



Deepfake: quando la finzione diventa credibile

Outline

- 1 Introduzione e definizioni
- 2 Storia del fenomeno
- 3 Case studies
- 4 IA Predittiva vs IA Generativa
- 5 IA Generativa
- 6 Truffe con Deepfake
- 7 Manipolazione dei Social Media
- 8 Interferenze Politiche
- 9 Normativa e policy
- 10 Conclusioni

Motivazione

- Negli ultimi 8 anni i modelli di generazione sono passati da semplici face-swap a pipeline multimodali (testo, immagine, audio, video) in tempo reale anche su smartphone.
- Clonare una voce convincente costa meno di 10 € in risorse cloud, rendendo l'IA accessibile a chiunque.
- Oggi i deepfake non sono più un “giocattolo”: servono nel marketing, nei videogame, nel cinema e, purtroppo, nella disinformazione.

Definizioni

Deepfake

Contenuto multimediale (video, audio, immagine) creato o alterato da modelli di Intelligenza Artificiale Generativa, con l'obiettivo di farlo sembrare reale. Il termine nasce su Reddit nel 2017 e oggi include qualunque manipolazione neurale sofisticata [21].

Intelligenza Artificiale

Disciplina che sviluppa sistemi capaci di funzioni cognitive umane (ragionamento, apprendimento, pianificazione) tramite algoritmi e grandi volumi di dati [19].

IA Generativa

Sottoinsieme dell'IA specializzato nella creazione di nuovi dati (immagini, testo, audio, video), a differenza dell'IA predittiva che “riempie spazi vuoti”.

Foundation Model

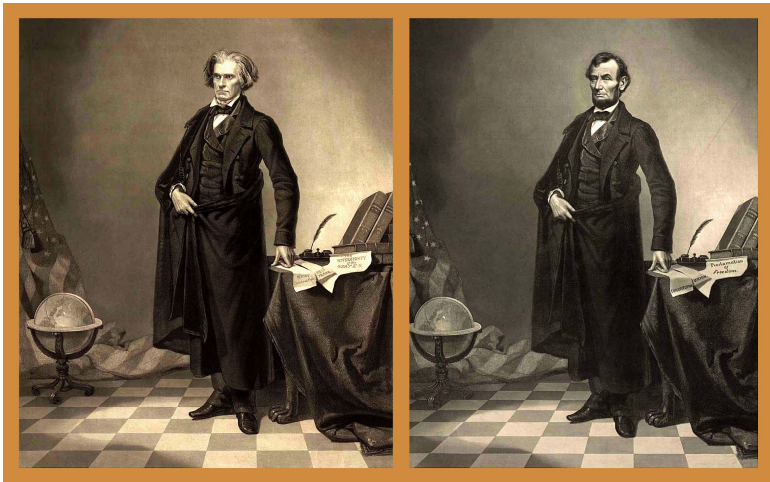
Modelli di grandi dimensioni (miliardi di parametri) addestrati su dataset generici e riadattabili a diversi compiti tramite prompt engineering (es. GPT-4o, LLaVA-3).

Normativa: Realistico vs Stilizzato

Secondo l'AI Act UE (§52), se il contenuto “appare reale” serve un’etichetta obbligatoria; se è manifestamente “stilizzato” (cartoon, illustrazioni), bastano avvertenze generiche.

Storia del fenomeno

Never trust quotes on the internet -Abraham Lincoln



Timeline 1860–2025 (1/3)

Origini pre-IA

- **1860 – Lincoln/Calhoun:** primi esperimenti di manipolazione fotografica mediante fusione di negativi su lastra, utilizzati come strumento di propaganda politica.
- **1917 – Fate di Cottingley:** due ragazze manipolano immagini con creature fantastiche; il caso diventa virale e crea la prima grande bufala fotografica.
- **1997 – Video Rewrite:** tecniche di warping labiale basate su Hidden Markov Models, antenato delle moderne manipolazioni video neurali.[5]

Timeline 1860–2025 (2/3)

Era delle GAN e social

- **2014:** Goodfellow et al. introducono le GAN, consentendo la generazione di immagini realistiche tramite competizione fra due reti neurali.
- **2016:** scandalo Cambridge Analytica dimostra il potere del *profiling* psicografico unito a messaggi persuasivi dinamici.[27]
- **2017:** nasce il subreddit /r/deepfakes; i primi face-swap pornografici si diffondono rapidamente.
- **2017:** Università di Washington pubblica un deepfake di Obama, primo video virale generato da IA.[22]

Timeline 1860–2025 (3/3)

Diffusione di massa e casi recenti

- **2020:** “Cheapfake” su Nancy Pelosi, manipolazione di velocità audio/video per creare falsi compromettenti.
- **2023:** Stable Diffusion 2&3 democratizzano immagini HD su GPU consumer; Google introduce SynthID, watermark latente resistente.[15]
- **2024:** robocall deepfake di Joe Biden nelle primarie del New Hampshire, uso di voice cloning per truffe politiche.[15]
- **2025:** LVD-XT genera video a 360° in 4 passi; prototipi di watermark hardware-level su ARM.
- **2025:** truffa CFO a Hong Kong con sei volti deepfake su Teams → 35 M USD; operazione “Spamouflage” a Taiwan smantellata da Meta.[9][18]

Case studies

Case study: Cambridge Analytica

Raccolta massiva di dati

- 2014: lancio dell'app “thisisyourdigitallife” sviluppata da Aleksandr Kogan.
- Oltre 270 000 utenti installarono volontariamente l'app, ma furono raccolti anche i dati dei loro amici (87 M profili) senza consenso diretto.
- Tipologie di dati acquisiti: like, preferenze, informazioni demografiche, network sociali e interazioni online.

Compliance apparente

- I dati venivano presentati come “uso accademico” da parte di ricercatori dell'Università di Cambridge (estrema alla vicenda)
- Facebook permise inizialmente l'accesso via API fino al 2015, prima di restringere drasticamente i permessi.

Case study: Cambridge Analytica

Profilazione psicografica & Microtargeting

- Applicazione del modello OCEAN per profilare personalità individuali.
- Segmentazione dell'elettorato in gruppi target con messaggi ultra-personalizzati (Brexit, Trump).
- A/B test su decine di varianti di annunci per massimizzare engagement.[27]

Ruolo dell'Università di Cambridge

- Kogan lavorava presso Cambridge, ma l'app fu sviluppata e distribuita senza alcuna approvazione o supervisione ufficiale.
- Cambridge ha preso le distanze, avviato indagini interne e dichiarato di non aver mai tratto vantaggio dai dati.

Case study: “Mia Ash”

Chi era “Mia Ash”?

- Synthetic persona creata nel 2016 dall’APT33 (OilRig).
- Foto e informazioni rubate da profili reali per dare credibilità.
- Profili LinkedIn e Facebook plausibili, con background professionale fittizio.
- Video preregistrati e audio deepfake per simulare colloqui di lavoro.[23]

Obiettivi dell’attacco

- Infiltrarsi in aziende del settore oil gas.
- Ottenere informazioni strategiche e credenziali di accesso.
- Stabilire un canale di comunicazione fidato per il payload.

Case study: “Mia Ash”

Modus operandi

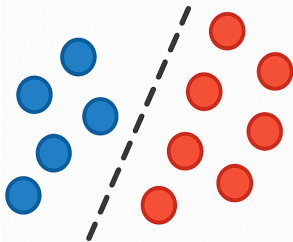
- Primo contatto: inviti su LinkedIn e messaggi diretti via Facebook.
- Chat video fake per guadagnare fiducia (deepfake audio/video).
- Invio di un documento Word malevolo che installa PupyRAT.
- Accesso persistente ai sistemi aziendali e movimento laterale.
- Evasione: utilizzo di tool di sistema per ridurre tracce (“living off the land”).

Implicazioni 2025

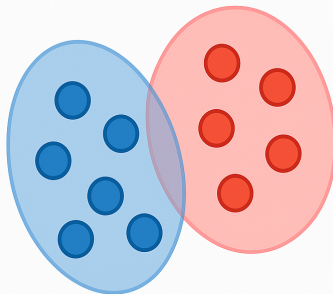
- Oggi creare synthetic personas richiede pochi click con tool IA multimodali.

IA Predittiva vs IA Generativa

IA Predittiva (Discriminativa)



IA Generativa



IA Predittiva – Modelli Classici

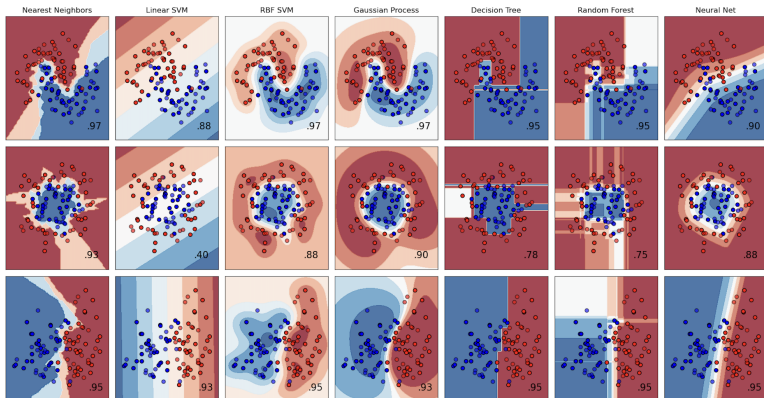
Obiettivo

Prevedere valori o classificazioni a partire da dati esistenti: “Cosa succederà?”

Tecniche più diffuse

Regressione lineare, Decision Tree, Random Forest, Neural Network offrono soluzioni consolidate per previsione e classificazione.

Possibili modelli



Dal “Cosa Succederà?” al “Cosa Posso Creare?”

Salto concettuale

Predittivo = riempire spazi vuoti con la risposta più probabile.

Generativo = inventare da zero contenuti plausibili mai esistiti.

Per i deepfake

Solo con l'IA generativa si possono creare volti, voci e scene altamente realistiche e convincenti.

IA Generativa

Cos'è l'IA Generativa?

Definizione

Modelli che apprendono lo “stile” di un dataset (immagini, testo, audio, video) e producono nuovi esempi simili a partire da semplici prompt.

Caratteristiche

Reinventano contenuti anziché selezionarli, interagiscono via testo/schizzi/audio e trovano applicazioni dall'arte alla disinformazione.

Famiglie di Modelli Generativi

GAN – Pittore vs Critico

Due reti (Generatore e Discriminatore) competono fino a generare contenuti indistinguibili dalla realtà.

VAE – Compressore Creativo

Encoder comprime l'input in uno spazio latente; Decoder lo ricostruisce, permettendo variazioni controllate.

Approcci Generativi

Diffusion – Scultura dal Rumore

Si parte da puro rumore e si rimuove progressivamente fino a emergere un'immagine coerente e nitida.

Transformer Autoregressivo

Genera token (parole, pixel, suoni) uno alla volta, preservando contesto e coerenza su lunghe sequenze.

IA Predittiva: modelli discriminativi (TECNICA)

Cosa fanno

- Stimano la distribuzione condizionale $P(y \mid x)$ (classificazione o regressione).
- Esempio: dato il profilo utente x , prevedere l'età o la categoria y .

Come si addestrano

- Minimizzano la *cross-entropy* (log-loss):

$$\mathcal{L} = -\mathbb{E}_{p_{\text{data}}(x,y)} [\log P(y \mid x)].$$

- Equivalente a minimizzare la divergenza di Kullback–Leibler:

$$\min D_{\text{KL}}(p_{\text{data}}(y \mid x) \parallel P(y \mid x)).$$

- Intuitivamente: avvicinano la *probabilità predetta* a quella *osservata*.

IA Generativa: modelli di probabilità (TECNICA)

Cosa fanno

- Stimano la distribuzione dei dati $P(x)$ o congiunta $P(x, y)$.
- Permettono di *campionare* nuovi esempi $x \sim P(x)$ (immagini, testo, audio).

Come si addestrano

- Massimizzano la *verosimiglianza* dei dati:

$$\max \mathbb{E}_{p_{\text{data}}(x)} [\log P(x)] \iff \min D_{\text{KL}}(p_{\text{data}}(x) \parallel P(x)).$$

- Intuitivamente: modellano *l'intera forma* dei dati, non solo le etichette.

Approfondimento: obiettivo GAN (TECNICA)

Min–Max avversariale

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$$

Spiegazione dei termini

- $\mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)]$: il discriminatore D cerca di assegnare alta probabilità (≈ 1) ai veri esempi x .
- $\mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$: D cerca di dare bassa probabilità (≈ 0) ai campioni sintetici $G(z)$.
- Il generatore G “inganna” D minimizzando $\log(1 - D(G(z)))$, cioè vuole che $D(G(z))$ sia alto (crede siano reali).

Approfondimento: obiettivo VAE (TECNICA)

Evidence Lower Bound (ELBO)

$$\log p(x) \geq \mathbb{E}_{z \sim q(z|x)} [\log p(x | z)] - D_{\text{KL}}(q(z | x) \| p(z))$$

Spiegazione dei termini

- $\mathbb{E}_{q(z|x)} [\log p(x | z)]$: termine di *ricostruzione*, misura quanto bene il decoder riproduce x dal codice z .
- $D_{\text{KL}}(q(z | x) \| p(z))$: regolarizzatore che avvicina la distribuzione latente $q(z | x)$ al prior $p(z)$ (tipicamente $\mathcal{N}(0, I)$).
- Massimizzare ELBO = migliorare la qualità delle ricostruzioni + mantenere il latente strutturato.

Come VAE genera nuovi esempi (TECNICA)

Fase di campionamento

- Si campiona $z \sim p(z)$ (il prior, es. $\mathcal{N}(0, I)$).
- Si genera $x' \sim p(x | z)$ tramite il decoder.
- Ogni x' è un nuovo dato plausibile secondo la distribuzione appresa.

Confronto con i discriminativi

- **Discriminativi:** non hanno latente, non possono campionare.
- **VAE:** modellano direttamente $P(x)$ attraverso ELBO, consentendo la creazione di nuovi campioni.

Implicazioni per i deepfake (TECNICA)

Perché servono modelli generativi

- Solo minimizzando $D_{\text{KL}}(p_{\text{data}}(x) \parallel P(x))$ si *creano* contenuti nuovi e realistici.
- I modelli discriminativi $P(y \mid x)$ non campionano dati: servono solo a *riconoscere* o *classificare*.
- Deepfake = generare pixel, audio o video che imitano la distribuzione reale dei media.

Applicazioni Diffuse

Stable Diffusion (immagini da testo), ChatGPT (testo conversazionale), VALL-E (clone vocale), Make-a-Video (brief video clip).

Truffe con Deepfake

Caso Emirati Arabi 2021

- Deepfake audio che replica in modo quasi perfetto la voce del CEO, creato da registrazioni pregresse.
- Il CFO, convinto dall'autenticità, dispone un trasferimento di 35 M USD verso un conto fraudolento.[6][11]
- L'attacco era supportato da:
 - Email di phishing estremamente mirate;
 - Clone vocale con accento, inflessioni e pause naturali.

Frodi Aziendali con Voice Cloning

Caso Europa 2019

- Deepfake audio del CEO generato a partire da pochi secondi di registrazione su Slack.
- Il reparto tesoreria, persuaso dalla corrispondenza vocale, ha eseguito un trasferimento di 220 k EUR su un conto estero.[10]
- Profilo vocale così fedele da ingannare anche gli analisti di sicurezza.

Kidnapping Scam

- Raccolta di pochi secondi di audio di un familiare da social media.
- Generazione di una chiamata deepfake in cui la voce simula un sequestro e chiede aiuto.
- Elevato realismo emotivo spinge le vittime a versare il riscatto immediatamente.
- Nel 2022 FBI e FTC hanno registrato un aumento del 300 % di questi attacchi, consigliando di adottare codici di verifica condivisi in anticipo.[13][16]

Falso Elon Musk 2022

- Un video deepfake mostrava Elon Musk presentare “BitVex”, piattaforma di trading cripto con “profitti garantiti” del 30
- L'alta qualità del volto e della voce clonata ha convinto migliaia di utenti a investire.
- Si è trattato di uno schema di pump-and-dump: il valore delle token BitVex è crollato dopo la smentita ufficiale di Musk.
- Dimostra come la fiducia nei personaggi pubblici possa essere sfruttata per manipolare i mercati.[26][3]

Manipolazione dei Social Media

Amplificazione automatica

- Le botnet sociali sono reti di account automatizzati che postano e condividono contenuti in massa secondo schemi prestabiliti.
- Uno studio dell'USC del 2020 ha rilevato che oltre il 20 % del traffico politico su Twitter negli USA proveniva da bot [25].
- Questi bot creano “echo chamber” virtuali ripubblicando ripetutamente gli stessi messaggi, dando l'illusione di un consenso ampio [30].
- Manipolano gli algoritmi delle piattaforme aumentando artificialmente i livelli di engagement e influenzando la visibilità dei contenuti.

Fabbriche di Like

- Le click farm impiegano operatori umani, mentre le like farm usano bot automatizzati per generare follower, like e visualizzazioni a pagamento.
- Un'inchiesta del 2022 in Bangladesh ha documentato giovani pagati pochi centesimi per produrre migliaia di interazioni fasulle.[8]
- Secondo l' "Imperva 2022 Bad Bot Report", i bot farm sfruttano proxy e account fake per eludere i filtri anti-frode delle piattaforme.[17]

Fabbriche di Like

- Gli algoritmi di raccomandazione amplificano questi segnali artificiali, promuovendo post e hashtag come autenticamente popolari.
- Implicazione politica: profili e campagne sponsorizzate ottengono visibilità ingiustificata, distorcendo il dibattito pubblico.

Identità Inventate

- I “sockpuppet” sono profili gestiti manualmente da operatori che usano foto e dati rubati per impersonare persone reali.
- Nel 2021 il Centre for Information Resilience ha smascherato 80 sockpuppet attivi su Twitter, Facebook e Instagram in India.[7][4]
- Questi account si sono infiltrati in community chiuse e gruppi di discussione per guadagnare fiducia.
- Una volta consolidata la reputazione, hanno iniziato a diffondere contenuti di propaganda nazionalista.

Comportamento Inautentico Coordinato

Operazioni su Larga Scala

- Il Coordinated Inauthentic Behavior (CIB) unisce bot e account umani in campagne sinergiche.
- Nel 2021 Meta ha rimosso 52 reti CIB operanti in oltre 30 Paesi, molte con link a governi esteri o gruppi privati.[18][1]
- Tecniche utilizzate:
 - Pagine esca e siti clone di testate giornalistiche.
 - Raccolta e rilancio sincronizzato di narrazioni su più piattaforme.
- Obiettivo: manipolare il dibattito pubblico creando l'illusione di un ampio consenso.

Interferenze Politiche

Propaganda AI-driven

- Dal 2023, Russia e altri attori esterni hanno diffuso video deepfake di funzionari USA; ad esempio, un falso rappresentante del Dipartimento di Stato annunciava attacchi militari inesistenti.[20]
- Siti pseudo-giornalistici come “D.C. Weekly” sono stati creati ad arte per diffondere fake news e minare la fiducia nelle fonti ufficiali.[28]
- Obiettivo: erodere la credibilità delle istituzioni e preparare il terreno a forme di influenza politica anche oltreoceano.

Interferenze Politiche: India

Deepfake in Dialetti Locali

- Nelle elezioni di Delhi 2020 il BJP ha diffuso video deepfake in hindi e haryanvi, lingue non fluentemente parlate dal leader.[2]
- Sincronizzazione digitale del labiale e voce sintetica hanno creato l'illusione di un discorso diretto alle comunità locali.

Sockpuppet Patriottici

- Reti di account falsi hanno veicolato messaggi nazionalisti contro dissidenti Sikh.[4]
- Foto rubate e narrazioni coordinate venivano usate per segmentare il pubblico e seminare discordia.

Interferenze Politiche: Taiwan

Operazione “Spamouflage”

- Campagne su YouTube, Instagram e Twitter con video deepfake contro la presidente Tsai Ing-wen.[29]
- Meta ha rimosso migliaia di account coinvolti in comportamenti inautentici coordinati.

Contromisure

- Introduzione di leggi anti-deepfake durante la campagna elettorale.
- Lancio di programmi di alfabetizzazione mediatica per cittadini e giornalisti.

Video Fake di Zelenskyj

- All'inizio dell'invasione russa del 2022, un deepfake di Zelenskyj annunciava la resa dell'Ucraina.[12]
- Il clip fu trasmesso brevemente da un canale TV compromesso, causando disorientamento.
- Il vero presidente Zelenskyj smentì ufficialmente via social, ripristinando la chiarezza.

Trasparenza e Responsabilità

Chi diffonde contenuti “verosimili” generati o manipolati da IA deve:

- 1 Etichettarli chiaramente come “sintetici / AI-generated”.
- 2 Conservare e rendere disponibili i metadati C2PA.
- 3 Rimuovere su richiesta delle autorità in caso di illeciti (terrorismo, pornografia non consensuale, ecc.).

Sanzioni

Fino a 35 M € o 7 % del fatturato globale annuo, a tutela dell'integrità dell'informazione.

Panoramica Internazionale

- **USA (FCC 2024):** vietate robocall con voce AI senza consenso scritto; multe fino a \$43 000 per chiamata.[14]
- **UK (Online Safety Act 2023):** reato specifico per deepfake pornografici; pene fino a 2 anni di detenzione.
- **Italia (DDL 1146/24-25):** obbligo di etichettatura dei contenuti sintetici; multa pari al 2 % del fatturato e obbligo di oscuramento del sito in caso di inadempienza.[24]

Sintesi dei Punti Chiave

- ❶ **Commodity IA:** i deepfake sono oggi strumenti accessibili a molti, non solo a pochi laboratori di ricerca.
- ❷ **Difesa multilivello:** combinare watermark hardware, standard C2PA, educazione digitale e tecniche forensi.
- ❸ **Rischio real-time:** live generation di audio/video in video-chiamata (Teams, Zoom) sarà il prossimo fronte d'attacco.

Riferimenti I

- [1] Atlantic Council DFRLab. *Disinformation in the Age of AI*. Rapp. tecn. Atlantic Council DFRLab, 2024.
- [2] BBC. “Indian Political Deepfakes in Dialects”. In: *BBC* (2020). Online edition.
- [3] BBC News. *Crypto Scam Flavored by Deepfake Elon Musk Video*. 2022. URL: <https://www.bbc.com/news/technology-61345678>.
- [4] BBC News. “Fake Sikh Accounts Exposed”. In: *BBC News* (2021). Online edition.
- [5] Christoph Bregler, Michael Covell e Malcolm Slaney. “Video Rewrite”. In: *Proceedings of ICASSP*. 1997.

- [6] Thomas Brewster. *Fraudsters Cloned Company Director's Voice In \$35 Million Heist Using AI*. 2021. URL: <https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/>.
- [7] Centre for Information Resilience. *Exposing Organised Sockpuppet Influence Operations*. Rapp. tecn. Centre for Information Resilience, 2021. URL: <https://informationresilience.org/sockpuppet-report-2021>.
- [8] Channel 4 News. "Bangladesh Like Farms Revealed". In: *Channel 4 News* (2022). Online investigation.

Riferimenti III

- [9] Communications Security Establishment Canada. *Hong Kong CFO Deepfake Scam*. Rapp. tecn. CSE Canada, 2025.
- [10] Jesse Damiani. *A Voice Deepfake Was Used To Scam A CEO Out Of \$243 000*. 2019. URL: <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>.
- [11] Dark Reading. *Deepfake Audio Scores \$35 Million in Corporate Heist*. 2021. URL: <https://www.darkreading.com/cyberattacks-data-breaches/deepfake-audio-scores-35-million-in-corporate-heist>.

Riferimenti IV

- [12] DFRLab. *Zelensky Fake Video Analysis*. Rapp. tecn. Atlantic Council DFRLab, 2022.
- [13] FBI. *FBI Warns of Deepfake “Kidnap” Scams Targeting Families*. 2022. URL:
<https://www.fbi.gov/news/press-releases/fbi-warns-of-deepfake-kidnap-scams-targeting-families>.
- [14] FCC. *AI Voice Regulation*. Rapp. tecn. Federal Communications Commission, 2024.
- [15] FCC. *Robocall Deepfake Biden*. Rapp. tecn. Federal Communications Commission, 2024.

- [16] FTC. *Consumer Alert: Deepfake Kidnapping Scams*. 2022.
URL:
<https://www.ftc.gov/news-events/blogs/consumer-alert/2022/09/deepfake-kidnapping-scams>.
- [17] Imperva Inc. *2022 Bad Bot Report*. Rapp. tecn. Imperva, 2022.
URL:
<https://www.imperva.com/resources/reports/bad-bot-report-2022.pdf>.
- [18] Meta. *CIB Reports*. Rapp. tecn. Meta Platforms, Inc., 2021.
- [19] NIST. *Definition of Artificial Intelligence*. NIST Special Publication. 2025.

- [20] NPR. *Fake State Dept Video*. 2023. URL: <https://www.npr.org/sections/codeswitch/2023/01/15/849572764/fake-state-department-video-deepfake>.
- [21] NSA. *Definition of Deepfake*. Technical definition by the U.S. National Security Agency. 2025.
- [22] Reuters. *Fake Obama Video*. 2017. URL: <https://www.reuters.com/article/usa-obama-deepfake-idUSKBN1D82A2>.
- [23] SecureWorks. *Mia Ash Report*. Rapp. tecn. SecureWorks, 2016.
- [24] Senato della Repubblica Italiana. *DDL 1146/24-25*. Rapp. tecn. Senato Italiano, 2025.

Riferimenti VII

- [25] Chengcheng Shao et al. “The Spread of Low-Credibility Content by Social Bots”. In: *Nature Communications* 11.1 (2020), 1–9. DOI: 10.1038/s41467-019-13886-0.
- [26] John Smith. *Deepfake Video of Elon Musk Promotes Fake Crypto “BitVex”*. 2022. URL: <https://www.forbes.com/sites/johnsmith/2022/05/10/deepfake-elon-musk-bitvex-scam>.
- [27] Vox. *Cambridge Analytica Explained*. 2018. URL: <https://www.vox.com/policy-and-politics/2018/3/23/17151944/cambridge-analytica-facebook-kogan-explained>.
- [28] Washington Post. “Fake “D.C. Weekly” Site Uncovered”. In: *The Washington Post* (2023). Online edition.

- [29] Washington Post. “Taiwan Deepfake Election Videos”. In: *The Washington Post* (2023). Online edition.
- [30] Savvas Zannettou et al. *Understanding Domestic and Foreign Influence Operations in Online Social Networks*. Rapp. tecn. Echolab, University of Twente, 2020.