Prova digitale e anti-forensics: cosa cambia con l’AI generativa

L’adozione di sistemi di intelligenza artificiale generativa sta producendo una trasformazione profonda nel modo in cui il diritto processuale penale interpreta, valuta e utilizza la prova digitale. Ciò che per anni è stato considerato un elemento relativamente stabile e verificabile ovvero il dato digitale come traccia dell’origine di un evento o di un’azione entra oggi in una fase di progressiva instabilità epistemica. La possibilità di generare contenuti artificiali altamente coerenti, completi di metadati plausibili e strutture tecniche formalmente corrette, incrina infatti uno dei presupposti storici della digital forensics: la corrispondenza tra traccia e origine.

Questo articolo si inserisce all’interno della serie di approfondimenti a cura di Cosimo de Pinto dal titolo **“Quando l’AI mente bene – Deepfake, prova digitale e anti-forensics nel processo penale”.** L’approfondimento analizza la frattura introdotta dall’AI generativa nel paradigma tradizionale della prova informatica, evidenziando come la “plausibilità sintetica” possa sostituirsi alla tracciabilità dell’origine come criterio implicito di affidabilità.

Nel testo si esamina la discontinuità tra anti-forensics classica e nuove tecniche abilitate dall’intelligenza artificiale, mettendo in luce il passaggio da strategie di occultamento o distruzione della prova a sistemi capaci di costruire evidenze artificiali credibili e difficilmente distinguibili da quelle autentiche. L’analisi si concentra inoltre sulle implicazioni processuali di tale evoluzione, con particolare attenzione ai rischi per la valutazione giudiziale della prova digitale e alle possibili ricadute sulla

presunzione di innocenza e sulla libertà personale.

La fine dell’innocenza digitale

L’AI generativa incrina il presupposto che la prova digitale rifletta sempre la propria origine.

La digital forensics si è sviluppata, nel corso dei suoi tre decenni di storia disciplinare, attorno a un presupposto implicito: i dati digitali conservano tracce della propria origine, e tali tracce possono essere estratte, interpretate e presentate come prova con ragionevole certezza scientifica. Un file aveva una biografia verificabile: nasceva in un momento preciso, recava l’impronta del software che lo aveva generato, i metadati del sistema operativo, le impronte dell’hardware. Su questo fondamento, ben descritto nella manualistica di riferimento, si è costruita l’intera architettura processuale della prova digitale nei sistemi giuridici occidentali.

Dal paradigma classico alla frattura epistemica

	Computer forensics classica	**Scenario AI generativa**
Presupposto di base	Le tracce digitali riflettono l’origine del dato	Un contenuto può simulare un’origine credibile
Obiettivo dell’analisi	Ricostruire eventi e attribuire provenienza	Verificare anche la possibilità di un’origine sintetica

	Computer forensics classica	Scenario AI generativa
Segni di manipolazione	Spesso lasciano anomalie residue	Possono imitare pattern statisticamente plausibili
Ruolo dell’esperto	Interpretare tracce e coerenze tecniche	Corroborare, autenticare e contestualizzare
Rischio giuridico	Errore tecnico circoscritto	Potenziale violazione della libertà personale

L’avvento dei modelli generativi di intelligenza artificiale, in particolare dei Large Language Models (LLM), dei modelli text-to-image, dei sintetizzatori vocali neurali e dei sistemi di video synthesis, ha introdotto nella catena della prova digitale una variabile che i framework forensi tradizionali non erano stati progettati per gestire: la plausibilità sintetica. Un documento, un’immagine, una registrazione audio o video possono oggi essere generati artificialmente in modo da superare non soltanto la percezione umana, ma anche una parte significativa delle analisi strumentali disponibili ai laboratori forensi. Come osservato dalla letteratura tecnica più recente, gli avversari non si limitano più a nascondere le prove: le costruiscono, le avvelenano e orientano l’investigatore verso una narrazione falsa.

Le conseguenze di questa trasformazione non sono confinabili all’ambito tecnico. Nei procedimenti penali, ogni errore nella valutazione di una prova digitale si traduce in una potenziale violazione della libertà personale. Un soggetto innocente può essere condannato sulla base di una conversazione WhatsApp mai avvenuta, di un’immagine fotografica manipolata, di un video che lo mostra in un luogo in cui non si è mai recato, di un log di sistema artatamente costruito per collocarlo in prossimità di un reato. L’AI non mente come un essere umano:

produce output statisticamente coerenti, difficili da smascherare con gli strumenti tradizionali, e capaci di ingannare persino analisti esperti.

> *Ogni errore nella valutazione di una prova digitale si traduce in una potenziale violazione della libertà personale.*

Il presente paper analizza le intersezioni tra AI generativa, tecniche di anti-forensics e il concetto giuridico di prova digitale, con attenzione alle implicazioni pratiche per consulenti tecnici, avvocati e magistrati. Vengono esaminate le principali superfici di attacco, le tecniche di evasione forense abilitate dall'AI, i framework di auditabilità esistenti, le lacune normative che rendono urgente un ripensamento del valore probatorio del dato digitale e, soprattutto, i rischi concreti per la libertà delle persone che derivano da un approccio acritico all'impiego dell'AI in sede processuale.

Metodo, contributo e domanda di ricerca

Metodo. Sintesi critica interdisciplinare che incrocia letteratura peer-reviewed, casi documentati, standard normativi e implicazioni processuali della prova digitale nell'era dell'AI generativa.

Contributo. Un quadro di sintesi tra tecnica di attacco, artefatto colpito, contromisure e rischio residuo; ricostruzione del nesso tra auditabilità tecnica e tenuta probatoria; proposta di principi guida per una digital forensics AI-resistant.

Domanda di ricerca. In che modo l'AI generativa trasforma la natura della prova digitale e quali condizioni metodologiche e giuridiche sono necessarie per conservarne l'affidabilità?

Questo paper affronta una delle sfide più critiche e meno sistematizzate dell'attuale panorama della sicurezza informatica applicata al diritto: l'impatto dell'intelligenza artificiale generativa sull'affidabilità delle prove digitali nei procedimenti penali. Non si tratta di uno scenario futuro. Le tecniche qui analizzate (data poisoning, log crafting, PRNU manipulation, deepfake, voice cloning, evasion attacks) sono operative già oggi, e i loro effetti sul sistema giustizia sono documentabili. Il taglio è deliberatamente duplice, tecnico e giuridico, perché duplice è il pubblico cui il problema si rivolge: professionisti della sicurezza, consulenti tecnici d'ufficio e di parte, magistrati e avvocati che si trovano a maneggiare prove digitali la cui affidabilità non può più essere data per scontata.

Sul piano del metodo, il lavoro è una sintesi critica di taglio interdisciplinare. Sul versante tecnico incrocia la letteratura peer-reviewed su anti-forensics e adversarial machine learning con casi documentati di rilevanza processuale o geopolitica; sul versante giuridico procede a un'analisi dottrinale e giurisprudenziale del diritto della prova, ancorata alla normativa italiana ed europea vigente e agli standard tecnici internazionali della famiglia ISO/IEC 27000. Le fonti sono state selezionate privilegiando, ove disponibili, gli atti originali (paper, whitepaper, testi normativi, massime ufficiali) rispetto alle fonti secondarie.

Per onestà epistemica, ogni affermazione tecnica viene qualificata secondo tre gradi di evidenza, esplicitati nel testo: risultato dimostrato sperimentalmente in letteratura; condotta documentata in un caso reale; rischio metodologico emergente, ipotizzato in via teorica ma non ancora consolidato in giudicato.

Il contributo originale è triplice: un quadro di sintesi che mette in relazione ciascuna tecnica di attacco con l'artefatto forense colpito, la contromisura di rilevazione e il rischio residuo (sezione 5); un raffronto comparato tra l'ordinamento

italiano e il modello angloamericano (sezione 7); e un framework forense resistente all'AI articolato in cinque principi operativi e quattro proposte normative (sezione 10). La domanda di ricerca che tiene insieme l'intero discorso è la seguente: a quali condizioni una prova digitale generata o mediata da sistemi AI può fondare un giudizio di responsabilità penale senza violare la presunzione di innocenza e il diritto a un equo processo?

Anatomia dell'anti-forensics tradizionale e la discontinuità dell'AI

Dalle tecniche che cancellano le tracce a un'AI che le costruisce: una rottura qualitativa, non solo quantitativa.

L'anti-forensics prima dell'AI

La disciplina dell'anti-forensics, intesa come l'insieme delle tecniche volte a ostacolare, fuorviare o rendere impossibile l'analisi forense di un sistema digitale, è quasi coeva alla nascita della computer forensics stessa. Harris la definisce come qualsiasi tentativo di compromettere la disponibilità o l'utilità delle prove in un processo forense.

Le tecniche classiche si articolano in quattro categorie principali:

- distruzione dei dati (data destruction) tramite sovrascrittura multipla secondo lo standard DoD 5220.22-M, degaussing o distruzione fisica del supporto;
- occultamento dei dati (data hiding) mediante steganografia, cifratura di volumi interi (VeraCrypt, BitLocker) e slack space exploitation;
- trail obfuscation mediante applicazioni portable,

timestamp manipulation (Timestomp) e cancellazione selettiva dei log di sistema;
- attacchi alla capacità di analisi attraverso exploit nei tool forensi (FTK, EnCase, Autopsy), file artatamente costruiti (crafted) e tecniche anti-VM e anti-sandbox.

La caratteristica comune di queste tecniche è che esse lasciano tracce della propria applicazione: assenza dove dovrebbe esserci presenza, anomalie statistiche, pattern inusuali. In questa prospettiva, la forensics tradizionale lavora spesso sulla deviazione rispetto all'atteso: l'anti-forensics classica mirava a cancellare, occultare o rendere non interpretabile l'evidenza. Nell'ecosistema AI la logica cambia radicalmente.

La discontinuità dell'AI: quando l'assenza non fa rumore

L'AI generativa introduce una rottura epistemica nel paradigma descritto. Il problema non è produrre contenuti falsi: è produrre contenuti coerenti con le aspettative statistiche di un analista forense. Un LLM avanzato istruito a generare la trascrizione di una conversazione WhatsApp produrrà testo con le peculiarità linguistiche del presunto autore, con riferimenti contestualmente plausibili e struttura temporale coerente; se integrato con un generatore di metadati, produrrà file con timestamp consistenti, encoding corretto, struttura interna del database SQLite di WhatsApp formalmente valida.

La differenza rispetto all'anti-forensics tradizionale è qualitativa, non soltanto quantitativa: non si tratta necessariamente di un'anti-forensics più efficiente, ma di una crisi del concetto stesso di autenticità digitale. Il bersaglio reale diventa ciò che il modello vede: dati, log, metadati, timestamp, immagini, testo, prompt, contesto RAG, workflow, plug-in, API e catena software. L'AI non mente in senso umano; produce un output coerente rispetto all'input e

al contesto che riceve. Se input e contesto sono manipolati, la conclusione può essere tecnicamente elegante e probatoriamente falsa.

> *L’AI non mente in senso umano; produce un output coerente rispetto all’input e al contesto che riceve.*

L’approfondimento ha mostrato come l’AI generativa stia alterando in modo strutturale il concetto stesso di prova digitale, spostando l’attenzione dalla tracciabilità dell’origine alla plausibilità sintetica del contenuto. Questo passaggio segna una discontinuità profonda per la digital forensics e apre scenari in cui la distinzione tra dato autentico e dato costruito diventa sempre più complessa da stabilire, con conseguenze dirette sul piano processuale e sulla tutela della libertà personale.

Il prossimo articolo approfondirà le principali tecniche anti-forensics assistite dall’AI insieme alle nuove superfici di attacco lungo la catena della prova digitale, dai documenti alle comunicazioni fino al codice generato automaticamente.

Per un inquadramento più completo e sistematico, è possibile approfondire il paper di **Cosimo de Pinto “Quando l’AI mente bene – Deepfake, prova digitale e anti-forensics nel processo penale”**: la lettura consente di comprendere in modo organico le dinamiche tra attacco, prova digitale e rischio probatorio, offrendo un riferimento utile sia per l’analisi tecnica sia per l’applicazione nel contesto processuale.

https://www.ictsecuritymagazine.com/articoli/prova-digitale/