SEGURIDAD Y ÉTICA

AI-powered suicide risk detection: capabilities, limits, and ethical fault lines

SentirIA Research Papers · 2026

Infraestructura de detección temprana en salud mental

Este documento es parte de la base científica de SentirIA,
plataforma de detección temprana y monitoreo continuo de deterioro en salud mental.
No constituye diagnóstico clínico. La evaluación es responsabilidad del profesional.

AI-powered suicide risk detection: capabilities, limits, and ethical fault lines

Automated suicide detection has advanced rapidly in technical sophistication, yet a fundamental mathematical constraint — the extreme rarity of suicide itself — means even the best models produce overwhelming false positives when deployed at scale. After 50 years and 365 studies, Franklin et al. (2017) found prediction accuracy barely exceeds chance. Transformer-based NLP models now achieve AUCs above 0.91 on research benchmarks, and systems like Meta's proactive detection and Crisis Text Line's triage algorithm process millions of messages daily. But Belsher et al. (2019) demonstrated that the positive predictive value (PPV) of the best models remains at or below 1% for suicide death — meaning over 99% of flagged individuals will not die by suicide. This tension between technical promise and epidemiological reality defines the field's central challenge, with profound implications for the millions of people whose digital communications are now subject to algorithmic surveillance.

---

1. How language reveals the suicidal mind

Two decades of research have identified a constellation of linguistic markers that distinguish suicidal individuals from controls, though differentiating acute crisis from chronic ideation remains an unsolved problem.

Absolutist language — words like "always," "never," "completely," "nothing" — represents one of the strongest linguistic signals. Al-Mosaiwi and Johnstone (2018) analyzed over 6,400 internet forum members and found absolutist words were elevated in suicidal ideation forums compared to controls (effect sizes $d > 3.14$) and, critically, elevated even relative to anxiety and depression forums ($d > 1.71$). This cognitive marker tracked severity more faithfully than negative emotion words, suggesting that *how* people think matters more than *what* they feel. Even recovery forums showed elevated absolutist language above control levels, suggesting persistent cognitive vulnerability.

First-person singular pronoun use ("I," "me," "my") is consistently elevated in suicidal language across text types. Stirman and Pennebaker (2001) documented this in suicidal poets' work, finding increased self-referential language and decreased collective pronouns ("we," "us") over time. Recent clinical transcript analysis (2024) revealed that objective first-person pronouns ("me" as object) were the single most influential predictor of suicidal ideation in machine learning models, outperforming subjective forms. This pronoun shift reflects social disconnection — supporting Durkheim's social integration theory and Joiner's interpersonal theory, where perceived burdensomeness ("my death is worth more than my life") and thwarted belongingness drive suicidal desire. Chu et al.'s 2017 meta-analysis of 66 studies confirmed perceived burdensomeness as the most robustly supported interpersonal predictor of ideation.

Temporal narrowing — the collapse of future orientation — may be the most clinically important marker for distinguishing acute from chronic risk. O'Connor et al. (2008, 2015) demonstrated that it is the *absence of positive future thinking*, rather than the presence of negative future thinking, that predicts suicidal behavior. Reduced future-tense verb usage and loss of references to goals, plans, and future events in text may signal the dangerous transition from chronic ideation to acute crisis. Beck's hopelessness construct mediates the depression-suicide relationship; when hopelessness is statistically controlled, depression ceases to predict suicide. Shneidman (1996) described this as "constriction" — a transient narrowing of affect and intellect into either/or thinking that characterizes the suicidal state.

A striking finding from Glenn et al. (2020) challenges intuitive assumptions: analyzing 189,478 text messages from 33 suicide attempters, explicit death-related words did not significantly differ between suicide attempt episodes and lower-risk periods. Instead, anger increased and positive emotion decreased as individuals approached an attempt. This underscores the need for models that detect subtle emotional trajectories rather than relying on explicit crisis language.

Digital behavioral markers add a temporal dimension

Beyond text content, digital phenotyping captures behavioral patterns that may signal escalating risk. De Choudhury et al. (2016) developed methodology using Reddit data from 440 users who transitioned to r/SuicideWatch, identifying shifts characterized by reduced social engagement, heightened self-attentional focus, and increased hopelessness, impulsiveness, and loneliness. Coppersmith's work at Qntfy demonstrated that effective predictive signals from social media are often clinically unintuitive — Crisis Text Line found "ibuprofen" was among the words most correlated with imminent risk, far more predictive than "suicide."

Nock et al. (2018) pioneered smartphone-based digital phenotyping of suicidal thoughts, identifying five distinct phenotypes based on frequency, intensity, and variability of ideation. More severe and persistent profiles (higher mean, lower variability) were most likely to lead to suicide attempts. A 2026 JMIR study applied vision-language models to passively captured smartphone screenshots, predicting momentary suicidal ideation with AUCs up to 0.83 — the first approach to analyze the semantic content of digital experience rather than behavioral metadata alone. Circadian patterns in online activity (login times, posting hours) represent another emerging signal; Coppersmith noted timing data provides "really interesting signals for proximal suicide risk." Anxiety-related crisis contacts peak at 11 PM, while self-harm contacts peak at 4 AM.

Despite these advances, Kleiman and Nock (2018) revealed that suicidal thoughts develop and subside rapidly, varying substantially within a single day, underscoring the need for temporally granular assessment. Most existing research classifies "suicidal vs. not suicidal" rather than distinguishing acute risk from chronic ideation — the transition that matters most clinically remains the least studied linguistically.

2. Inside the detection systems already watching us

Several major platforms and organizations have deployed AI-assisted suicide detection systems, though all maintain human-in-the-loop architectures. Their technical approaches range from simple keyword matching to deep learning ensembles.

Crisis Text Line's evolving triage algorithm

Crisis Text Line (CTL) has processed over 4 million text conversations and 129+ million messages since 2013. Its AI triage system classifies incoming texts into three risk tiers — high (suicidal thoughts + plan + access to means), medium (suicidal thoughts or self-harm), and normal — moving high-risk texters to the front of the counselor queue within 25 seconds on average. Approximately 8% of engaged conversations are classified as high risk.

CTL's technical architecture evolved through three generations. An initial rules-based "Code Orange" system gave way to a Pointwise Mutual Information (PMI) model using N-gram features, developed because standard NLP tools performed poorly on informal text-message language. The current system uses two binary classification models for suicidal risk and ongoing self-harm, achieving a reported recall of 0.89. Training data comprises 2.8 million annotated conversations from January 2016 through August 2020, with ground truth from post-conversation counselor surveys. The system combines language features with metadata like time of day and conversational context — finding, for instance, that the crying-face emoji is 4× more likely to appear in crisis conversations than the word "suicide," and that "numbs" paired with "sleeve" matches cutting behavior with 99% confidence.

CTL's 2022 data controversy exposed significant ethical tensions. A for-profit spinoff, Loris.ai (in which CTL held a 53% interest), received anonymized crisis conversation data to train commercial customer-service software. After Politico's exposé, CTL ended the arrangement and requested data deletion. Former board chair danah boyd acknowledged that a Terms of Service is not meaningful consent for people in crisis — a point with broad implications for all systems processing crisis communications.

Meta's proactive scanning architecture

Facebook deployed proactive suicide detection globally in November 2017 (excluding the EU, where GDPR prohibits it). The system scans all posts, comments, and Live streams without requiring user reports. Its architecture represents a sophisticated ensemble approach: N-gram-based linear regression classifiers for post and comment text are combined with DeepText neural network classifiers trained on word-meaning vector spaces, all fed into a random forest meta-classifier. Text features account for approximately 80% of model importance, with contextual features (reaction types, time of day, post location) contributing the remaining 20%.

A crucial innovation was Meta's approach to negative training examples. Rather than using random non-suicidal posts, engineers trained on posts that were *reported* as potentially suicidal but *cleared* by human reviewers — such as "I have so much homework I want to kill myself." This harder negative set dramatically improved the classifier's ability to distinguish genuine distress from figurative language. Friend and community response patterns (comments like "Are you OK?" and sad reactions) serve as additional strong signals.

In its first year, the system generated approximately 3,500 wellness check referrals — roughly 10 per day. Meta reported flagging 20× more cases and reaching twice as many users with support materials compared to its previous reporting-only system. However, Meta has refused to publish accuracy data, including false positive rates or outcomes of wellness checks. Health law professor Mason Marks characterized it as a "black box of algorithms" with the power to trigger police visits without accountability.

VA's REACH VET and other deployed systems

The VA's Recovery Engagement and Coordination for Health–Veterans Enhanced Treatment (REACH VET) uses a Lasso regression model with 61 variables extracted from electronic health records, developed by Ronald Kessler at Harvard Medical School. Running monthly batch processing on all patients with VHA contact in the prior 24 months, it identifies those in the top 0.1% risk tier (~6,500–6,800 patients per month). A 2021 JAMA study demonstrated REACH VET was associated with greater treatment engagement, new safety plan documentation, and a 5% reduction in documented suicide attempts — making it one of the few systems with published outcomes data.

Other deployed systems include Koko (1,300+ keywords plus LLMs for distress identification, integrated with Discord, Tumblr, and TikTok), Woebot (regex-based pattern matching for crisis detection within a CBT chatbot framework, now discontinued as of mid-2025), and school monitoring tools like Gaggle and GoGuardian that scan student communications on school-issued devices. Gaggle's performance data from one school district revealed that approximately two-thirds of alerts were false positives, while RAND (2023) found no independent research comprehensively showing these tools measurably lower suicide rates.

---

3. Keyword matching versus contextual understanding

The evidence consistently demonstrates that contextual NLP approaches outperform keyword-based methods, but the magnitude of improvement depends heavily on task complexity, and both approaches face fundamental limitations when deployed at population scale.

Where keywords fall short

LIWC (Linguistic Inquiry and Word Count), the dominant dictionary-based approach, operates as a

bag-of-words system — counting words in psychologically meaningful categories without considering word order, context, or sentence-level semantics. In clinical interview analysis (JMIR Med Inform, 2023; n=305), LIWC features fed into random forest models achieved AUC 0.76–0.89 for detecting high suicide risk, with sensitivity of 0.69–0.85 and specificity of 0.73–0.84. These represent respectable performance in structured clinical settings.

However, LIWC's context-blindness creates systematic errors. Pennebaker himself acknowledged that "programs such as LIWC ignore context, irony, sarcasm, and idioms" — "Not happy" scores as positive emotion, "I'm mad about him" codes as anger. A 2025 reliability study found LIWC precision fell to as low as 49.6% for some categories on online forum data. O'Dea et al. (2015) found that only 14% of keyword-matched suicide-related tweets were "strongly concerning," with 29% "safe to ignore" entirely. Standard dictionaries capture only ~66% of domain-specific vocabulary, and they cannot adapt to rapidly evolving online language (e.g., "unalived," "sewer-slide," "s3lf h@rm").

The transformer advantage

Transformer-based models represent the current state of the art, with dramatic performance improvements on benchmark tasks. RoBERTa-CNN achieved 98.0% accuracy on the Reddit SuicideWatch dataset. Fine-tuned BERT models achieved F1 = 0.93 for binary classification of harmful versus protective suicide-related content on Twitter. The ensemble SI-BERT approach reached ROC-AUC of 0.91. These models excel precisely where keyword approaches fail: handling sarcasm, metaphor, indirect references, and context-dependent meaning. A 2025 study found that adding broader conversational context improved fine-grained classification (self vs. other vs. hyperbole) from 89% to 91% accuracy.

The CLPsych shared tasks provide the most controlled comparisons. In the 2019 task (Reddit suicide risk prediction), a BERT model processing all user posts performed best across all subtasks, though the gap narrowed for simpler binary flagging tasks where even logistic regression approached 97% accuracy. The critical finding: the contextual advantage is most pronounced for fine-grained risk stratification and implicit or indirect language. For simple binary detection with explicit language, traditional methods remain competitive.

Meta-analytic evidence on overall performance

Kusuma et al.'s 2022 meta-analysis of 56 studies (54 ML models) found a pooled AUC of 0.86, sensitivity of 0.66, and specificity of 0.87 for machine learning suicide prediction across all approaches. A 2025 JMIR meta-analysis of 42 studies on adolescent suicide prediction (104 ML models, 1.4 million participants) found combined AUC around 0.82. Critically, meta-regressions in Kusuma et al. did not show significant performance variation by model type, suggesting that data quality and task formulation matter more than algorithmic sophistication. The performance gap between approaches widens mainly on harder, more nuanced classification tasks.

Zero-shot large language models show promise but do not yet surpass fine-tuned smaller transformers. GPT-3.5 zero-shot achieved accuracy of 88% but F1 of only 73% on Reddit suicide

risk assessment, compared to fine-tuned ALBERT's F1 of 86.9%. A 2025 BMJ Mental Health study found that reasoning models running on consumer hardware can approximate the performance of larger models for suicide risk stratification from clinical notes. The emerging consensus favors hybrid architectures — using LIWC features alongside transformer embeddings, or two-stage systems with efficient BERT triage followed by LLM analysis for ambiguous cases — to balance accuracy, interpretability, and computational efficiency.

---

4. The base rate wall that no algorithm can breach

The most fundamental challenge in suicide detection is not algorithmic but epidemiological. The annual suicide rate of approximately 13–14 per 100,000 (0.014%) in the general U.S. population creates a mathematical constraint that no amount of model improvement can overcome at the individual prediction level.

The arithmetic of rare events

Consider a hypothetical test with 99% sensitivity and 99% specificity — far better than any existing tool — applied to 1 million people at the general population base rate. It would correctly identify ~139 of the 140 people who will die by suicide (true positives), but would also incorrectly flag ~9,999 people who will not (false positives). The resulting PPV is just 1.4% — meaning 98.6% of positive flags are false alarms. With the more realistic performance of existing tools (90% sensitivity, 90% specificity) applied to a clinical population with 1% prevalence, PPV reaches only 8.3%, meaning over 90% of high-risk flags remain false positives.

Belsher et al. (2019), in a systematic review of 17 cohort studies and 64 prediction models across 14+ million participants, found that while global classification accuracy was good (AUC ≥ 0.80 in most models), predictive validity for a positive result for suicide mortality was extremely low — PPV ≤ 0.01 in most models. Their conclusion was stark: "To date, suicide prediction models produce accurate overall classification models, but their accuracy of predicting a future event is near 0."

Clinical risk tools fare no better

Large et al.'s 2016 meta-analysis of 53 samples from 37 studies of risk-assessed psychiatric patients found pooled sensitivity of 56%, specificity of 79%, and an odds ratio of 4.84 for suicide among high-risk versus lower-risk patients. The clinical implications are sobering: approximately 95% of patients classified as "high-risk" did not die by suicide, while 44% of suicides occurred among patients classified as "lower-risk." A separate inpatient-specific meta-analysis found PPV of risk categorization at just 0.43%. Runeson et al. (2017) reviewed 15 prediction instruments and found none achieved even the modest benchmark of 80% sensitivity with 50% specificity. Chan et al. (2016) found PPV of clinical risk scales ranged from 1.3% to 16.7%, concluding: "The use of these scales, or an over-reliance on the identification of risk factors in clinical practice,

may provide false reassurance and is, therefore, potentially dangerous."

The C-SSRS, widely considered the gold standard for suicide assessment, illustrates the distinction between assessment and prediction. It achieves excellent concurrent validity (100% sensitivity, 96–100% specificity for classifying past behavior) but much weaker prospective prediction: sensitivity of 53.9% and specificity of 75.6% for suicide within 31 days in emergency department settings. The PHQ-9 Item 9 achieves sensitivity of 74.7% against the C-SSRS, but with a PPV of only 27.5% — three-quarters of positive screens are false positives.

Machine learning has not solved the problem

Walsh et al. (2017) achieved an impressive AUC of 0.84 for predicting suicide attempts using EHR data, with accuracy improving closer to the event. But this was a case-control design (suicide attempts vs. non-suicidal self-injury), not population-level prediction — when applied to broader populations, PPV drops dramatically due to base rates. The Army STARRS program, one of the largest military suicide prediction efforts, achieved AUC of 0.78 for suicide attempts among transitioning service members; the top 20% predicted risk captured 60.9% of attempts. Even at this concentration of risk, the absolute numbers of false positives vastly exceed true positives at population scale.

An intriguing counterpoint from Haghish and Czajkowski (2023) found that ML "false positives" in a Norwegian adolescent study had significantly higher rates of future first-time suicide attempts compared to true negatives — suggesting some false positives may represent genuine but not-yet-manifest risk. This reframing has implications for how we evaluate prediction systems, but does not resolve the fundamental base rate constraint.

Large articulated the sharpest critique: "Suicide is a sufficiently rare event and our current tools sufficiently blunt that 'high-risk' groups will contain mainly false positives and the majority of those who die by suicide will have been categorised as low risk." He advocated abandoning "misguided attempts at risk prediction" in favor of engagement with individual patients.

---

5. When algorithms trigger welfare checks and other ethical fractures

The ethical landscape of automated suicide detection encompasses privacy, consent, bias, autonomy, and the tangible harms that flow from system errors — particularly in a domain where false positives can trigger coercive interventions.

The consent paradox in crisis detection

A foundational ethical tension pervades all deployed systems: people in crisis cannot meaningfully consent to algorithmic surveillance. CTL's Terms of Service — the consent mechanism for people texting in suicidal distress — runs to 4,000+ words. Meta scans virtually

every post without specific consent for mental health data processing, which is why the system is banned in the EU under GDPR. Barak-Corren et al.'s "Protecting Life While Preserving Liberty" framework argues that patients with decisional capacity should have the right to opt out of AI monitoring, but acknowledges that opt-in systems may miss at-risk individuals who fail to consent. The Hastings Center for Bioethics has argued that healthcare institutions have a duty to disclose when patient records are subjected to automated suicide risk screening.

False positives carry real-world consequences

The consequences of false positives extend far beyond statistical inconvenience. AI-triggered welfare checks have resulted in documented fatalities: Atatiana Jefferson (2019), a 28-year-old Black woman shot by police during a wellness check; John Albers (2018), a 17-year-old killed by police responding to social media concerns about self-harm; and Trevor Mullinax (2021), shot at 50 times during a suicide welfare check. A study published in Psychiatric Services found a "lack of any real medical literature" on the clinical effectiveness of police welfare checks for suicide risk management. Additional harms include involuntary psychiatric hospitalization, stigmatization, employment consequences, loss of firearms rights, damaged therapeutic relationships, and clinician "alarm fatigue" — reflexive click-through of warnings within the first 30 days of deployment.

Algorithmic bias compounds existing disparities

AI suicide detection systems inherit and amplify existing healthcare biases. An APA-cited study by Yates Coley at Kaiser Permanente found that suicide prediction algorithms performed worse for BIPOC populations due to smaller training datasets and structural factors — Black Americans are approximately 10–12% less likely to be diagnosed with depression despite similar symptom burden. A controlled experiment by Adam et al. in Nature Communications Medicine demonstrated that prescriptive recommendations from biased AI systems created racial and religious disparities in emergency mental health decisions: both clinicians and non-experts chose police intervention more often for racial minorities. However, descriptive (rather than prescriptive) framing of AI recommendations mitigated this bias — a finding with direct design implications.

The chilling effect may undermine prevention

Perhaps the most paradoxical harm is that surveillance-based detection may deter the very help-seeking behavior it aims to protect. Edwards (2013) found that states with duty-to-warn laws experienced an increase in teen suicides of approximately 9%, suggesting mandatory reporting requirements deter disclosure. Crisis line workers report that approximately 75% of high-risk texters terminate conversations when asked for identifying information during escalation protocols. The awareness that AI systems monitor communications for suicide risk could discourage vulnerable individuals from expressing distress online — precisely the behavior these systems depend on detecting.

Regulatory frameworks remain underdeveloped

The FDA has not yet authorized any generative AI-based mental health device. Its Digital Health Advisory Committee, which held focused sessions on AI mental health devices in 2024 and 2025, emphasizes human oversight requirements and equitable performance across populations. The EU AI Act (effective August 2024) classifies mental health AI as likely high-risk, requiring fundamental rights impact assessments, data governance standards, and human oversight provisions, with penalties up to €35 million or 7% of annual turnover. The Canada Protocol provides the most concrete ethical checklist: 38 items across five categories (system description, privacy/transparency, security, health-related risks, and biases) designed specifically for AI in suicide prevention.

The Tarasoff duty-to-protect framework, which requires mental health professionals to protect potential victims when aware of serious threats, does not straightforwardly apply to digital platforms — social media companies are not licensed clinicians, and no "therapeutic relationship" exists between a platform and its users. Whether algorithmic prediction constitutes "foreseeable risk" that triggers Tarasoff-like obligations remains legally unresolved across jurisdictions. Twenty-nine U.S. states have mandatory duty-to-warn provisions, 17 have permissive duty, and 4 have not recognized such a duty at all.

Conclusion: navigating between mathematical limits and moral imperatives

The field of automated suicide detection sits at a crossroads where genuine technical progress collides with irreducible epidemiological constraints and unresolved ethical questions. Three insights emerge from synthesizing the evidence.

First, the transition from "who is at risk" to "when someone is at risk" represents the field's most important frontier. Within-person temporal models — tracking linguistic and behavioral trajectories over time — hold more promise than cross-sectional classification. The shift from absolutist language and pronoun changes to temporal narrowing and loss of future orientation may mark the critical acute-risk window. Digital phenotyping approaches that capture circadian shifts, social withdrawal, and emotional trajectory changes in real time are beginning to operationalize this distinction, though most remain at pilot scale.

Second, the base rate problem reframes the fundamental question from "can we predict?" to "what should we do with predictions?" The same algorithm may have clinical utility when paired with low-burden interventions (offering resources, prompting safety planning) but cause net harm when triggering coercive responses (involuntary holds, police welfare checks). The asymmetric cost framework — traditionally assuming false negatives are always worse than false positives — requires reassessment when false positives can result in stigma, trauma, or death during wellness checks.

Third, the absence of published accuracy data from the largest deployed systems — particularly Meta's — represents a critical accountability gap. Systems that process billions of communications and trigger thousands of real-world interventions annually operate without independent validation, peer review, or outcome tracking. The VA's REACH VET stands as a notable exception, demonstrating that transparent, ethically governed deployment with published outcomes data is feasible. The field needs standardized reporting requirements, independent audits, and mandatory disaggregated performance data across demographic groups. Until these exist, the gap between technical capability demonstrated in research and responsible real-world deployment will remain the defining challenge of AI-assisted suicide prevention.