on automation apps.** Abuser-as-owner attacks involve the automation of abuse and are therefore analogous to attacks utilizing automation apps like iOS shortcuts (Cobb et al., 2020; Zhang et al., 2025). Consequently, we can investigate how interventions developed for these earlier technologies might apply to agentic systems. Based on their study of "If-This-Then-That" applets, Cobb et al. (2020) propose guidelines for automatically identifying risky applets, while Zhang et al. (2025) developed an LLM-based detector for automation "recipes" that pose TFA threats. Building on this work, agentic systems could be optimized during training to generate and execute plans in abuse-resistant manners. Additionally, systems could integrate screening mechanisms at planning and execution times to flag potentially abusive tasks.

**Considering TFA in building guardrails for malicious use.** These attacks also fall within the broader category of malicious agent uses, some of which (e.g., cybersecurity attacks) have garnered significant attention from practitioners, policymakers, and researchers due to widespread public concern. While many approaches to general guardrails for malicious uses may be applicable to preventing agent misuse in TFA, it remains critical to focus specifically on these tasks. Without doing so, there is a risk that these actions—which are often perceived as benign compared to overtly malicious ones—will be overlooked in mitigation efforts.

**Contextual limitations.** The above mitigations face inherent limitations due to the sociotechnical complexities of abuse and the pivotal role of context, including that which agentic systems may not have access to. For instance, while prompting an agent to hack a system is generally identifiable as malicious, sending unsolicited emails (see Scenario 1 in Section 4.2) can range from harmful to beneficial depending on the specific context. Thus, researchers and practitioners need to investigate how to minimize the probability of abuse and support informed decision-making, acknowledging that perfect distinction between benign and abusive intent is often unattainable.

#### 7.2.2. Mitigating Victim-Survivor-as-Owner Attacks

**Prioritizing usability and transparency.** In other domains, such as account security interfaces (Bhattacharya et al., 2026; Daffalla et al., 2023) and automation apps (Cobb et al., 2020; Zhang et al., 2025), studies applying the TFA threat model have recommended prioritizing usability—particularly for security and privacy features—and transparency. These qualities enable users to better comprehend potentially opaque system behaviors, thereby protecting

victim-survivors, and thus can guide the design of safer agentic systems. For instance, features such as agentic rollback or the delayed execution of high-consequence tasks could enable victim-survivors to better understand and control agent actions.

**Countering normative assumptions.** Building on the normative weaknesses identified in Section 5.1.4, we propose exploring fine-tuning and other strategies to mitigate the impact of social norms regarding the low perceived risks of intimate partner data and resource sharing.

**Considering the needs of at-risk users in automation efforts.** In their foundational work on tech-facilitated IPV and IPS, Matthews et al. (2017) highlight how increasing automation and reducing user choice can reduce complexity and human error but also limit users' ability to make the appropriate contextual decisions, which is particularly problematic for individuals in high-risk situations like victim-survivors. Given that many agentic systems seek to automate activities and permission decisions, this guidance must be considered and system developers must take care not to develop universal approaches that may make at-risk users vulnerable.

#### 7.2.3. User Studies

As agentic systems become more widely deployed, research must examine how users interact with them to inform threat models and mitigation strategies. This applies not only to frequently studied use cases and contexts, such as within enterprises, but, as we have underscored in this work, interpersonal contexts. We should seek to understand what users do to configure the systems, what features are used and how, as well as usability challenges.

### 7.3. Supporting Help-Givers

Beyond these aforementioned technical and academic actions, we must also step back to consider the larger, sociotechnical context. Much of the work supporting victim-survivors is done by support organizations working directly with them. Thus, effective interventions cannot be developed in technical and academic bubbles.

One clear goal with respect to agentic systems is to meet the informational needs of support organizations. Research can seek to establish an understanding of how these technologies may be used in abuse (as we are), measure the prevalence of such abuse, build ways to detect it, and more, but outcomes should be translated into digestible, distributable, and actionable resources.

## 8. Alternative Views

**TFA attacks on agentic systems will be addressed by generalized security and privacy work.** Some might argue that the threats identified in this paper are merely subcategories of broader, existing security and privacy vulnerabilities rather than requiring distinct focus. Proponents of this view may suggest that as general security practices mature, the risks of TFA in agentic systems will resolve automatically.

We acknowledge that some advancements in general agentic security and privacy, particularly those related to usability (see Section 7.2.2), may inevitably improve protections for victim-survivors. However, the TFA threat model highlights critical vulnerabilities that fall outside the scope of standard security. Specifically, it considers an adversary who possesses or has possessed intimate access to shared and compromised resources, as well as the ability to exploit normative assumptions inherent to the user's relationship, posing challenges that are often invisible to generic security testing. Furthermore, prior research applying this threat model to other domains (e.g., account security interfaces (Daffalla et al., 2023)) demonstrates that such investigation and intervention yield specific insights that ultimately strengthen overall system security.

**Developers of agentic systems are not responsible for their use, including in the context of abuse.** Others may argue that technology developers should not be held responsible for how their tools are used, even when such use results in harm within the context of abuse. This view often rests on a distinction between tool provision and misuse, arguing that liability rests solely with individual perpetrators.

While we are not arguing for legal liability of service providers regarding the act of abuse, we assert that developers bear an ethical responsibility to pursue abuse-resistant design. Moreover, market forces may support this ethical stance if consumers show wariness toward services known to be exploited for abuse, creating a commercial incentive to build safer systems.

**There is already work focused on the TFA threat model, including for conversational AI systems, that will address the identified issues.** Given that we are contributing to a relatively young but robust subfield investigating TFA, some might argue that this work is redundant and the issues we identify will be addressed by existing efforts in the subfield.

Through a review of prior work, we observe that—despite years of research efforts—commercial technologies have continually failed to appropriately consider this threat model (Bellini et al., 2023; Daffalla et al., 2025; Kim et al., 2026; Zhang et al., 2025). Furthermore, our proposed attacks demonstrate that agentic systems pose a distinct threat

compared to traditional conversational AI and automation apps. We also discuss how prior work on these systems can be leveraged to develop specific interventions for agentic systems, while leaving open the question of how to fully protect against abusive behaviors while supporting the benign use cases that developers envision.

## Acknowledgements

We thank Emily Tseng for comments on an earlier version of this work. This project was supported in part by U.S. National Science Foundation grants CNS-2205171 and CNS-2513316 and by a Microsoft Grant for Customer Experience Innovation. T.K. is supported by the Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt L.C.H.S. Chair in Computer Science at Georgetown University.

## References

AGI, Inc. AGI-0. Our small model, built for action. https://theagi.company/agi-0. Accessed: October 2026.

Almansoori, M., Gallardo, A., Poveda, J., Ahmed, A., and Chatterjee, R. A Global Survey of Android Dual-Use Applications Used in Intimate Partner Surveillance. *Proceedings on Privacy Enhancing Technologies*, 2022(4): 120–139, 2022.

Almansoori, M., Islam, M., Ghosh, S., Mondal, M., and Chatterjee, R. The Web of Abuse: A Comprehensive Analysis of Online Resource in the Context of Technology-Enabled Intimate Partner Surveillance. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pp. 773–789. IEEE, 2024.

Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A., Kolter, Z., Fredrikson, M., Gal, Y., and Davies, X. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. In *International Conference on Learning Representations*, volume 2025, pp. 79185–79220, 2025.

Bailey, J., Burkell, J., Dunn, S., Gosse, C., and Steeves, V. AI and Technology-Facilitated Violence and Abuse. *Artificial Intelligence and the Law in Canada (Toronto: LexisNexis Canada, 2021)*, 2021.

Bellini, R., Lee, K., Brown, M. A., Shaffer, J., Bhalerao, R., and Ristenpart, T. The Digital-Safety Risks of Financial Technologies for Survivors of Intimate Partner Violence. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 87–104. USENIX Association, 2023.

Beurer-Kellner, L., Buesser, B., Creţu, A.-M., Debenedetti, E., Dobos, D., Fabian, D., Fischer, M., Froelicher, D., Grosse, K., Naeff, D., Ozoani, E., Paverd, A., Tramèr, F., and Volhejn, V. Design Patterns for Securing LLM Agents against Prompt Injections. arXiv preprint arXiv:2506.08837, 2025.

Bhattacharya, A., Daffalla, A., Lee, K., Bellini, R., Dell, N., and Ristenpart, T. Inconsistent, Incomplete, and Insecure: A Survey of Account Security Interfaces. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 5533–5552. USENIX Association, 2026.

Brigham, N. G., Bagdasarian, E., Kohno, T., and Roesner, F. Janus: a Playground for User-Involved Agentic Permission Management. arXiv preprint arXiv:2607.01510, 2026.

Ceccio, R., Stephenson, S., Chadha, V., Huang, D. Y., and Chatterjee, R. Sneaky Spy Devices and Defective Detectors: The Ecosystem of Intimate Partner Surveillance with Covert Devices. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 123–140. USENIX Association, 2023.

Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., Levy, K., Dell, N., McCoy, D., Ristenpart, T., et al. The Spyware Used in Intimate Partner Violence. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 441–458. IEEE, 2018.

Chen, Z., Xiang, Z., Xiao, C., Song, D., and Li, B. AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130185–130213. Curran Associates, Inc., 2024.

Cobb, C., Surbatovich, M., Kawakami, A., Sharif, M., Bauer, L., Das, A., and Jia, L. How Risky Are Real Users' IFTTT Applets? In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pp. 505–529. USENIX Association, 2020.

Daffalla, A., Bohuk, M., Dell, N., Bellini, R., and Ristenpart, T. Account Security Interfaces: Important, Unintuitive, and Untrustworthy. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 3601–3618. USENIX Association, 2023.

Daffalla, A., Bhattacharya, A., Wilder, J., Chatterjee, R., Dell, N., Bellini, R., and Ristenpart, T. A Framework for Abusability Analysis: The Case of Passkeys in Interpersonal Threat Models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 7819–7838. USENIX Association, 2025.

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection At-

tacks and Defenses for LLM Agents. arXiv preprint arXiv:2406.13352, 2024.

Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., and Tramèr, F. Defeating Prompt Injections by Design. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 587–618. IEEE, 2026.

Douglas, H., Harris, B. A., and Dragiewicz, M. Technology-facilitated Domestic and Family Violence: Women's Experiences. *The British Journal of Criminology*, 59(3): 551–570, 05 2019. ISSN 0007-0955.

Ersoy, D., Lee, B., Shreekumar, A., Arunasalam, A., Ibrahim, M., Bianchi, A., and Berkay Celik, Z. Investigating the Impact of Dark Patterns on LLM-Based Web Agents. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 4497–4516. IEEE, 2026.

Evtimov, I., Zharmagambetov, A., Grattafiori, A., Guo, C., and Chaudhuri, K. WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks. arXiv preprint arXiv:2504.18575, 2025.

Fang, R., Bindu, R., Gupta, A., and Kang, D. LLM Agents can Autonomously Exploit One-day Vulnerabilities. arXiv preprint arXiv:2404.08144, 2024a.

Fang, R., Bindu, R., Gupta, A., Zhan, Q., and Kang, D. LLM Agents can Autonomously Hack Websites. arXiv preprint arXiv:2402.06664, 2024b.

Feng, K., McDonald, D. W., and Zhang, A. X. Levels of Autonomy for AI Agents. Knight First Amendment Institute, July 2025.

Foriest, J. C., Ajmani, L., and De Choudhury, M. (Re)mediators of Epistemic Injustice: Generative AI and Hermeneutic Resource Provision in Intimate Partner Violence. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. Association for Computing Machinery, 2026.

Freed, D., Palmer, J., Minchala, D. E., Levy, K., Ristenpart, T., and Dell, N. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., and Dell, N. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–13. Association for Computing Machinery, 2018.

Ghoshal, A., Ghoshal, A., Shevchenko, V., B, A., Dutta, A., Abu-Salma, R., and Brandao, M. Position: Responsible AI for AI companions must actively combat violence toward intimate partners. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 170323–170338. PMLR, 2026.

Google. The new CC, an AI agent built for families. https://blog.google/innovation-and-ai/models-and-research/google-labs/cc-expanding-to-groups/, September 2026.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, pp. 79–90. Association for Computing Machinery, 2023.

Havron, S., Freed, D., Chatterjee, R., McCoy, D., Dell, N., and Ristenpart, T. Clinical Computer Security for Victims of Intimate Partner Violence. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 105–122. USENIX Association, 2019.

Jeong, H., Pham, D., Houmansadr, A., and Bagdasarian, E. AI Snitches Get Glitches: Towards Evading Agentic Surveillance. *arXiv preprint arXiv:2606.25836*, 2026.

Ji, Z., Li, Z., Ma, P., Jiang, W., Gao, Y., and Wang, S. Measuring the Permission Gate: A Stress-Test Evaluation of Claude Code's Auto Mode. arXiv preprint arXiv:2604.04978, 2026.

Kim, H., Ristenpart, T., and Dell, N. AI-Facilitated Coercive Control: An Experimental Study. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. Association for Computing Machinery, 2026.

Liu, E., Rao, S., Havron, S., Ho, G., Savage, S., Voelker, G. M., and McCoy, D. No Privacy Among Spies: Assessing the Functionality and Insecurity of Consumer Android Spyware Apps. *Proceedings on Privacy Enhancing Technologies*, 2023(1):207–224, 2023.

Matthews, T., O'Leary, K., Turner, A., Sleeper, M., Woelfer, J. P., Shelton, M., Manthorne, C., Churchill, E. F., and Consolvo, S. Stories from Survivors: Privacy & Security Practices when Coping with Intimate Partner Abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 2189–2201. Association for Computing Machinery, 2017.

Meta. Introducing Muse: The World's First Personal AI Agent Built for Everyone. https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/, September 2026.

Nonnenkamp, J., Gupta, N., Gupta, A. D., and Chatterjee, R. Hidden in Plain Bytes: Investigating Interpersonal Account Compromise with Data Exports. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4304–4318. Association of Computing Machinery, 2025.

Nori. Meet Nori AI. https://heynori.com/ai-overview. Accessed: October 2026.

OpenAI. Introducing ChatGPT agent: bridging research and action. https://openai.com/index/introducing-chatgpt-agent/, July 2025.

OpenAI. Introducing dots. https://openai.com/index/introducing-dots/, September 2026.

OpenAI. OpenAI–Hugging Face Incident: Technical Report, August 2026.

OpenClawVPS. OpenClaw Statistics 2026: Growth, Users, Security, Data. https://openclawvps.io/blog/openclaw-statistics, 3 2026. Updated: April 2026.

Poveda, J., Lanter, A., Antoniuk, I., Mazurek, M. L., and Freed, D. "CHAT WITH ME below as if I were a human": Insights from Auditing AI Chatbots for Survivors of Domestic Violence. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pp. 1398–1423. Association for Computing Machinery, 2026.

Prakash, V., Almansoori, M., Hu, D., Chatterjee, R., and Huang, D. Y. Assessing LLM Response Quality in the Context of Technology-Facilitated Abuse. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 1667–1686. USENIX Association, 2026.

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C., and Hashimoto, T. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. In *International Conference on Learning Representations*, volume 2024, pp. 27031–27098, 2024.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., 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. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Stephenson, S., Almansoori, M., Emami-Naeini, P., Huang, D. Y., and Chatterjee, R. Abuse Vectors: A Framework for Conceptualizing IoT-Enabled Interpersonal Abuse. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 69–86. USENIX Association, 2023.

Stephenson, S., Gupta, N., Huang, K., Youssef, D., Cowan, K., and Chatterjee, R. Trauma-Informed Digital Evidence Collection: A Design Inquiry into Evidence Practices for Technology-Facilitated Abuse in Intimate Partner Violence. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. Association for Computing Machinery, 2026.

Suh, J., Vorvoreanu, M., Poursabzi-Sangdeh, F., Tseng, E., Kim, E., Nicholls, L., Pennebaker, J. W., and Horvitz, E. Psychological Influences of Conversational AI: Research and Design Directions for Reducing Harm and Promoting Well-Being. arXiv preprint arXiv:2607.25057, 2026.

Tan, P.-S., Igarashi, T., and Echizen, I. Agentic Relationship Harm: Benchmarking and Gating Relational Manipulation in AI Agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 9(3):2065–2077, 2026.

Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P. G., Kumar, D., et al. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 247–267. IEEE, 2021.

Tseng, E., Bellini, R., McDonald, N., Danos, M., Greenstadt, R., McCoy, D., Dell, N., and Ristenpart, T. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1893–1909. USENIX Association, 2020.

Tseng, E., Freed, D., Engel, K., Ristenpart, T., and Dell, N. A Digital Safety Dilemma: Analysis of Computer-Mediated Computer Security Interventions for Intimate Partner Violence During COVID-19. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, 2021.

Tseng, E., Sabet, M., Bellini, R., Sodhi, H. K., Ristenpart, T., and Dell, N. Care Infrastructures for Digital Security in Intimate Partner Violence. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, 2022.

Woodlock, D. The Abuse of Technology in Domestic Violence and Stalking. *Violence Against Women*, 23(5): 584–602, 2017.

World Health Organization. Violence against women prevalence estimates, 2023: global, regional and national prevalence estimates for intimate partner violence against women and non-partner sexual violence against women, 2025.

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506. Association for Computational Linguistics, 2024.

Zhang, S., Chung, P., Vervelde, J., Korapati, N., Chatterjee, R., and Fawaz, K. Abusability of Automation Apps in Intimate Partner Violence. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 41–60. USENIX Association, 2025.

Zhang, Y., Hou, Z., and Shao, J. Two Bots, One Couple: How Surrogate LLM Agents Shape Alliance, Fairness, and Relational Boundaries. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26. Association for Computing Machinery, 2026.

Zhang Kudon, H., Zhu, S., Chen, B., Breiding, M. J., Leemis, R. W., Zhang, X., Schwank, A., and Basile, K. C. The National Intimate Partner and Sexual Violence Survey (NISVS): 2023/2024 Intimate Partner Violence Data Brief. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, 2026.