FONDAZIONE BRUNO KESSLER
MAGAZINE

La sfida per l’integrità dei contenuti digitali nell’era dei deepfake e della GenAI

11 Marzo 2026

Perché identità digitale, crittografia e standard di provenienza sono strumenti chiave per garantire autenticità, tracciabilità e fiducia

Fino a un paio di anni fa era spesso ancora possibile intuire che un’immagine o un audio fossero **manipolati**: un dettaglio visivo fuori posto, un’imperfezione nella voce. Oggi, con l’evoluzione degli strumenti di **GenerativeAI**, l’aspetto percettivo è sempre più convincente e la rilevazione di falsi sempre più complessa. È anche su questo fronte che lavora il Centro Cybersecurity di FBK, che studia come garantire autenticità e tracciabilità dei contenuti digitali in uno scenario in rapida trasformazione.

Accanto alle applicazioni positive, questi strumenti rendono possibile la creazione di contenuti falsi, i cosiddetti deepfake. Le conseguenze sono evidenti e possono colpire singole persone o aziende: **truffe** e **scam** sempre più sofisticati, casi di videochiamate **manipolate** che hanno portato dirigenti a trasferire ingenti somme di denaro, processi di onboarding aziendale da remoto completamente falsificati, fino alla diffusione di materiale diffamatorio o pornografico. Mentre queste tecnologie diventano sempre più accessibili e di supporto in diverse applicazioni, si sta sviluppando anche un mercato parallelo di servizi pensati per facilitare attività di cybercrime.

In questo scenario, l’obiettivo di sicurezza non è solo garantire la **confidenzialità dei dati**, ma anche la loro **integrità**, intesa in senso ampio: quando un’immagine viene alterata rispetto al suo stato originale, è violata la sua integrità tecnica; ma esiste anche un livello più sottile, quello dell’ **integrità semantica** dell’informazione trasmessa dal contenuto, che dipende anche dal tipo di trasformazioni che un’immagine od un video hanno subito e dal contesto in cui sono inseriti. Per esempio, ridimensionamento o filtraggio leggero di un’immagine possono essere considerate modifiche ad impatto limitato rispetto all’aggiunta/rimozione di oggetti o la combinazione di più dati visivi diversi, con questi ultime trasformazioni la semantica del contenuto può essere manipolata e radicalmente alterata. Per questo verificare in modo trasparente l’origine e le trasformazioni di un’informazione visiva, sia essa un’immagine od un video, per poter valutare la loro affidabilità è oggi più che mai fondamentale: se non possiamo stabilire con garanzie ragionevoli se e come siano stati manipolati, diventa in generale difficile mantenere fiducia nei contenuti digitali.

Da anni la ricerca sviluppa tecniche per individuare contenuti manipolati analizzandone le caratteristiche statistiche. Oggi però si lavora anche tramite un **approccio complementare**: non limitarsi a riconoscere i falsi a posteriori, ma costruire un sistema di **trust fin dalla creazione** del contenuto. *"L'idea è associare ai contenuti dei metadati crittograficamente verificabili, legati in modo univoco all'immagine o al video",* afferma il direttore del Centro Cybersecurity **Silvio Ranise**. *"Questi metadati vengono firmati attraverso identità digitali certificate, così da poter individuare non solo chi ha creato l'immagine od il video ma anche chi ne ha modificato il contenuto effettuando quali operazioni".*

In questa direzione si muove anche la specifica **C2PA** (Coalition for Content Provenance and Authenticity), che punta a creare ecosistemi in cui i contenuti siano tracciati e firmati idealmente dalla loro acquisizione fino al consumo finale. La sfida è soprattutto l'adozione, perché questi sistemi diventino parte integrante delle piattaforme.

Cecilia Pasquini, ricercatrice del Centro Cybersecurity di FBK, è coinvolta in due working group europei impegnati nell'elaborazione di un code of practice per rendere operativi gli obblighi di trasparenza previsti dall'articolo 50 dell'AI Act. *"Stiamo lavorando per definire approcci tecnologici appropriati per fare marking e labelling dei contenuti generati dall'AI. L'obiettivo è individuare strumenti tecnici efficaci ed interoperabili, dai metadati firmati al watermarking, che rendano possibile una reale tracciabilità dei contenuti digitali"*, spiega Cecilia Pasquini.

Il tema si inserisce in un quadro normativo in evoluzione. Anche a livello nazionale si moltiplicano le iniziative, con il riconoscimento del deepfake come reato e un crescente collegamento tra **AI compliance** e **cybersecurity**. In Italia, infatti, la responsabilità per la vigilanza sul mercato dell'AI e la compliance con l'AI Act è stata affidata all'Agenzia per la Cybersicurezza Nazionale (ACN). Questa decisione posiziona l'ACN al centro della strategia nazionale di sicurezza e innovazione digitale, con implicazioni dirette nella gestione dei deepfake. Essendo l'**AI Act** un quadro normativo che impone obblighi di trasparenza e labelling per i contenuti generati dall'intelligenza artificiale (come i deepfake che imitano persone), il ruolo dell'ACN non si limita alla difesa informatica, ma si estende alla garanzia che gli strumenti di AI rispettino rigorosi standard di **etica e sicurezza**. Ciò significa che l'ACN è chiamata a spingere l'uso dell'AI difensiva per la rilevazione e l'attribuzione dei contenuti manipolati, trasformando la lotta ai deepfake da una reazione tecnologica a posteriori a una gestione sistemica e normativa che mira a stabilire la fiducia e la provenienza del contenuto digitale fin dalla sua creazione: *"è anche in questa prospettiva di supporto all'azione istituzionale che le nostre attività sia a supporto della specifica C2PA che di partecipazione ai working group Europei devono essere viste"*, conclude il direttore del Centro Cybersecurity Silvio Ranise.

LINK

https://magazine.fbk.eu/it/news/la-sfida-per-lintegrita-dei-contenuti-digitali-nellera-dei-deepfake-e-della-genai/

TAG

- #ai compliance
- #AI difensiva
- #ai-act
- #C2PA
- #crittografia
- #cybercrime
- #cybersicurezza
- #dati
- #deepfake
- #genai
- #ia generativa
- #identità digitale
- #innovazione digitale
- #integrità
- #intelligenzaartificiale
- #labelling
- #marking
- #metadati
- #sicurezza
- #trust

AUTORI

- Michela Antino