future tense

Rapid Risk Assessment Using AI-Enhanced Due Diligence: *Trust and Safety Risk Identification via Public Signals and Human Judgment*

Author: Liana H. Meyer
Affiliation: Independent Researcher, Future Tense
Date: January 2026
Version: 1.0 Preprint
ORCID: 0009-0002-4587-8039
DOI: 10.6084/m9.figshare.31160749

About This Paper: A rapid trust-and-safety risk assessment demonstrating how AI-assisted open-source research, combined with human judgment, can support ethical and proportionate disengagement when formal vetting is unavailable.

Preprint Status: This manuscript has not yet been peer reviewed.

Rapid Risk Assessment Using AI-Enhanced Due Diligence:
Trust and Safety Risk Identification via Public Signals and Human Judgment

Liana H. Meyer

Abstract

This case study examines the use of AI-assisted due diligence to conduct a rapid trust-and-safety risk assessment in the absence of formal vetting mechanisms. The assessment was triggered by unsolicited professional outreach via a public networking platform and conducted under time pressure using public information. Open-source records included material relevant to safety considerations, along with documentation referencing two established California risk assessment instruments indicating placement in the highest risk category within those frameworks.

AI tools were used to accelerate information gathering and synthesis, (1) while all interpretation and decision-making remained human-led. The process combined language and communication analysis, open-source background review, credibility and consistency checks, and structured risk framing. Formal risk classifications were interpreted cautiously alongside other signals.

The outcome was a precautionary decision to disengage, selected as a proportionate harm-prevention measure under uncertainty. This case illustrates how AI-enhanced due diligence—when bounded by ethical safeguards, human accountability, and proportional reasoning—can support responsible decision-making in time-sensitive trust-and-safety situations. As professional interactions increasingly occur in informal digital spaces, such approaches represent an emerging component of practical governance.

As a qualitative practice case study, this paper contributes applied insight into AI-assisted risk governance in informal digital interaction contexts

Keywords: AI governance, trust and safety, risk assessment, due diligence, recidivism risk, human-in-the-loop, public-source intelligence

Context

In a professional role involving external partnerships within a mission-driven organizational setting, unsolicited outreach from unfamiliar contacts was a routine occurrence. During a period of increased reliance on online professional networking platforms, one message framed as a collaboration opportunity prompted caution due to its urgency, tone, and lack of prior relationship or institutional context.

Several observable signals in the individual's communication style and profile raised questions about credibility, intent, and potential reputational exposure. The situation unfolded in late 2025 over several days, requiring a rapid credibility assessment without the benefit of formal vetting processes. These conditions led to a preliminary review of publicly available information before considering any further engagement.

Problem Definition

The challenge was to determine whether an unsolicited contact posed safety or reputational risk using publicly available information, under time pressure, and without formal vetting tools—while ensuring the assessment avoided speculation or defamation.

Method & Judgment Applied

A structured assessment combined AI-assisted research with human judgment. The approach was intentionally defensive, prioritizing the identification and interpretation of potential risk signals.

Language and Framing Review: The tone, language, and self-presentation in the initial outreach were analyzed for anomalous or concerning patterns. Indicators included undue familiarity for a first contact, boundary-testing requests, and inconsistencies in stated identity or purpose.
Open-Source Background Review: Publicly available records and digital footprint information were examined. AI tools were used to accelerate search, aggregation, and summarization, enabling efficient review of relevant documentation while preserving human oversight of interpretation.
Credibility and Consistency Assessment: Information from the outreach was cross-checked against public records for omissions, contradictions, or credibility gaps. Professional markers (e.g., coherence of online presence) were considered in context rather than as standalone proof of legitimacy.
Structured Risk Framing: Potential best- and worst-case scenarios were mapped, with decision weight placed on safety and reputational exposure rather than potential opportunity. The absence of formal vetting increased reliance on precaution. (2)
Synthesis and Judgment: Signals from these steps were synthesized into an overall qualitative risk judgment. Public records included documented results from established California risk assessment frameworks (LS/CMI and Static-99R) indicating placement in the highest recidivism-risk category. The underlying criteria referenced in those assessments were reviewed to understand category meaning. These classifications were interpreted cautiously and in context. (3) AI supported information gathering and synthesis, while interpretation and judgment remained human-directed.

Ethics & Safeguards

Ethical safeguards were central to this assessment due to the sensitivity of personal risk information.

Privacy and Data Boundaries: The review was limited strictly to publicly available sources, including public records, reporting, and open social media. (4) No confidential, private, or restricted data were accessed.
Human Oversight and Accountability: AI tools supported search and summarization; conclusions and risk judgments remained human-led.

Version 1.0 • January 2026 • DOI: 10.6084/m9.figshare.31160749

Governed Use of AI in Due Diligence: AI-assisted research can accelerate preliminary screening, but must operate within defined ethical and procedural boundaries under human oversight.
Precaution as a Legitimate Risk Control: When credible but incomplete signals indicate potential harm, disengagement may be the most proportionate control measure.
Why This Is Increasingly Relevant: Professional interaction is increasingly decentralized, digital, and cross-jurisdictional. As a result, more risk decisions occur in informal spaces without clear governance scaffolding.

Outcomes & Findings

The rapid assessment resulted in a precautionary decision to disengage, based on multiple converging risk signals.
Public record indicators: Open-source records included information relevant to safety considerations.
Formal risk classifications: Publicly available documentation referenced two established assessment instruments (LS/CMI and Static-99R) indicating placement in the highest risk category within those frameworks. These were interpreted cautiously and in context.
Credibility and boundary concerns: Communication analysis identified boundary and credibility concerns. Taken together, these signals suggested a level of potential safety and reputational risk inconsistent with continued engagement. The proportionate response was to discontinue contact and avoid further involvement. Disengagement served as a precautionary risk-control measure by limiting exposure. No further interaction occurred following the decision.

Governance / Risk Implications

Individual Judgment Within Organizational Risk Systems: This case illustrates how individual due diligence functions as a frontline element of organizational risk governance. (5) When credible external assessment frameworks signal elevated risk, disregarding those signals would be inconsistent with prudent safety and reputational risk management.
Informal Channels as Structural Governance Gaps: Most formal vetting systems apply to hiring, partnerships, or procurement—not unsolicited or informal professional outreach. Professionals operating on digital platforms often serve as de facto risk gatekeepers without clear policy guidance.
Risk Typology: Reputational, Safety, and Ethical Exposure: The risks in this case were multi-dimensional. Governance frameworks must address operational, reputational, and ethical risk surfaces created by informal engagement.

Governed Use of AI in Due Diligence: AI-assisted open-source research can increase the speed and breadth of preliminary screening, but it must operate within defined procedural and ethical constraints. (6) Human oversight, documented reasoning, and proportional interpretation are necessary to prevent overreach, bias amplification, or false certainty. This case highlights the importance of treating AI as a support tool within a governed decision process rather than as an autonomous evaluator.

Precaution as a Legitimate Risk Control: The case reinforces the role of precaution in trust-and-safety governance. When credible but incomplete signals indicate potential harm, disengagement may be the most proportionate control measure. Effective governance frameworks should support good-faith precautionary decisions that prioritize safety and reputational protection while maintaining fairness and restraint.

Why This Is Increasingly Relevant: Professional interaction is increasingly decentralized, digital, and cross-jurisdictional, while formal vetting structures remain tied to traditional institutional processes. As a result, more risk decisions occur in informal spaces without clear governance scaffolding. Organizations that fail to adapt their risk guidance to these environments leave staff unsupported and institutional exposure unmanaged.

Implications for Practice

AI as a Rapid Screening Support Tool: AI can support rapid preliminary screening when formal vetting is unavailable or too slow.

Human Judgment as the Decisive Layer: Decisions involving safety, ethics, or reputational exposure should remain human-led.

Communication Patterns as Early Risk Signals: Communication behaviors—such as boundary-pushing familiarity, undue urgency, or inconsistencies in self-presentation—can serve as early indicators of potential concern.

Use Structured, Documented Processes: Clear documentation of signals, uncertainties, and rationale supports accountability and defensibility.

Applicability Beyond This Case: These practices extend beyond individual safety screening to broader trust-and-safety, partnership vetting, digital engagement, and reputational risk contexts.

Key Governance Insight

Insight: AI-assisted research can accelerate early risk detection when decisions remain firmly human-led.

Risk/trade-off: Precautionary decisions may close uncertain opportunities, but failing to act on credible signals can create outsized safety and reputational risk.

Practical takeaway: Use AI for synthesis, not judgment, and apply structured risk framing that prioritizes harm prevention and documents decision rationale.

From Case Insight to Organizational Practice

Organizations can strengthen trust and safety governance by:

- Establishing clear screening and escalation guidance for unsolicited or informal professional contacts
- Training and equipping employees to apply these standards consistently
- Clarifying that AI supports research and synthesis, while accountability and decisions remain human-led
- Documenting rationale and aligning decisions with risk appetite

Limitations

The analysis was limited to publicly available information and probabilistic risk indicators, leaving personal context, intent, and potential rehabilitation unknown. As in many rapid due-diligence settings, conclusions relied on proxy signals that may favor precaution. The findings reflect a defensible judgment under constraint rather than a comprehensive or predictive assessment.

Conclusion

This case demonstrates that AI-assisted due diligence can support rapid, safety-oriented decision-making when formal vetting mechanisms are unavailable—provided the process remains governed by human judgment, ethical restraint, and proportional response. By treating public risk indicators as contextual signals rather than certainties and prioritizing harm prevention over potential opportunity, the assessment reached a defensible decision under conditions of uncertainty.

As professional interactions increasingly take place in low-friction digital environments, practitioners and organizations alike need practical, ethically grounded methods for navigating ambiguous risk. This case offers one model: AI as support, humans as accountable decision-makers, and precaution applied with discipline. Sound judgment in these moments is not only a personal safeguard but a growing governance necessity.

Version 1.0 • January 2026 • DOI: 10.6084/m9.figshare.31160749

References

(1) OECD. OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development, 2019.
https://oecd.ai/en/ai-principles
(2) European Commission. Communication on the Precautionary Principle. Commission of the European Communities, 2000.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52000DC0001
(3) Public Safety Canada. Actuarial Risk Assessment Tools and Their Limitations. Government of Canada, research summary.
https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/index-en.aspx
(4) U.S. Department of Homeland Security. Privacy Impact Assessment for Open Source Intelligence Activities. DHS, 2017.
https://www.dhs.gov/publication/privacy-impact-assessment-open-source-intelligence-osint
(5) Institute of Risk Management (IRM). A Risk Practitioner's Guide to ISO 31000: Risk Management. IRM, overview resources.
https://www.theirm.org/knowledge-and-resources/iso-31000-risk-management-standard/
(6) National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce, 2023.