



Deepfake

Outline

- ① Definizioni
- ② Breve storia
- ③ Fondamenti teorici/computazionali
- ④ Difesa: come rilevarle ?
- ⑤ Aspetti legali
- ⑥ Manipolazione

Definizioni

Deepfake e Intelligenza Artificiale - definizioni

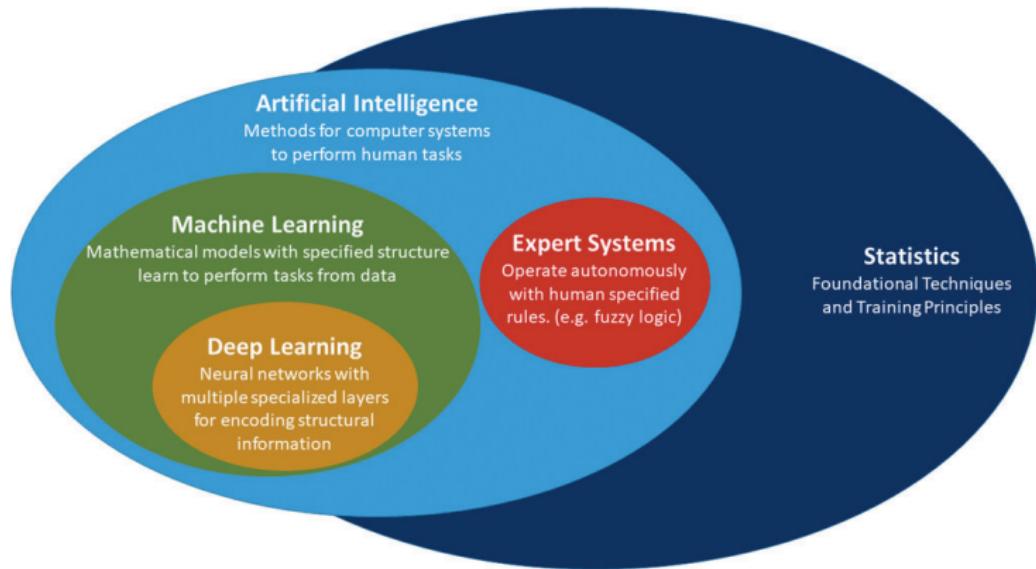
Deepfake (NSA) [1]

Contenuto multimediale creato sinteticamente o manipolato utilizzando una qualche forma di tecnologia meccanica o di deep learning.

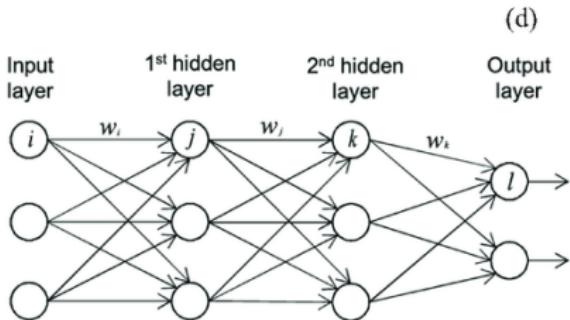
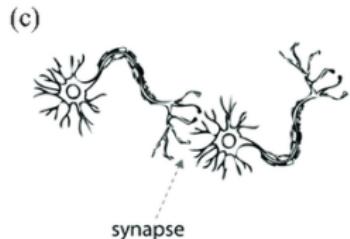
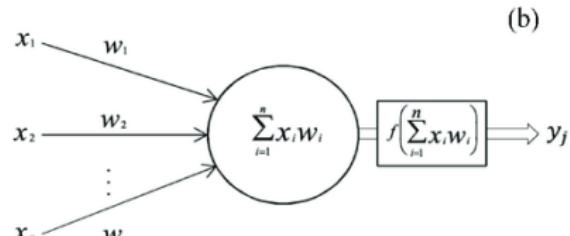
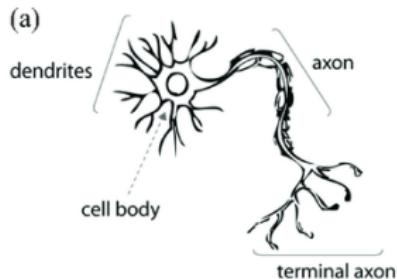
Intelligenza artificiale (NIST) [2]

Una branca dell'informatica dedicata allo sviluppo di sistemi di elaborazione dati che svolgono funzioni normalmente associate all'intelligenza umana, come il ragionamento, l'apprendimento e l'auto-miglioramento

Deepfake e Intelligenza Artificiale - tassonomia [3]



Deepfake e Intelligenza Artificiale - reti neurali [4]



Breve storia

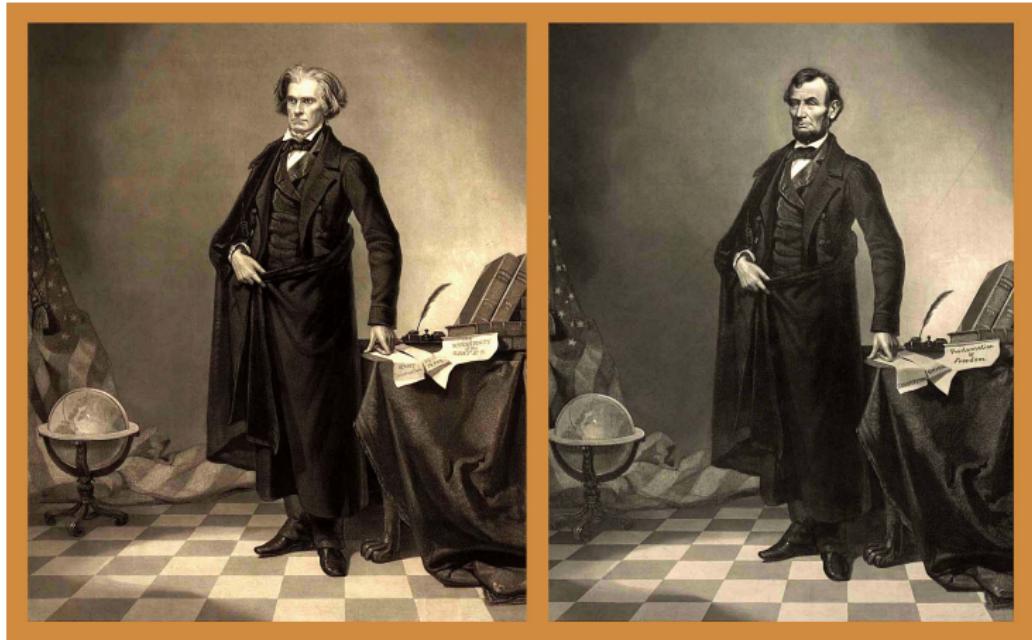
Storia dei deepfake

“Never trust
anything you
see on the
internet”

Abraham Lincoln



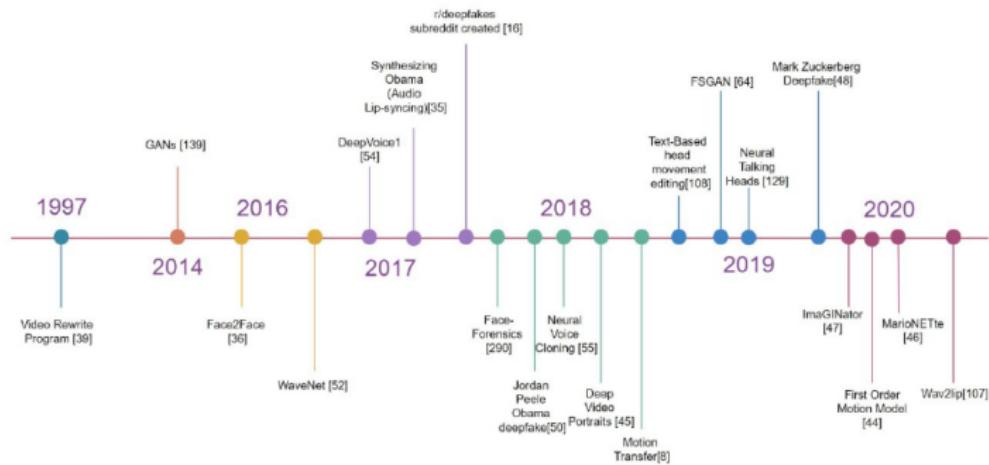
Storia dei deepfake: 1860 Calhoun/Lincon [5]



Storia dei deepfake: 1997 Video Rewrite (GAN) [6]



Deepfake e Intelligenza Artificiale - tassonomia [18]



Storia dei deepfake: 2017 B.Obama [7]

Print subscriptions Sign in Search jobs Search Europe edition

Support the Guardian

Fund independent journalism with €10 per month

Support us →

News Opinion Sport Culture Lifestyle More ▾

Technology

This article is more than 6 years old

The future of fake news: don't believe everything you read, see or hear

A new breed of video and audio manipulation tools allow for the creation of realistic looking news footage, like the now infamous fake Obama speech

Olivia Solon in San Francisco
Wed 26 Jul 2017 07.00 CEST

Share

Synthesizing Obama: Learning Lip Sync from Audio Copia link

Without Re-timing With Re-timing (Our Result)

Guarda su YouTube

The University of Washington's Synthesizing Obama project took audio from one of Obama's speeches and used it to animate his face in an entirely different video

Advertisement

veeam

NEW DEMO

Unlock Kubernetes Application Mobility

CLICKABLE DEMO

Storia dei deepfake: 2017 Imitazione della voce [8]

BANKRUPTCY CENTRAL BANKING CYBERSECURITY PRIVATE EQUITY SUSTAINABLE BUSINESS VENTURE CAPITAL

WSJ PRO CYBERSECURITY

Home News Research Archive Newsletters Events

WSJ PRO

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

By Catherine Stipp Updated Aug. 30, 2019 12:52 pm ET | WSJ PRO

Share AA Resize



PHOTO: SIMON DAWSON/BLOOMBERG NEWS

Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 (\$243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

The CEO of a U.K.-based energy firm thought he was speaking on the phone with his boss, the chief executive of the firm's German parent company, who asked him



Draußen bleiben,
trocken bleiben
Wasserdichter Schutz,
hergestellt ohne PFCs.

Regenbekleidung entdecken

MUST READS FROM CYBERSECURITY

- Port of Rotterdam Tests Quantum Network to Defend Against Hacks
- Otta Hack Update

Storia dei deepfake: 2020 N.Pelosi [8]



World

Fact check: “Drunk” Nancy Pelosi video is manipulated

By **Reuters**

August 3, 2020 7:23 PM GMT+2 · Updated 4 years ago



U.S. Speaker of the House Nancy Pelosi, joined by Senate Minority Leader Chuck Schumer, speaks to reporters in the U.S. Capitol in Washington, U.S. July 29, 2020. REUTERS/Erin Scott [Purchase Licensing Rights](#)

A video circulating on social media shows House Speaker Nancy Pelosi speaking in a slurred and awkward manner. One popular post boasts 91,000 shares on Facebook and bears a caption reading: “This is unbelievable, she is blown out of her mind, I bet this gets taken down!” The video, however, has been manipulated to make Pelosi appear drunk and incoherent.

A viral example of the video is visible [here](#).

Storia dei deepfake: Software disponibili [18]

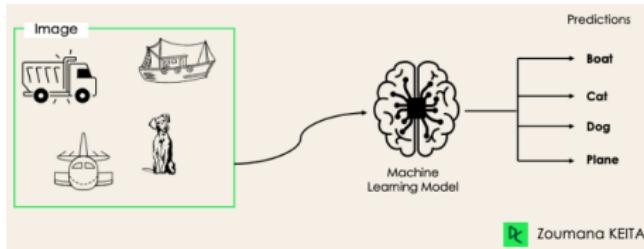
Tool	Type	Reference/Developer	Technique
Cheat fakes			
Adobe Premiere	Commercial Desktop Software	Adobe	Audio Video Editing, AI-powered video refacing
Corel VideoStudio	Commercial Desktop Software	Corel	Proprietary AI
Lip-sync			
dynalips	Commercial Web App	www.dynalips.com	Proprietary
cryctalk	Commercial Web App	www.realmon.com/cryctalk/	Proprietary
Wave2Lip	Open source implementation	github.com/Rudalabs/Wave2Lip	GAN with pre-trained discriminator network and visual quality loss function
Facial Attribute Manipulation			
FaceApp	Mobile App	FaceApp Inc	Deep generative CNNs
Adobe	Commercial Desktop Software	Adobe	DNNs + filters
Rosebud	Commercial Web App	www.rosebud.ai/	Proprietary AI
Face Swap			
ZAO	Mobile app	Momo Inc	Proprietary
REFACE	Mobile app	Neocortext, Inc	Proprietary
Reflect	Mobile app	Neocortext, Inc	Proprietary
Impressions	Mobile app	Synthesized Media, Inc.	Proprietary
FakeApp	Desktop App	www.malavida.com/en/soft/fakeapp/	GAN
FaceSwap	Open source implementation	faceswapweb.com/	Employed two pairs of encoder-decoder. Shared encoder parameters.
DFaker	Open source implementation	github.com/dfaker/df	For face reconstruction DSSIM loss function [44] is utilized. Keras library-based implementation.
DeepFaceLab	Open source implementation	github.com/iperon/DeepFaceLab	- provide several face extraction methods, e.g. dlib, MTCNN, SIFT etc. - Extend different Face swap model i.e. H64, H128, LIAEF128, SAE [33]
FaceSwapGAN	Open source implementation	github.com/taosunlu/faceswap-GAN	Uses two loss functions namely adversarial loss and perceptual loss to the auto-encoder.
DeepFake-Tf	Open source implementation	github.com/StromWine/DeepFake-tf	Same as DFaker however, used tensor-flow for implementation.
Facesswapweb	Commercial Web App	faceswapweb.com/	GAN
Face Recreament			
Face2Face	Open source implementation	web.stanford.edu/~zhifeng/papers/CVPR2016_Face2Face/page.html	Uses 3DMM and ML technique
Dynamixyz	Commercial Desktop Software	www.dynamixyz.com/	Machine-learning
FaceiT3	Open source implementation	github.com/idev3/faceit_live3	GAN
Face Generation			
Generated Photos	Commercial Web App	generated photos/	StyleGAN
Video Synthesis			
Overdub	Commercial Web App	www.descript.com/overdub	Proprietary (AI based)
Reespecter	Commercial Web App	www.reespecter.com/	Combined traditional digital signal processing algorithms with proprietary deep generative modeling techniques
SV2TTS	Open source implementation	github.com/Continuum/Real-Time-Voice-Cloning	LSTM with Generalized end-to-end loss
ResembleAI	Commercial Web App	www.resemble.ai/	Proprietary (AI based)
Voxery	Commercial Web App	www.voxery.com/	Proprietary AI and deep learning
VoiceApp	Mobile app	Zoeei AB	Proprietary (AI-based)

Fondamenti teorici/computazionali

Reti neurali [20, 19, 9]

Distribuzione di probabilità condizionata - idea

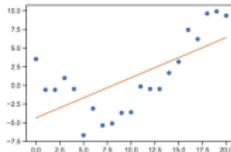
Abbiamo una serie di record che contengono delle variabili input (dimensione del fiore, colore etc..) e una variabile di output (caso semplice, può essere generalizzato). Non sappiamo a priori come input e output siano legati (altrimenti avremmo risolto il problema !), ma siamo interessati a trovare una funzione di probabilità che associa degli input a degli output. $p(y = c|x; \theta) = f_c(x; \theta)$



Reti neurali [20, 19, 9]

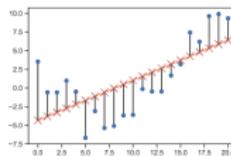
Regressione lineare

$$f(x; \theta) = b + wx$$



Regressione polinomiale

$$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$



Non-linear feature extraction

$$f(\mathbf{x}; \mathbf{w}; \mathbf{V}) = \mathbf{w}^T \phi(\mathbf{x}; \mathbf{V})$$

No free lunch theorem

Non esiste un modello che abbia performance migliori in tutti i casi possibili. Ciò è dovuto al fatto che ogni modello fa delle assunzioni.

Deep neural network

$$f(\mathbf{x}; \theta) = f_L(f_{L-1}(\dots(f_1(x))\dots))$$

Reti neurali [20, 19]

Perceptrons

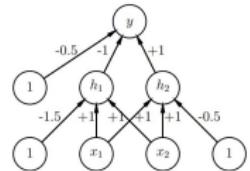
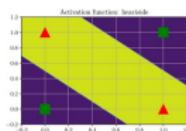
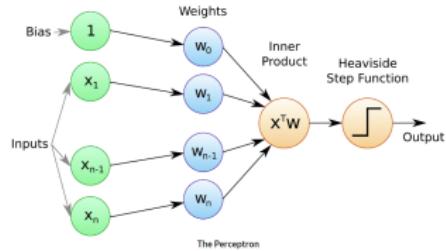
$$\begin{aligned}f(\mathbf{x}; \theta) &= \\&= 1(\mathbf{w}^T \mathbf{x} + b \geq 0) \\&= H(\mathbf{w}^T \mathbf{x} + b)\end{aligned}$$

MLP

$$z_I = f_I(z_{I-1}) = \varphi(\mathbf{b}_I + \mathbf{W}_I z_{I-1})$$

$$z_{kl} = \varphi_l \left(b_{kl} + \sum_{j=1}^{K_{l-1}} w_{lkj} z_{jl-1} \right)$$

$$a_I = b_I + \mathbf{W}_I z_{I-1}$$

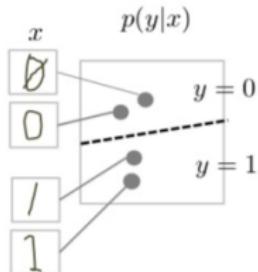


Modelli generativi [10, 20, 19]

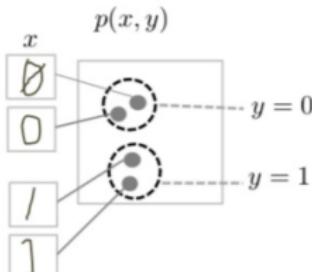
Modello generativo

Un modello generativo descrive come viene generato un set di dati utilizzando un modello probabilistico ($p(x) \quad x \in X$). Campionando da questo modello, siamo in grado di generare nuovi dati.

- Discriminative Model



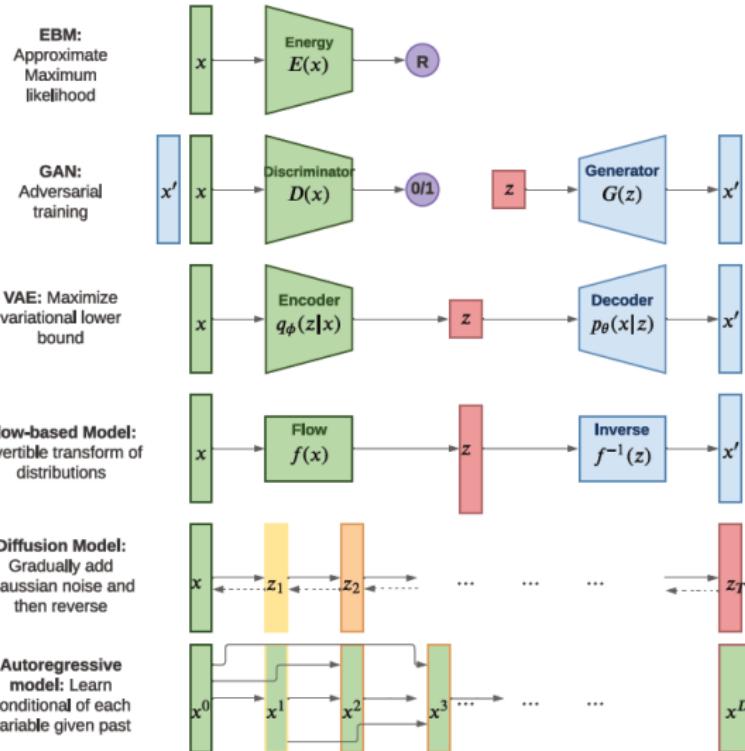
- Generative Model



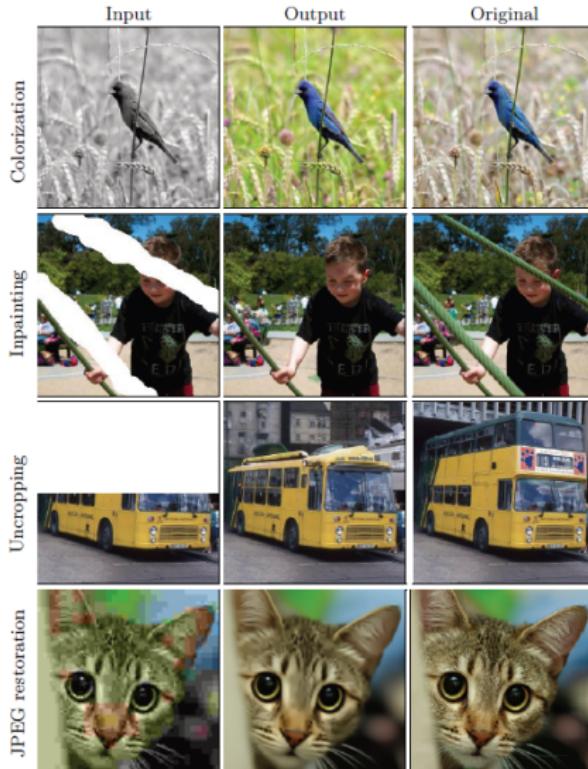
Tipologia di modelli generativi

- Deep generative models (DGM): viene usata una deep neural network. Comprende i Variational Autoencoder (VAE) e le Generative adversarial network oltre che, ad esempio, i Diffusion models e gli Energy based models (EBM)
- Probabilistic graphical models: viene usato un grafo causale

Modelli generativi [20, 19]



Modelli generativi: esempi [20, 19]



Autoencoders [20, 19]

Autoencoder

Encoder (f_e) + Decoder f_d

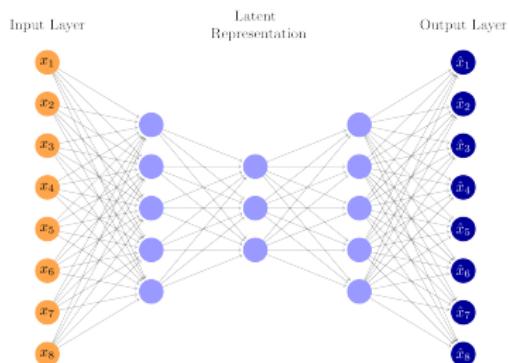
$$f_e : x \rightarrow z \quad f_d : z \rightarrow x$$

$r(x) = f_d(f_e(x))$ (rec. function)

$\mathcal{L}(\theta) = \|r(x) - x\|^2$ (loss function)

Funzionamento

L'unità in mezzo agisce come collo di bottiglia tra l'input e la sua ricostruzione, in modo da applicare una compressione.



Autoencoders [20]



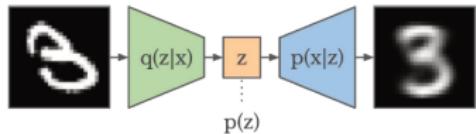
Variational autoencoders [20, 19]

Modello generativo

Un modello generativo descrive come viene generato un set di dati utilizzando un modello probabilistico ($p(x) \quad x \in X$). Campionando da questo modello, siamo in grado di generare nuovi dati.

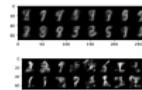
Differenza rispetto agli AE

Rispetto agli autoencoder i variational autoencoder possono essere visti come una versione probabilistica di un autoencoder deterministico. In questo modo si può avere una IA generativa



Remark

La VAE, per come è costruita, è in grado di convertire punti casuali in output, mentre il decoder di AE (deterministico) funziona solo per un punto che è già presente nel training.



VAR generation [20, 19]



Generative adversarial networks [20, 19]

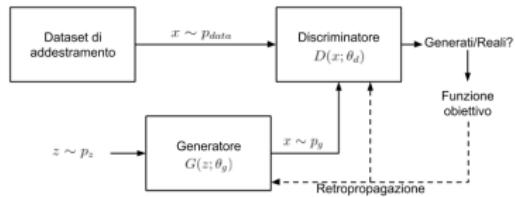
Idea

Ho due reti neurali

- Generatore: creare nuovi dati
- Discriminatore: distinguere i dati nuovi da quelli veri

Apprendimento

$$\min_G \max_D E_{x \sim p_{\text{data}}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$



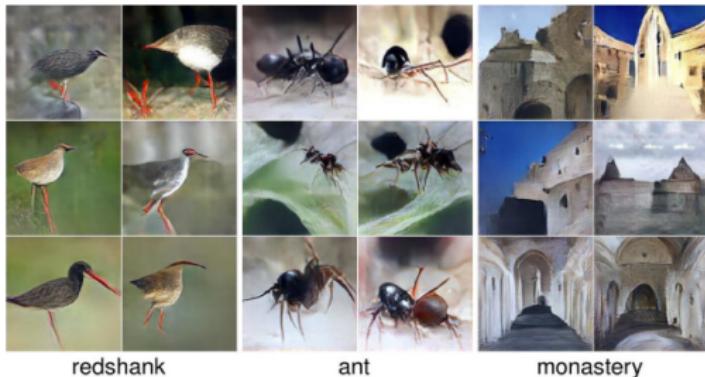
Algorithm 26.2: GAN training algorithm

```
1 Initialize  $\phi, \theta$ 
2 for each training iteration do
3   for  $K$  steps do
4     Sample minibatch of  $M$  noise vectors  $z_m \sim q(z)$ 
5     Sample minibatch of  $M$  examples  $x_m \sim p(x)$ 
6     Update the discriminator by performing stochastic gradient descent using this gradient:
    
$$\nabla_{\theta} \frac{1}{M} \sum_{m=1}^M [g(D_{\phi}(x_m)) + \nabla_{\phi} h(D_{\phi}(G_{\theta}(z_m)))]$$

7     Sample minibatch of  $M$  noise vectors  $z_m \sim q(z)$ 
8     Update the generator by performing stochastic gradient descent using this gradient:
    
$$\nabla_{\theta} \frac{1}{M} \sum_{m=1}^M l(D_{\phi}(G_{\theta}(z_m)))$$

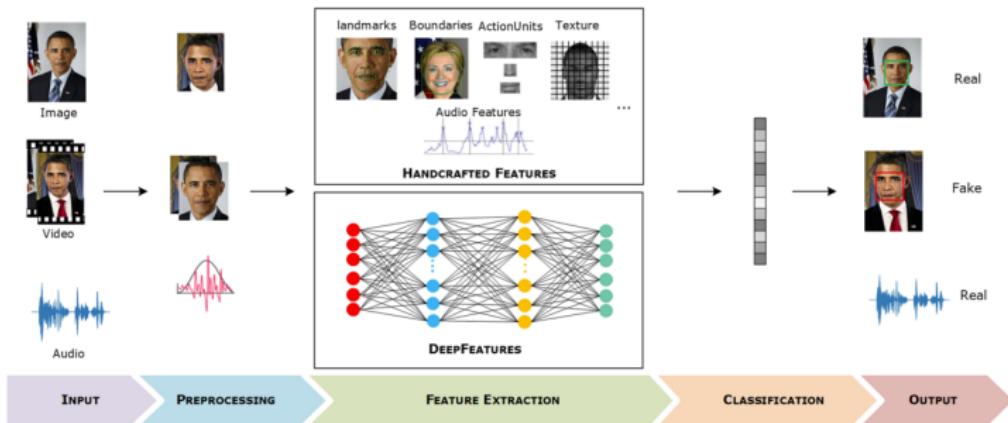
9 Return  $\phi, \theta$ 
```

VAR generation [21]

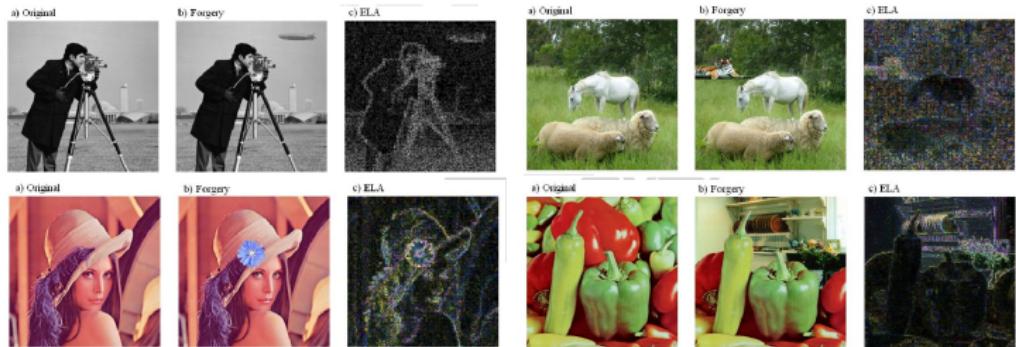


Difesa: come rilevarle ?

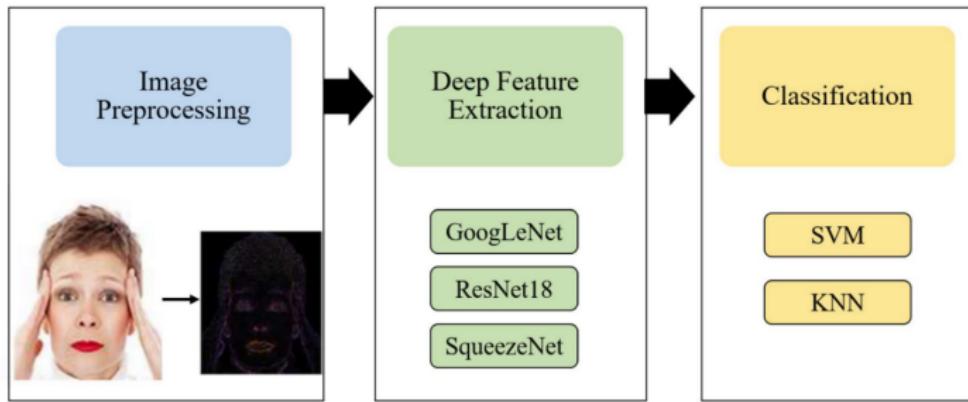
Metodi [18]



Error level analysis [17]



Error level analysis + Deep Learning [22]



Aspetti legali

70a) A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. **In the light of those impacts, the fast technological pace and the need for new methods and techniques to trace origin of information, it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human.** Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate

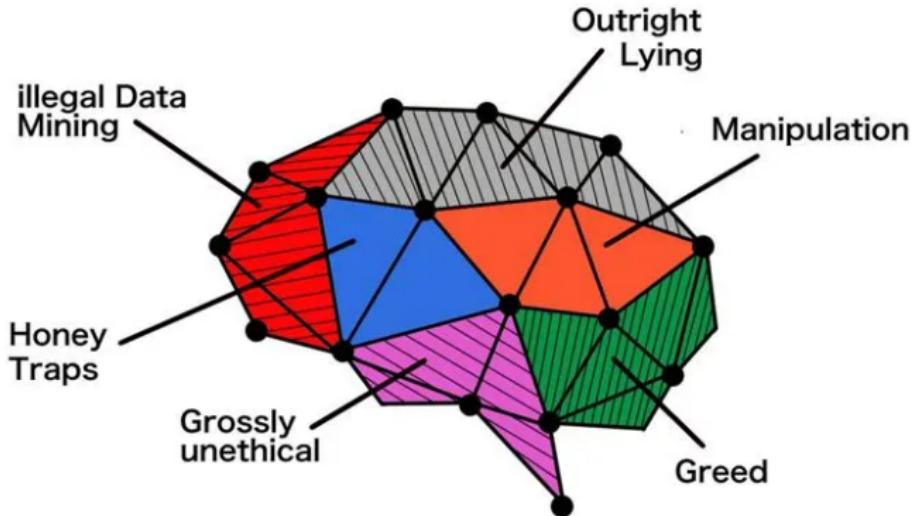
Conseguenze di un uso improprio dei deepfake [11]

Uso improprio dei deepfake

- Diffamazione, attraverso la creazione di contenuti che denigrano o danneggiano la reputazione di un individuo
- Furto di identità, che rileva nel caso della creazione di video o audio in cui un soggetto viene rappresentato come un'altra persona
- Violazione della privacy, attraverso la condivisione non consensuale di informazioni relative a una persona

Manipolazione

Cambridge Analytica [12]



Cambridge Analytica

Cambridge Analytica [13]

How was Facebook users' data misused?

- 1 In 2014 a Facebook quiz invited users to find out their personality type
- 2 The app collected the data of those taking the quiz, but also recorded the public data of their friends
- 3 About 305,000 people installed the app, but it gathered information on up to 87 million people, according to Facebook
- 4 It is claimed the data was sold to Cambridge Analytica (CA), which used it to psychologically profile voters in the US
- 5 CA denies it broke any laws and says it did not use the data in the US presidential election
- 6 Facebook sends notices to users telling them whether their data was breached

CA denies any wrongdoing. Facebook has apologised to users and says a "breach of trust" has occurred.

BBC

Cambridge Analytica [14]

≡ MENU CERCA

LA STAMPA

IL QUOTIDIANO ABBONATI [ACCEDI](#)

Economia

Lavoro Agricoltura TuttoSoldi Finanza Borsa Italiana Fondi Obbligazioni

Meta paga 725 milioni di dollari per risolvere class action Cambridge Analytica

TELEBORSA

Pubblicato il 23/12/2022
Ultima modifica il 23/12/2022 alle ore 13:03

cerca un titolo



Meta Platforms (società che controlla Facebook, Instagram e WhatsApp) ha accettato di pagare 725 milioni di dollari per risolvere una class action che accusava il colosso dei social media di consentire a terzi, tra cui Cambridge Analytica, di accedere alle informazioni personali degli utenti. La società fondata da Mark Zuckerberg non ha ammesso alcun illecito come parte dell'accordo, che è soggetto all'approvazione di un giudice federale di San Francisco, ma ha sottolineato che l'accordo è "nel migliore interesse della nostra comunità e degli azionisti".

LEGGI ANCHE

04/04/2024 
New York: risultato positivo per Meta Platforms

19/04/2024

Cambridge Analytica [15]

ANSA.it › Tecnologia › Internet & Social ›

Garante Privacy, multa 1 mln a Facebook per Cambridge Analytica

Garante Privacy, multa 1 mln a Facebook per Cambridge Analytica

Società attraverso app ha avuto accesso a dati 87 milioni di utenti

Redazione ANSA

28 giugno 2019

16:54

ANALISI

 Suggerisci

 Facebook

 Twitter

 Altri



 Stampa

 Scrivi alla redazione



Garante Privacy, multa 1 mln a Facebook per Cambridge Analytica © ANSA/AP

SCOPRI

FINO A -50%

su VINI, BIRRE e SPIRITS

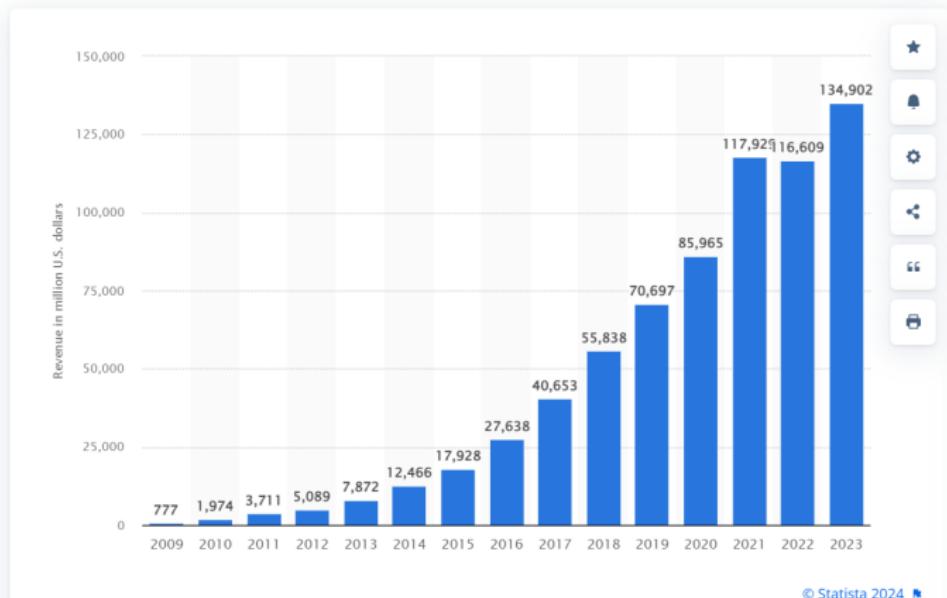


CLICCA PER INGRANDIRE 

Cambridge Analytica [16]

Internet > Social Media & User-Generated Content

Annual revenue generated by Meta Platforms from 2009 to 2023 (in million U.S. dollars)



Bibliography I

- [1] [https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/.](https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/)
- [2] [https://csrc.nist.gov/topics/technologies/artificial-intelligence.](https://csrc.nist.gov/topics/technologies/artificial-intelligence)
- [3] [https://journals.ametsoc.org/view/journals/bams/103/5/BAMS-D-20-0234.1.xml.](https://journals.ametsoc.org/view/journals/bams/103/5/BAMS-D-20-0234.1.xml)
- [4] [https://www.researchgate.net/figure/A-biological-neuron-in-comparison-to-an-artificial-neural-network-a-human-neuron-b_fig2_339446790.](https://www.researchgate.net/figure/A-biological-neuron-in-comparison-to-an-artificial-neural-network-a-human-neuron-b_fig2_339446790)
- [5] [https://www.nationalgeographic.com/photography/article/digitally-manipulated-ai-altered-photo-images.](https://www.nationalgeographic.com/photography/article/digitally-manipulated-ai-altered-photo-images)

Bibliography II

- [6] [https://www.historyofinformation.com/detail.php?id=4792.](https://www.historyofinformation.com/detail.php?id=4792)
- [7] [https://www.reuters.com/article/idUSKCN24Z2B1/.](https://www.reuters.com/article/idUSKCN24Z2B1/)
- [8] [https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.](https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402)
- [9] [https://www.datacamp.com/blog/classification-machine-learning.](https://www.datacamp.com/blog/classification-machine-learning)
- [10] [https://developers.google.com/machine-learning/gan/generative?hl=it.](https://developers.google.com/machine-learning/gan/generative?hl=it)
- [11] [https://www.smartius.it/data-it-law/reato-usare-deepfake/.](https://www.smartius.it/data-it-law/reato-usare-deepfake/)

Bibliography III

- [12] https://www.ansa.it/sito/notizie/tecnologia/internet_social/2019/06/28/garante-privacy-multa-1-mln-a-facebook-per-cambridge-analytica_ec9234f5-caae-4485-9beb-e092137a4d67.html.
- [13] <https://www.bbc.com/news/business-43983958>.
- [14] <https://finanza.lastampa.it/News/2022/12/23/meta-paga-725-milioni-di-dollari-per-risolvere-class-action-cambridge-analytica/NzVfMjAyMi0xMi0yM19UTEI>.
- [15] https://www.ansa.it/sito/notizie/tecnologia/internet_social/2019/06/28/garante-privacy-multa-1-mln-a-facebook-per-cambridge-analytica_ec9234f5-caae-4485-9beb-e092137a4d67.html.
- [16] <https://www.statista.com/statistics/268604/annual-revenue-of-facebook/>.

Bibliography IV

- [17] Daniel Cavalcanti Jeronymo, Yuri Cassio Campbell Borges e Leandro dos Santos Coelho. "Image forgery detection by semi-automatic wavelet soft-thresholding with error level analysis". In: *Expert Systems with Applications* 85 (2017), pp. 348–356.
- [18] Momina Masood et al. "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward". In: *Applied intelligence* 53.4 (2023), pp. 3974–4026.
- [19] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: <http://probml.github.io/book2>.
- [20] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

Bibliography V

- [21] Anh Nguyen et al. "Plug & play generative networks: Conditional iterative generation of images in latent space". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4467–4477.
- [22] Rimsha Rafique et al. "Deep fake detection and classification using error-level analysis and deep learning". In: *Scientific Reports* 13.1 (2023), p. 7422.