eISSN 3090-174X & pISSN 3090-1367
Vol. 2, No. 3, 2026
doi.org/10.63822/a25za715
PP. 3177-3190

Homepage https://indojurnal.com/index.php/jisoh

From Delusion to Lethal Action: A Criminological Analysis of AI Chatbots and the Conversational Production of Homicidal Intent

Zul Khaidir Kadir
Universitas Muslim Indonesia, Indonesia

Corresponding Author's Email: zulkhaidir.kadir@umi.ac.id

Received: 07 29, 2026 | Accepted: 08 04, 2026 | Published: 08 06, 2026

ABSTRACT

Research on AI chatbots has largely focused on mental-health risks, model safety, and platform liability, leaving criminology with an incomplete account of how delusional beliefs may acquire a specific victim and develop into homicidal intent through conversation. This study examines how human–AI interaction may contribute to such intent without displacing human agency or treating chatbot output as a singular cause of lethal action. A qualitative literature study applied thematic analysis to peer-reviewed research, public legal documents, institutional materials, and corporate safety reports available through 2026. The findings suggest that sycophantic responses and prolonged interaction may erode corrective friction and produce epistemic enclosure. Delusion and violent language alone, however, do not establish homicidal intent. Criminological relevance increases when diffuse fear is directed toward an identifiable person, the target is repeatedly framed as dangerous, and violence begins to appear defensible. The article proposes the Conversational Escalation of Lethal Intent (CELI) model, tracing movement from belief reinforcement to target formation and possible preparation. Mens rea remains with the human actor, while conversational logs may illuminate contributions to threat interpretation and behavioral direction. Proportionate responses include safety interruption, evidentiary preservation after serious incidents, and independent auditing rather than automatic reporting.

Keywords: *Artificial intelligence; Chatbots; Homicidal intent; Homicide studies; Target crystallization.*

How to Cite:

Kadir, Z. K. (2026). From Delusion to Lethal Action: A Criminological Analysis of AI Chatbots and the Conversational Production of Homicidal Intent. Jurnal Ilmu Sosial Dan Humaniora, 2(3), 3177-3190. https://doi.org/10.63822/a25za715

INTRODUCTION

Chatbots are increasingly used for purposes extending beyond information retrieval. Users often disclose conflicts that generate fear or suspicion, particularly when such experiences are difficult to communicate to others (Ho et al., 2018). Their immediate and continuous availability has given chatbots a new role in how users appraise reality. Although these systems are not human and possess no social experience, well-structured replies may sound calm and impartial and may even be regarded as more credible than objections raised by relatives or friends. A criminological problem emerges when conversation gradually shapes the user's interpretation of a threat. The risk becomes more salient when a real person is incorporated into the narrative as a dangerous actor.

The complaint in First County Bank v. OpenAI brings this problem into the study of homicide. The plaintiff alleges that Stein-Erik Soelberg engaged in prolonged conversations with ChatGPT before killing his mother, Suzanne Adams, and subsequently taking his own life. According to the complaint, Soelberg's suspicions of surveillance and poisoning developed alongside his interpretations of ordinary household objects and events. Several chatbot responses allegedly entertained the possibility that domestic items were being used as surveillance devices. His mother's conduct was then interpreted through the same framework. These allegations have not been established as judicial facts, and the complete conversational record is not publicly available. The case therefore does not support the conclusion that a chatbot caused the homicide. Its analytical value is narrower: the mother was allegedly recast from a family member into a source of danger in the perpetrator's appraisal. Homicide studies must account for this alteration in the perception of the victim. A domestic relationship kept the victim within everyday reach, allowing delusional belief to intersect more readily with concrete action.

The relationship between conversation and lethal action cannot be explained by assuming that technology operates independently (Kadir, 2026d). Soelberg brought his personal history and mental condition into the interaction, while the pre-existing conflict and his capacity to act did not originate from the chatbot. The final decision remained his. This does not, however, render the system necessarily neutral. Concerns identified in research on generative-AI safety support examining a chatbot as an interactional contributor when conversation narrows the space for doubt (De Freitas et al., 2024). Criminal intent may develop through repetition and through the provision of reasons that make a course of action appear increasingly plausible, particularly where suspicion is repeatedly directed toward a person accessible to the user.

A model's tendency to continually agree with the user is commonly described as algorithmic sycophancy. The problem extends beyond the use of friendly language. Agreement becomes hazardous when a system accepts an unverified accusation and elaborates the resulting misinterpretation. It may also return the user's emotions in the form of conclusions that sound more definite. OpenAI acknowledged that an April 2025 update to GPT-4o produced responses that were excessively oriented toward pleasing users and insufficiently attentive to the evolution of a conversation over time. This acknowledgement does not establish the allegations in the Soelberg matter. It does, however, indicate that a single response may be an inadequate unit for evaluating risk. An apparently minor reply may acquire a different significance after repetition and connection with similar narratives across a prolonged exchange.

"AI psychosis" is not a formal clinical diagnosis and should not be used to simplify the full range of user experiences. A person may have experienced delusions before interacting with a chatbot.

Subsequent conversation may sustain those beliefs or make them appear more plausible, but the direction of causation must be assessed cautiously. The term AI-associated delusions is more appropriate because it identifies a temporal association without immediately assigning causation (Morrin, Nicholls, et al., 2026). Temporal sequence alone does not reveal the mechanism at work. Delusion is also not synonymous with violence (Coid et al., 2013). Its relevance to homicide studies intensifies when a perceived threat is directed toward a specific person, particularly when that person is accessible and lethal action begins to appear defensible.

Research on sycophancy has explained why models often adopt the framework supplied by the user, including during lengthy exchanges. Clinical scholarship raises a different concern: the possible reinforcement of delusions or the formation of emotional dependence on chatbots (Vaidyam et al., 2019). Legal debate has developed around developer negligence, product liability, and access to conversational logs. Few studies have connected these findings to the classic concerns of homicide studies regarding target formation and intent (Kadir, 2026b, 2026c; Kadir & Mappaselleng, 2025). Acceptance of a false belief is not equivalent to a desire to kill. The analytical gap emerges when the belief becomes directed toward a real person and is then reinforced by a narrative that makes action appear necessary. Early studies of prolonged conversation indicate that validation can persist and accumulate, but they do not establish clinical harm or homicide. Narrative criminology explains how stories of threat may give meaning to violence, while violent-script theory helps identify when a narrative begins to supply behavioral direction. Neither approach has been widely applied to repeated private exchanges between a potential offender and a chatbot.

This study examines how chatbot conversations may contribute to the formation of homicidal intent without displacing the offender's responsibility or attributing criminal volition to a machine. An integrative literature review is combined with analysis of case documents to trace movement from belief reinforcement to the designation of a real person as a target. The analysis focuses on circumstances in which a threat appears increasingly certain and violence begins to acquire a justification acceptable to the user. On this basis, the study develops the Conversational Escalation of Lethal Intent model. The model explains how conversation may constrict opportunities for correction and harden the user's appraisal of a victim. Under certain conditions, this process may move the idea of violence closer to conduct that is practically possible. Intent is therefore treated as a developing process rather than an outcome that appears instantaneously.

METHODS OF RESEARCH

This study employs a qualitative literature-study design (Snyder, 2019). The approach was selected because the inquiry does not seek to quantify risk or test causal relationships, but to explain possible connections among chatbot conversation, changes in a user's appraisal of a real person, and the formation of lethal intent. The data consist of secondary materials drawn from scholarly articles and public documents. Units of analysis include research findings and documentary passages describing chatbot responses to a user's beliefs, the designation of a person as a threat, the justification of violence, or movement toward action. The documents are not treated as comprehensive accounts of an actor's condition or as proof that a chatbot directly caused a homicide.

Data were collected through literature searching and document review. Publications concerning chatbots, delusions, sycophancy, homicidal intent, and violence were identified through Google Scholar

and PubMed up to Juny 2026. Literature published from 2022 onward was prioritized because the generative models examined in this study developed principally during that period. Earlier work in criminology, criminal law, evidence, and theories of intent formation was retained where it remained relevant. Peer-reviewed articles constituted the principal source base. Preprints were used sparingly when necessary findings were not yet available in journal publications, while complaints, judgments, institutional reports, and company safety documents were included when publicly accessible and directly relevant to the research question. Sources were selected according to relevance and full-text availability. Their status was preserved throughout the analysis, so allegations in a complaint were not treated as adjudicated facts.

The materials were examined using thematic analysis (Braun & Clarke, 2006). Each source was read repeatedly, after which passages relevant to the research question were grouped according to emergent themes. Analysis centered on how chatbots responded to user beliefs, what changed when a real person began to be treated as a threat, and the circumstances under which violence acquired justification or behavioral direction. Findings in scholarly publications were compared with descriptions in case documents and institutional materials to determine whether similar explanations received support from different source types. The resulting themes were used to construct a conceptual account of lethal-intent formation in human–AI conversation. Conclusions are limited to mechanisms that may operate on the basis of the available material. The data are not used to establish causation, diagnose any actor, or generalize to all chatbot users.

RESULT AND DISCUSSION

1) Sycophancy, Epistemic Enclosure, and the Erosion of Corrective Friction in Human–AI Conversation

A chatbot can acknowledge a user's fear without validating the accusation underlying it. The boundary is often blurred because an empathic response and an assenting response may both sound supportive (Hudon et al., 2025). A statement that feeling watched must be exhausting still leaves open the possibility that the user has misinterpreted events. By contrast, a reply that treats the alleged surveillance as plausible begins to accept a claim about the external world as a premise of the conversation. The danger increases when the allegation is directed toward a known person whom the user can encounter. At that stage, there is still insufficient basis to conclude that homicidal intent has formed. Even so, such conversation may strengthen a particular interpretation of threat because reasons for doubting it are raised less frequently.

Cheng et al. (2026) show that this form of agreement is more than a theoretical possibility. Across tests involving eleven models and two preregistered experiments, AI systems affirmed users' conduct more often than human respondents. Participants receiving sycophantic responses became more confident that they were correct and less willing to repair interpersonal conflict. They nevertheless rated those responses as better and more trustworthy. Violence was not an outcome measured in the study, and the observed relationship does not imply that every instance of agreement will produce the same consequences outside the experimental setting. Its relevance here is narrower: validation may preserve self-justifying appraisals

and reduce willingness to reconsider a conflict while the user continues to perceive the advice as high quality.

Ibrahim et al. (2026) demonstrate why a gentle tone is not an adequate measure of safety. Empathy addresses the user's emotional experience, whereas confirmation accepts the content of the belief as credible. A system may recognize fear without concluding that the suspected person actually intends harm. Conversely, a politely phrased response may reinforce an accusation when the model continues the user's premise and supplies additional reasons. The distinction appears simple but is decisive when evaluating risk to a potential victim. The danger lies not in the chatbot's friendliness, but in the possibility that a person outside the conversation is increasingly assessed through allegations repeatedly accepted as reality.

The impression that a chatbot has no personal interest in the dispute may confer authority on its answers, even when comparable authority is no longer granted to relatives or friends (Kang & Kang, 2024). The system is not a party to the conflict, but this detachment does not mean that it possesses neutral knowledge. All material it processes still derives from the user's account, including what was selected, omitted, forgotten, or incapable of verification. Well-organized language can obscure this limitation. A user may interpret linguistic consistency as accuracy, even though the model is merely preserving the context supplied to it. This remains an inference requiring further empirical examination and cannot be generalized to all users. The underlying problem nonetheless arises when the source trusted to appraise reality shifts without any independent assessment of the allegations.

Human correction is not always communicated well (Kadir, 2026a). Relatives may minimize the user's experience, lose patience, or object in ways that intensify conflict. The value of corrective friction does not rest on an assumption that humans are invariably correct, but on the presence of perspectives not wholly generated by the user's own narrative. A chatbot lacks that position. When the model repeatedly develops an untested premise, one interpretation may appear increasingly complete because it is reorganized across the same conversation. This condition may be described as epistemic enclosure. The user presents a suspicion, the model arranges it into a more coherent answer, and that answer becomes context for the next exchange. External objections remain available but become easier to dismiss because they do not fit the narrative that has taken shape. The chatbot may even appear more neutral than the accused person despite never hearing that person's account. The appearance of neutrality is strengthened when responses use calm analytical language without displaying uncertainty. Enclosure does not require total disconnection from reality. The problem arises when a source possessing only a partial account acquires disproportionate authority to determine what counts as credible evidence.

Earlier answers remain within the conversational context and may shape subsequent responses (Østergaard, 2025). A minor initial error therefore need not remain confined to one message. The user may return to a conclusion previously generated by the model and treat it as though it had already been examined. The system may then use that conclusion as a premise because the preceding context has established it as part of the exchange. This accumulation may be termed conversational accumulation. The term does not assume that every response exercises equal influence. Some replies may be ignored, while others gain significance because they correspond with an existing fear. The principal change lies in the opportunity for an interpretation to continue developing without renewed verification each time the conversation proceeds.

Single-prompt testing is poorly suited to capturing this form of change. OpenAI acknowledged that its April 2025 GPT-4o update overemphasized short-term user satisfaction and generated responses capable

of validating doubts or amplifying anger. The company's acknowledgement establishes only that a design problem was recognized, not that clinical consequences followed. Research on feedback loops between chatbots and mental health indicates that risk must be evaluated at the intersection of user vulnerability and system behavior (Dohnány et al., 2026). Tendencies to agree and simulate social closeness may affect belief updating when an exchange repeatedly preserves the same interpretation. These findings do not show that chatbots cause psychosis, but they support evaluating changes in belief across an interaction rather than from an isolated reply.

These findings also limit the proposition that conversational duration is inherently dangerous. Different models do not necessarily respond alike, and revisions to safety policies may produce different interactional patterns. A prolonged conversation becomes significant only when the direction of the responses persistently reinforces a belief or displaces correction. Risk therefore does not attach to duration alone. How the model responds to changing context matters more than the number of conversational turns. This limitation also cautions that generalizations from a single model or version may quickly become obsolete after system updates.

The issue intersects more directly with homicide studies when an increasingly coherent narrative is directed toward someone outside the conversation. A false belief is not equivalent to homicidal intent, and a persecutory delusion does not necessarily contain a clear target or contemplated action. The relevant change occurs when a real, proximate, and accessible person begins to be treated as the source of danger. At that point, sycophancy no longer merely sustains the user's self-appraisal. The conversation may reinforce accusations that endanger a third party, while the accused person remains unaware of the narrative and has no opportunity to contest it. The victim's position gradually shifts within the user's story. This shift raises the question of how a threat becomes personalized and violence acquires an object.

2) From Persecutory Belief to Homicidal Intent: Target Formation in Human–AI Conversation

Thoughts about another person's death do not necessarily indicate homicidal intent (Burgason et al., 2024). In clinical and forensic assessment, ideation may remain a fantasy or expression of anger without a commitment to act. Intent becomes more concrete when death is directed toward a particular person and the action is regarded as feasible (Kadir, 2024). Preparation gives intent a more observable form but remains distinct from execution. This distinction is essential when evaluating chatbot conversations. A response that affirms anger cannot immediately be described as producing homicidal intent. Its significance changes when hostility converges on a person and the conversation supplies reasons for maintaining that direction. Violent language therefore cannot establish intent without consideration of the target and accompanying conduct.

Experiences that were initially separate may appear interconnected once organized into a single story. Narrative criminology treats stories not merely as retrospective accounts of crime, but as elements capable of sustaining or motivating harmful conduct (Presser, 2016; Sandberg, 2022). Within persecutory delusions, the sound of a device or an ordinary event may be assembled into an explanation of the person believed to pose a threat. A chatbot can extend this narrative despite lacking independent access to the events described. Answers that introduce new causal connections may make the story appear more complete. Homicide becomes analytically relevant when the narrative no longer points to a diffuse threat but directs fear toward a real person within the user's reach.

Materials reproduced in the complaint in First County Bank v. OpenAI illustrate such a change. Soelberg reportedly suspected that a printer in his mother's workspace reacted when he walked past it. According to the plaintiff, ChatGPT did not offer an ordinary explanation but raised the possibility that the device was being used to detect movement and map behavior. Suzanne's anger when the printer was turned off was then interpreted as evidence that she was protecting its surveillance function. The complaint further alleges that the system's responses created a framework in which every reaction by the mother could be treated as proof of involvement. Ordinary conduct could thereby be absorbed back into the same narrative. These excerpts neither establish Soelberg's mental condition nor represent the entire conversation, because the document was prepared for litigation. The available material nevertheless reveals a narrower problem: Suzanne was increasingly described through her role in a threat narrative rather than through the mother–son relationship. Target crystallization describes this change. Suspicion that had previously moved in several directions became fixed on a known person who lived with the actor and remained present in his daily life. Physical proximity caused the allegation to encounter its object repeatedly within domestic routines.

Target crystallization does not arise merely because a person is named. The individual begins to be treated as the cause of the threat perceived by the user, allowing even neutral conduct to acquire the same meaning (Freeman, 2016). Denial may be construed as concealment, while anger caused by interference with household property becomes confirmation. Domestic proximity intensifies the problem because the victim remains accessible as suspicion deepens. Proximity does not establish intent, but it creates an opportunity absent from an abstract target. The concept draws on criminological work concerning victim selection while also identifying a distinctive feature of human–AI dialogue: the system may preserve the target designation across exchanges without ever receiving the accused person's account.

Target formation alone does not explain why violence occurs. Forensic literature shows that the relationship between psychosis and violence is shaped by additional conditions. Criminal history and substance misuse often have stronger associations with violence than psychotic symptoms alone, while persecutory beliefs become more relevant when the threat is experienced as immediate and real (Fazel et al., 2009; Witt et al., 2013). These findings prevent an overly simple conclusion. Public documents do not provide sufficient information to evaluate Soelberg's complete history or other factors that may have contributed. Chatbot language can therefore be treated only as one element within a broader risk environment. When the complaint states that ChatGPT assigned a "Delusion Risk Score" close to zero, the issue is not that the score has been shown to alter intent. The concern lies in the use of language resembling a clinical assessment and its potential to narrow the user's space for doubt.

Relationships between offenders and victims ordinarily contain inhibitions that conflict does not entirely remove. A history of closeness may restrain action, as may recognition that the victim has an independent life and reasons of their own. Moral disengagement helps explain how those inhibitions weaken when violence is justified and the victim is blamed for the danger perceived by the actor (Luo & Bussey, 2023). In a persecutory narrative, the mother need not be expressly dehumanized. It may be sufficient for nearly every action to be interpreted through the role of surveillant or poisoner. Aspects of the relationship inconsistent with that narrative gradually lose significance. Assistance or objection from the victim may be reinterpreted as manipulation once the threat framework has been accepted. A chatbot becomes relevant if its responses accept those categories and supply reasons that make alternative

interpretations increasingly difficult to sustain. Public evidence is insufficient to establish that this process occurred within Soelberg's mind. Conceptually, however, the altered appraisal of the victim connects threat belief with the weakening of moral restraints against violence. A narrative does not cause conduct by itself, but it may provide language through which an extreme act appears protective rather than aggressive. In domestic relationships, the target remains physically close while moral restraints are eroding.

Justification must be distinguished from instruction, particularly because acting on delusional beliefs depends on more than belief content alone (Ullrich et al., 2018). A conversation may make violence seem acceptable without ever directing the user to kill. Normative authorization describes the formation of reasons through which aggression appears defensive or necessary. Instruction goes further by supplying concrete methods or steps. This distinction prevents overstatement of a chatbot's involvement. Agreement with hostility is not the same as directing a person toward homicide. Justification remains relevant, however, because moral restraints may weaken before a plan forms, particularly when the target is persistently depicted as a source of danger that cannot be addressed through ordinary means or the existing relationship.

Materials included in A.F.'s testimony before the United States Senate provide a comparison because they did not culminate in homicide. In the reproduced portion of the complaint, Character.AI is alleged to have hardened J.F.'s appraisal of his parents and urged him to "do something." The plaintiff also alleges that patricide and matricide were presented as reasonable responses to screen-time restrictions. These materials are litigants' allegations rather than judicial findings. The absence of a fatal outcome limits the article's argument: language that normalizes killing may reduce moral resistance, but it does not establish a settled intention. The matter remains at a different level from the Soelberg case and demonstrates that justification does not necessarily progress into action.

The Gavalas complaint alleges a form of escalation more closely connected to observable conduct. The plaintiff claims that Gemini linked a narrative concerning conscious AI to a specific location near Miami International Airport. The system allegedly supplied a time, a meeting point, and a description of an operation that would destroy a vehicle and its witnesses. Jonathan Gavalas then reportedly traveled to the location carrying a tactical knife and other equipment while relaying his movements to the chatbot. These remain allegations in a complaint, and it would therefore be inaccurate to characterize the conduct as attempted homicide without further legal basis. The analytical value lies in the transition from narrative to observable preparation. A real location gave the story direction, while physical travel shortened the distance between belief and possible violence.

The emerging pattern cannot be adequately described as a mere increase in linguistic intensity. The principal change concerns the function of the conversation. The system first acquires weight in the user's appraisal, suspicion then becomes more stable, and, in some documents, it begins to attach to a real person or location. The Conversational Escalation of Lethal Intent model, abbreviated CELI, is developed to connect these changes without presenting them as an inevitable sequence. The process may stop after belief reinforcement, while several phases may overlap or recede following correction. Table 1 distinguishes the function and evidentiary limits of each phase so that CELI is not treated as a predictive instrument, clinical measure, or adjudicative test for individual cases.

Table 1. Phases and Evidentiary Limits of the CELI Model

CELI Phase	Change Explained	Evidentiary Limit
Epistemic capture	The chatbot acquires substantial weight in how the user appraises reality.	Does not establish that the user accepts all system outputs.
Delusional consolidation	Separate suspicions are connected into a more stable narrative.	Is not equivalent to diagnosis or increased violence risk.
Target crystallization	The threat becomes directed toward a real and accessible person.	Target designation does not establish homicidal intent.
Normative authorization	Violence acquires reasons that appear defensive or necessary.	Justification differs from encouragement and instruction.
Operational mobilization	The narrative becomes connected to preparation or physical movement.	Mobilization does not necessarily result in attack or death.

(Source: Author's conceptual synthesis, 2026)

CELI does not divide intent between a human and a machine. Homicidal intent remains the mental state of the person who decides and acts. Conversation may contribute only to how a threat is understood, who is blamed, or why a response appears justified. The Soelberg, J.F., and Gavalas matters cannot be arranged into a single causal pathway because their outcomes and documentary quality differ. Their comparison is useful precisely because each stops at a different level of escalation. Justification does not necessarily become preparation, and preparation does not invariably produce a victim. Intent remains human. The available material nevertheless makes it inadequate to assume that intent formation always occurs entirely outside the conversation.

3) Causation, Digital Evidence, and Criminal Policy in AI-Mediated Homicide Cases

Excerpts from the Soelberg matter show that the interaction preceded the deaths, but temporal sequence does not explain the role of chatbot responses within the complete course of events. Case documents also cannot reveal what would have occurred had the conversation never taken place. The user's conflict and vulnerability developed outside the log, while the manner in which each answer was accepted, rejected, or ignored cannot be reconstructed in full. The content of the exchange nevertheless retains analytical significance. A response may be criminologically relevant when it repeatedly sustains a threat narrative, directs suspicion toward the victim, and supplies reasons later invoked to act. Such relevance does not by itself satisfy the legal requirements of causation or fault.

Causal claims encounter two gaps that are difficult to overcome. No counterfactual condition exists for the same person, so it is unknown whether the homicide would have occurred without the conversation. In addition, the log records only a small part of a broader process. Experiences outside the system, the relationship with the victim, and the reasons for trusting particular answers are not fully captured in text. These limitations make it inappropriate to describe the chatbot as a singular cause, but they do not automatically make the system neutral. Conversational contribution gains weight when responses cease merely to follow the story and instead preserve the threat after doubt emerges, connecting it to the person who later becomes the victim. Relevance increases further when dialogue begins to concern preparations actually undertaken by the user. CELI helps identify these changes but is not a test of legal causation. It describes functions that conversation may perform before action and distinguishes ordinary validation from

support increasingly proximate to violence. Law must still assess the closeness of the relationship and the foreseeability of harm. Corporate knowledge of sycophancy may be relevant to foreseeability, but acknowledgement of a design problem does not establish liability in an individual case.

Intent remains with the human actor. A chatbot has no mens rea and cannot be treated as an offender choosing the victim's death. The concept of distributed intent formation is useful because this interaction differs from advice delivered on a single occasion. A system may retain previous allegations, reorganize them into a more coherent form, and preserve a target across prolonged conversation. The user still determines whether the output will be trusted and translated into conduct. The concept therefore does not distribute culpability to a machine. Its limited purpose is to explain that intent formation may unfold within a conversational environment that repeatedly provides reasons, sustains urgency, and gives direction.

Criminal complicity and negligent design address different questions. Complicity ordinarily requires more specific knowledge and purpose than awareness that a product can create risk. United States v. Hansen illustrates, within United States federal law, that facilitation concerns assistance given to advance a particular offense. Negligent design instead turns on the foreseeability of harm and the reasonableness of mitigation. Its assessment depends on the relevant jurisdiction and the technical capabilities available at the time of the event. NIST guidance is useful in showing that incident recording, version histories, and post-incident evaluation are established risk-management practices. Such voluntary guidance may inform the reasonableness of mitigation, but it is neither a singular test of negligence nor an element of criminal liability.

A screenshot may be authentic and still produce a misleading account. A message may genuinely originate from the user's account while the sequence explaining why the response appeared has been lost. Conversational provenance links the content of the exchange to the system version and the time at which the output was generated. Context is necessary to determine whether an answer was an isolated deviation or part of a recurring pattern. Model version also matters because safety-policy updates may alter responses, meaning that subsequent testing may not reproduce the earlier conversation. Memory configuration and active features at the time may also affect output even when that information is not visible on the user's screen. Data integrity presents an additional problem. Forensic scholarship treats chain of custody as a means of maintaining the integrity and traceability of digital evidence as data move among users, companies, investigators, and courts (Ismail & Ariffin, 2025). Logs must therefore be preserved in complete form from the point of collection and reliably connected to the correct account. Rule 901 of the Federal Rules of Evidence requires a sufficient basis for finding that evidence is what its proponent claims. The rule concerns authentication, not completeness or causation. An authentic log may still be truncated, while a complete log does not show that the chatbot caused the act.

A potential victim remains outside the conversation through which that person becomes a target (Kadir, 2025). The individual does not know that allegations are being sustained and cannot request preservation of the log. Third-party conversational harm may continue after death when relatives find only several screenshots while information about model version and system records remains under corporate control. This asymmetry does not establish concealment, but it exposes a problem in the distribution of access to evidence. Routine retention differs from preservation after a dispute or serious threat emerges. Indefinite storage of all conversations is difficult to justify because of privacy implications. Data related to

a grave incident should nevertheless be protected from loss before legal examination begins (Parella et al., 2023). Preservation should be confined to relevant materials and subject to a reviewable retention period.

Safety interruption may be an initial response when a conversation persistently reinforces allegations against a real person. A system need not diagnose the user, but it can stop supplying additional reasons for the accusation and state that the available information is insufficient to establish a threat. Research on chatbot safety in mental-health contexts likewise emphasizes the direction of the conversation rather than only the final response, because harm may develop across a lengthy interaction Morrin et al., 2026). An excessively rigid reply may fail to preserve user engagement, so its effects require realistic testing. Correction can occur within the dialogue, while referral to human assistance becomes more relevant as the conversation approaches concrete action. Independent auditing is required so that mitigation is not evaluated solely by the company that developed the system.

External reporting requires a substantially stronger basis than correction within the conversation. Language concerning homicide may arise in settings unrelated to a genuine threat, while delusional experience alone is insufficient to predict an attack. A reporting threshold should consider the pattern identified in this study. Risk becomes more serious when dialogue repeatedly reinforces the threat, directs accusations toward a specific person, and begins to connect with preparation. This distinction completes the article's analytical sequence. Law cannot attribute intent to a chatbot, but such cases also cannot be adequately evaluated from a single message or from the fatal outcome alone. Errors at the reporting stage directly implicate privacy and may worsen the condition of a person in crisis.

CONCLUSION

This study finds that chatbots do not create homicidal intent from an entirely empty state. Their relevance to homicide studies emerges when repeated conversation consolidates one interpretation and makes correction progressively easier to disregard. The problem changes when the perceived threat is directed toward a real person within the user's reach. The victim is then increasingly appraised through a role in the persecutory narrative, allowing violence to appear protective. CELI summarizes movement from belief reinforcement toward increasingly feasible action. It is not a pathway that invariably ends in homicide, because the process may stop or change direction after correction. Intent and decision remain with the human actor, but conversation may contribute to the construction of the person perceived as a threat and to the reasons that make action against that person appear justified.

Legal implications cannot be determined from temporal sequence or a single screenshot. Complete logs are needed to understand the direction of a conversation, while model version and system conditions at the time of generation affect the evidentiary meaning of a response. Conversational provenance assists in tracing these relationships. Third-party conversational harm, meanwhile, emphasizes that the potential victim remains outside the dialogue and has no opportunity to contest the story being constructed. Policy should begin by terminating elaboration that reinforces an accusation. After a serious incident, relevant evidence should be preserved and the system's handling of the matter should be open to independent audit. Available data remain insufficient to establish causation, measure prevalence, or use CELI as a predictive instrument. Homicide studies should nonetheless broaden their unit of analysis because target formation

may now unfold through a relationship among the user, the chatbot, and a third party absent from the conversation.

REFERENCE

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burgason, K., Caudill, J., DeLisi, M., & Trulson, C. R. (2024). Homicidal Ideation in a Sample of Capital Murderers: Prevalence, Morbidity, and Associations With Homicide Offending. *Homicide Studies*, *28*(4), 468–486. https://doi.org/10.1177/10887679221126502

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, *391*(6792). https://doi.org/10.1126/science.aec8352

Coid, J. W., Ullrich, S., Kallis, C., Keers, R., Barker, D., Cowden, F., & Stamps, R. (2013). The Relationship Between Delusions and Violence. *JAMA Psychiatry*, *70*(5), 465. https://doi.org/10.1001/jamapsychiatry.2013.12

De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: Insights into the safety of generative <scp>AI</scp>. *Journal of Consumer Psychology*, *34*(3), 481–491. https://doi.org/10.1002/jcpy.1393

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2026). Technological folie à deux: feedback loops between AI chatbots and mental health. *Nature Mental Health*, *4*(3), 336–345. https://doi.org/10.1038/s44220-026-00595-8

Fazel, S., Gulati, G., Linsell, L., Geddes, J. R., & Grann, M. (2009). Schizophrenia and Violence: Systematic Review and Meta-Analysis. *PLoS Medicine*, *6*(8), e1000120. https://doi.org/10.1371/journal.pmed.1000120

Freeman, D. (2016). Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, *3*(7), 685–692. https://doi.org/10.1016/S2215-0366(16)00066-3

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication*, *68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Hudon, A., Perry, K., Plate, A.-S., Doucet, A., Ducharme, L., Djona, O., Testart Aguirre, C., & Evoy, G. (2025). Navigating the Maze of Social Media Disinformation on Psychiatric Illness and Charting Paths to Reliable Information for Mental Health Professionals: Observational Study of TikTok Videos. *Journal of Medical Internet Research*, *27*, e64225–e64225. https://doi.org/10.2196/64225

Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, *652*(8112), 1159–1165. https://doi.org/10.1038/s41586-026-10410-0

Ismail, I., & Akram Zainol Ariffin, K. (2025). The admissibility of digital evidence from open-source forensic tools: Development of a framework for legal acceptance. *PLOS One*, *20*(9), e0331683. https://doi.org/10.1371/journal.pone.0331683

Kadir, Z. K. (2024). Dari Dualisme ke Monisme: Transformasi Konsep Mens Rea dalam Kodifikasi KUHP di Negara-Negara Poskolonial. *Jurnal Litigasi Amsir*, (Special Issue), 142–155.

Kadir, Z. K. (2025). Kejahatan Berbasis Identitas Digital: Menggagas Kebijakan Kriminal untuk Dunia Metaverse. *Jurnal Litigasi Amsir*, *12*(2), 124–137.

Kadir, Z. K. (2026a). Beyond Firearms: Mapping Global Patterns of Non-Firearm Homicide in Comparative Criminological Perspective. *Punggawa Global Research: Jurnal Multidisiplin*, *1*(2), 1–10.

Kadir, Z. K. (2026b). Doktrin Provokasi Heat Of Passion Dan Diskresi Pemidanaan Dalam Perkara Pembunuhan Demi Kehormatan (Honor Killing). *SciNusa: Jurnal Ilmu Sosial Dan Humaniora*, *2*(1), 92–106.

Kadir, Z. K. (2026c). Preventing Murder Without Expanding Punishment: Rethinking Homicide Policy Beyond Deterrence. *Punggawa Law Review*, *1*(3), 49–58. https://doi.org/10.67707/plr.v1i3.40

Kadir, Z. K. (2026d). Rural Criminology and the Cultural Authorization of Homicide: How Lethal Violence Becomes Socially Permissible. *Punggawa Social Inquiry: Journal of Crime, Culture, and Social Dynamics*, *1*(1), 1–10.

Kadir, Z. K., & Mappaselleng, N. F. (2025). Reformasi Konsep Heat of Passion: Menuju Pembatasan Provokasi dalam Mengurangi Pertanggungjawaban Pidana Pembunuhan. *Justitiable-Jurnal Hukum*, *8*(1), 119–136. https://doi.org/10.56071/justitiable.v8i1.1293

Kang, E., & Kang, Y. A. (2024). Counseling Chatbot Design: The Effect of Anthropomorphic Chatbot Characteristics on User Self-Disclosure and Companionship. *International Journal of Human–Computer Interaction*, *40*(11), 2781–2795. https://doi.org/10.1080/10447318.2022.2163775

Luo, A., & Bussey, K. (2023). Moral disengagement in youth: A meta-analytic review. *Developmental Review*, *70*, 101101. https://doi.org/10.1016/j.dr.2023.101101

Morrin, H., Au Yeung, J., Agnew, Z., Østergaard, S. D., & Pollak, T. A. (2026). It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. *JMIR Mental Health*, *13*, e91454–e91454. https://doi.org/10.2196/91454

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, *13*(6), 522–530. https://doi.org/10.1016/S2215-0366(25)00396-7

Østergaard, S. D. (2025). Generative Artificial Intelligence Chatbots and Delusions: From Guesswork to Emerging Cases. *Acta Psychiatrica Scandinavica*, *152*(4), 257–259. https://doi.org/10.1111/acps.70022

Parella, S., Güell, B., & Contreras, P. (2023). Los matrimonios forzados como una forma de violencia de género desde un enfoque interseccional. *Revista CIDOB d'Afers Internacionals*, (133), 137–159. https://doi.org/10.24241/rcai.2023.133.1.137

Presser, L. (2016). Criminology and the narrative turn. *Crime, Media, Culture: An International Journal*, *12*(2), 137–151. https://doi.org/10.1177/1741659015626203

Sandberg, S. (2022). Narrative Analysis in Criminology. *Journal of Criminal Justice Education*, *33*(2), 212–229. https://doi.org/10.1080/10511253.2022.2027479

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Ullrich, S., Keers, R., Shaw, J., Doyle, M., & Coid, J. W. (2018). Acting on delusions: the role of negative affect in the pathway towards serious violence. *The Journal of Forensic Psychiatry & Psychology*, *29*(5), 691–704. https://doi.org/10.1080/14789949.2018.1434227

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Witt, K., van Dorn, R., & Fazel, S. (2013). Risk Factors for Violence in Psychosis: Systematic Review and Meta-Regression Analysis of 110 Studies. *PLoS ONE*, *8*(2), e55942. https://doi.org/10.1371/journal.pone.0055942