UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Diritto Pubblico, Internazionale e Comunitario (DiPIC)

Dipartimento di Ingegneria dell'Informazione (DEI)

Corso di Laurea Triennale in

DIRITTO E TECNOLOGIA

ASPETTI LEGALI RIGUARDANTI L'INTELLIGENZA ARTIFICIALE GENERATIVA MULTIMEDIALE

Relatore: Prof. Milani Simone

Laureanda: Candotto Elena

Matricola: 2010483

Anno accademico 2022-2023

A mia nonna Aurora

La mia più grande fonte di ispirazione

Abstract

Negli ultimi anni, grazie allo sviluppo degli algoritmi di *deep learning*, l'intelligenza artificiale generativa consente a chiunque di creare contenuti multimediali falsi altamente realistici, come immagini, video e audio.

L'ampia diffusione di questi contenuti suscita numerose preoccupazioni, poiché sono spesso utilizzati per fini malevoli. Pertanto, è di fondamentale importanza comprendere appieno queste tecnologie, quali sono gli ambiti di applicazione e quali implicazioni legali ne derivano, fornendo delle possibili tutele legali sia sul piano nazionale che sul piano europeo.

Inoltre, data la crescente difficoltà nel distinguere ciò che è falso da ciò che è reale, diventa essenziale analizzare l'efficacia degli algoritmi di rilevamento e le capacità di rilevazione umana, in particolare per quanto riguarda i deepfake audio, per contrastare gli abusi derivanti da queste tecnologie e preservare l'integrità delle comunicazioni.

Indice

Lista delle figure

Introduzione

L'intelligenza artificiale (AI) è una disciplina che si occupa di sviluppare sistemi e algoritmi che sono in grado di simulare le capacità cognitive umane, quali il ragionamento, l'apprendimento, la pianificazione e la creatività.

Lo sviluppo dell'automazione ha portato all'emergere due teorie fondamentali: AI forte e AI debole. L'intelligenza artificiale forte, anche conosciuta come intelligenza artificiale generale (AGI), mira ad emulare completamente le capacità cognitive umane, con l'obiettivo di creare sistemi esperti che pensano, comprendono e agiscono come l'essere umano. L'intelligenza artificiale debole, d'altro canto, è progettata per svolgere attività specifiche e compiti limitati, analizzando i dati e prendendo decisioni, spesso con una precisione maggiore a quella umana.

Mentre l'AI forte è ancora un obiettivo lontano dalla realizzazione ed è attualmente focalizzata sulla ricerca teorica, in quanto nessun sistema è stato in grado di passare completamente il test di Turing[1], l'AI debole è ampiamente utilizzata e offre molte applicazioni pratiche. Tra queste, vi è la creazione di nuovi dati sintetici con proprietà simili a quelle dei dati reali, facendo uso delle tecniche generative.

[1] Il test di Turing, sviluppato nel 1950 da Alan Turing, è un esperimento che permette di verificare se la macchina sa esibire un comportamento intelligente che sia indistinguibile da quello umano.

Capitolo 1

Tecniche generative

Le tecniche generative sono un insieme di modelli utilizzati nell'ambito dell'intelligenza artificiale che, tramite l'impiego di algoritmi di *machine learning*[2], sono progettate per apprendere la struttura sottostante di un determinato set di dati e creare nuovi dati sintetici[3] con proprietà simili a quelle dei dati reali.

A differenza delle tecniche discriminative, utilizzate per la classificazione dei dati e il riconoscimento di *pattern*, le tecniche generative sono in grado di generare dati mai visti prima o modellare la distribuzione di probabilità[4] dei dati complessi. Considerando una rete che genera immagini sintetiche con un contenuto semantico specifico, data una classe Y (es. gatto) e un vettore di rumore pseudo-casuale ξ, la rete genera un'immagine X effettivamente riconducibile alla classe Y, modellando la distribuzione di probabilità $P(X|Y)$, ovvero la distribuzione di tutte le immagini che rappresentano Y (un gatto, seguendo l'esempio). Nel caso in cui il modello sia a classe singola, per generare l'output di un solo tipo si riduce a modellare $P(X)$.

Le tecniche generative mirano alla massimizzazione della *likelihood*, ovvero la massimizzazione della probabilità di generare dati che seguono una distribuzione il più possibile vicina a quella del *dataset* di training.

Nel panorama attuale, i modelli generativi possono essere classificati in macrocategorie in base alla loro struttura, alla metodologia di apprendimento e

[2] Il termine *machine learning*, ovvero apprendimento automatico, è stato coniato per la prima volta nel 1959 da Arthur Lee Samuel, pioniere nel campo dell'intelligenza artificiale, però ad oggi la definizione più accettata dalla comunità scientifica è quella di Tom Mitchell, direttore di dipartimento della Machine Learning presso la Carnegie Mellon University, il quale definisce apprendimento automatico un qualsiasi programma che apprende una classe di *task* T dall'esperienza E, utilizzando una misura della performance P se l'esperienza E migliora le prestazioni dell'attività T misurate rispetto a P

[3] Dati creati artificialmente a partire dai dati reali tramite l'impiego dell'intelligenza artificiale

[4] Modello matematico che associa una probabilità ad ogni modalità osservabile di una variabile aleatoria

all'uso a cui sono destinati: i Variational AutoEncoders, le Reti Neurali Generative, tra cui le Generative Adversarial Networks, i modelli di diffusione, ed altri modelli.

1.1 Variational AutoEncoders

Una delle prime tecniche generative ad essere introdotte sono i Variational AutoEncoders. Per comprendere al meglio questi modelli è opportuno introdurre il concetto di AutoEncoder.

1.1.1 AutoEncoders

Gli AutoEncoders (AEs) sono una particolare tecnica di *deep learning*[5] che è progettata per codificare i dati di input in una rappresentazione compressa e significativa, quindi decodificarli in modo tale che l'input ricostruito sia il più possibile simile a quello originale. Sono composti da due reti neurali artificiali[6], l'encoder e il decoder, e dallo spazio latente z [Fig. 1.1]. L'encoder comprime i dati ricevuti in input in una rappresentazione codificata, più piccola dei dati in input, il decoder viene utilizzato per ricostruire i dati decomprimendo le rappresentazioni, e il collo di bottiglia, o spazio latente, contiene le rappresentazioni compresse.

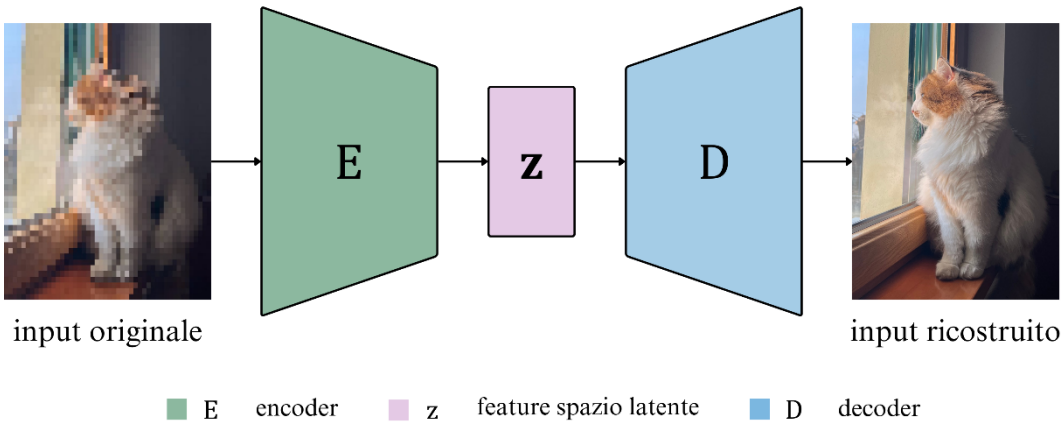

Figura 1.1: Architettura dell'Autoencoder

[5] Il *deep learning*, ovvero apprendimento profondo, è una sottocategoria dell'apprendimento automatico che, utilizzando algoritmi basati su reti neurali artificiali composte da molteplici strati di unità di elaborazione, modella dati altamente complessi per generare dati di *input* mai visti prima

[6] Le reti neurali artificiali, o simulate, sono un modello computazionale di *deep learning* ispirato al funzionamento dei neuroni biologici umani. Una rete neurale di base ha neuroni interconnessi su tre livelli: il *layer* di *input*, molteplici *layers* "nascosti" e il *layer* di *output*

L'obiettivo principale degli AutoEncoders (Bank et al., 2021) è quello apprendere in modo non supervisionato[7] una rappresentazione informativa dei dati che può essere utilizzata per varie implicazioni, come il *clustering*.

1.1.2 Architettura e funzionamento

I Variational AutoEncoders (VAEs), introdotti nel dicembre 2013 da Diederik P. Kingma e Max Welling (Kingma & Welling, 2022), sono modelli a variabili latenti continue che impiegano due reti neurali artificiali, l'encoder e il decoder, per apprendere una distribuzione $P'(X)$ che approssimi la distribuzione $P(X)$ di un certo *dataset*. L'encoder $P(Y|X)$ riceve un dato in input, lo codifica e lo comprime, generando un vettore delle medie e un vettore delle varianze. Successivamente, tramite il processo di *sampling,* viene generato un vettore z, che mappa spazio latente. Il decoder $(X|Y)$ decodifica e ricostruisce il dato in modo che sia il più possibile simile al dato ricevuto in input dall'encoder [Fig. 1.2].

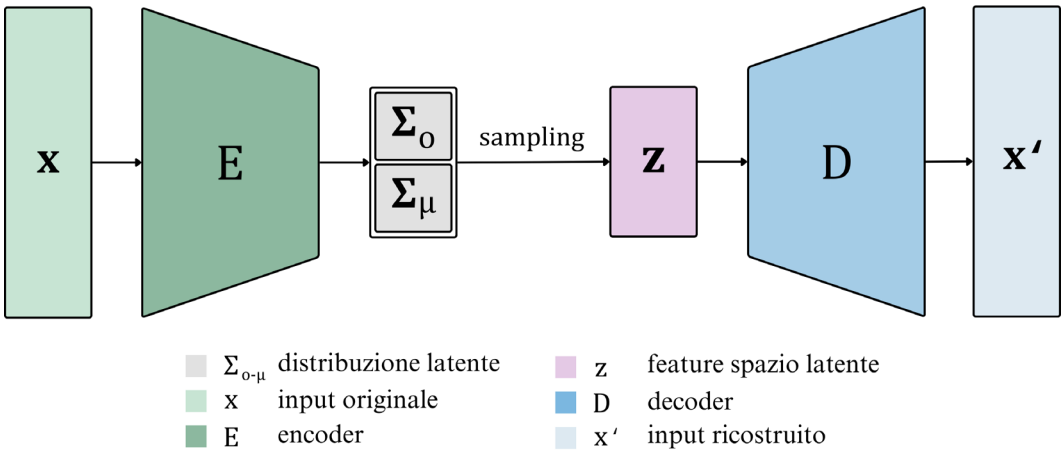

Figura 1.2: Architettura del Variational AutoEncoder (VAE)

A differenza degli AutoEncoders, dove lo spazio latente è sparso, i Variational AutoEncoders presentano parametri ben definiti per ogni input ricevuto, in modo tale che la distribuzione sia normale e la rappresentazione dello spazio latente sia continua e uniforme.

[7] L'apprendimento non supervisionato si riferisce a modelli in cui viene fornito al sistema un solo *dataset,* senza specificare il risultato da considerare, con lo scopo di apprendere la struttura nascosta o sottostante dei dati senza fare l'uso di etichette, a differenza dell'apprendimento supervisionato, il quale etichetta i dati per l'addestramento della rete

1.2 Generative Adversarial Networks

Le Generative Adversarial Networks (GANs), proposte per la prima volta nel 2014 da Ian Goodfellow e i suoi colleghi presso l'Università di Montréal (Goodfellow et al., 2014), sono una tecnica di *deep learning* che ha l'obiettivo di generare dati sintetici indistinguibili da quelli reali, tramite l'utilizzo di due reti neurali artificiali antagoniste che competono tra di loro: la rete generativa (o generatore) e la rete discriminativa (o discriminatore). La rete generativa cerca di generare dati sempre più convincenti e realistici per ingannare la rete discriminativa, mentre la rete discriminativa cerca di distinguere sempre più accuratamente i dati reali da quelli sintetici prodotti dalla rete generativa.

1.2.1 Architettura

Il generatore, o decoder, accetta in input un vettore di rumore pseudo casuale e produce una nuova istanza di dati, i quali assomigliano il più possibile ai dati di training, mentre il discriminatore, o encoder, cerca di distinguere come falsi i dati sintetici generati dal generatore e come veri i dati reali presenti nel *dataset* di training [Fig. 1.2].

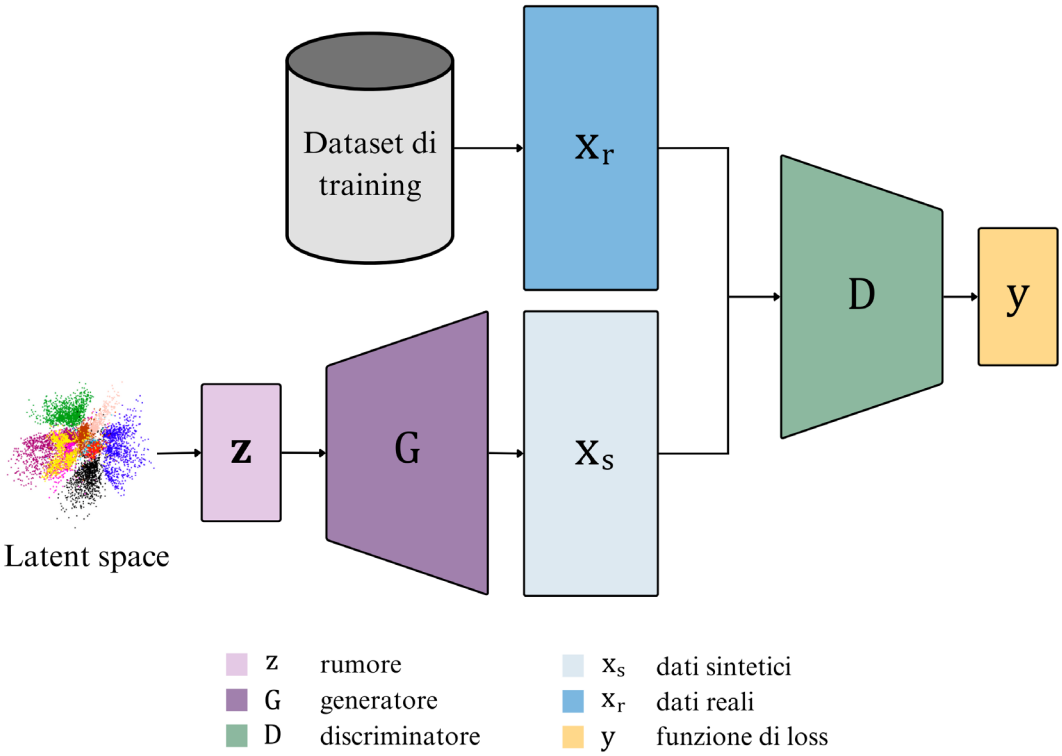

Figura 1.2: Architettura di una Generative Adversarial Network (GAN)

Analizzare l'architettura del generatore e del discriminatore è di fondamentale importanza per comprendere al meglio questo modello e come esso influisce sulle prestazioni. Il numero e il tipo di strati di neuroni artificiali e le funzioni di attivazione[8] utilizzate possono influenzare la capacità del modello di apprendere rappresentazioni significative dei dati di input e generare output di alta qualità.

1.2.2 Convolutional Neural Networks

La maggior parte delle GAN oggi si basano su un approccio standardizzato chiamato Deep Convolutional GAN (DCGAN), formalizzato nel 2015 da Alex Radford nel documento "*Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*".

Nelle DCGANs, il generatore e il discriminatore sono due Convolutional Neural Network (CNN), o reti neurali convoluzionali. Le CNN sono un tipo di algoritmo di *deep learning* che, a differenza delle reti neurali tradizionali, ha prestazioni elevate con input quali immagini, voce o audio. Una CNN è composta da tre principali *layer*: il *convolutional layer*, dove avviene la maggior parte del calcolo, il *pooling layer*, il quale consente la riduzione delle dimensioni a un quarto delle originali e il *full- connected layer*, ovvero ogni nodo del livello di output si connette direttamente a un modo del livello precedente.

Una caratteristica importante di queste reti è quella che presentano neuroni nascosti con tutti i *bias* e pesi condivisi uguali.

1.2.3 Addestramento delle reti

La prima fase dell'addestramento riguarda la preparazione del *dataset*: i dati devono essere normalizzati[9] e i valori devono essere nella stessa scala, in modo tale da poterli utilizzare sia per addestrare il generatore che il discriminatore.

Le due reti, addestrate in modo iterativo, vengono poste in competizione in un gioco a somma zero, con funzioni-obiettivo opposte. Il discriminatore è un classificatore

[8] Una funzione di attivazione determina quali neuroni devono essere attivati

[9] Eliminare la ridondanza informativa e le dipendenze incoerenti tramite la creazione di tabelle e la definizione di relazioni tra di esse

binario[10] che analizza i pixel[11] dell'immagine in input, estrae i *pattern*, li confronta con i modelli appresi, assegna un'etichetta e restituisce un valore y tra 0 e 1: 0 se l'immagine è riconosciuta come falsa, 1 se è riconosciuta come vera. Il generatore viene addestrato utilizzando l'algoritmo di *back propagation* tramite l'utilizzo dei *feedback* ricevuti dal discriminatore, il quale indica se i dati sintetici prodotti dal generatore sono più o meno realistici.

L'alternanza continua tra addestramento del generatore e del discriminatore si protrae finché non viene raggiunto l'equilibrio di Nash, ovvero quando una delle due reti non cambia la propria azione indipendentemente da ciò che può fare l'altra rete. Una volta terminato l'addestramento il modello viene valutato utilizzando dati di test per verificare la sua capacità di generare dati sintetici realistici.

L'addestramento di una GAN può richiedere molto tempo e risorse di calcolo, soprattutto se il *dataset* è molto grande o complesso. Inoltre, è possibile incontrare diverse sfide durante l'addestramento, come il collasso del generatore (la produzione di una varietà limitata di *sample*) o l'*overfitting*[12] del discriminatore, i quali richiedono una soluzione di problemi adeguata.

1.2.4 Applicazioni

Le GANs trovano applicazione in molte aree: la generazione di immagini, la sintesi di suoni e la creazione di testo. È possibile dividere queste applicazioni nelle seguenti aree: generare esempi per *dataset* di immagini, generare fotografie di volti umani, generare fotografie realistiche, generare personaggi dei cartoni animati, traduzione da immagine a immagine, traduzione da testo a immagine, traduzione semantica da immagine a foto, generazione della vista frontale del viso, generazione nuove pose umane, traduzione da foto a Emoji, modifica fotografica,

[10] Algoritmo di apprendimento supervisionato che classifica le nuove osservazioni in due classi

[11] Il pixel è la più piccola unità rappresentabile sullo schermo del computer

[12] Non è presente una regola generalizzata poiché non ci sono abbastanza dati. L'*overfitting* è il sovra-adattamento del modello statistico al campione di dati osservato, ovvero il modello si adatta troppo strettamente al *dataset* di training

invecchiamento del viso, fusione di foto, super risoluzione, fotoritocco, traduzione di abbigliamento, previsione video, generazione di oggetti 3D.

1.3 Diffusion models

Tra i modelli generativi, i modelli di diffusione sono una famiglia di modelli generativi probabilistici che utilizzano funzioni di base per imparare a modellare la distribuzione dei dati di input. Un modello di diffusione è un modello a variabile latente che mappa lo spazio latente facendo uso di una catena di Markov[13]. Tale catena rappresenta una serie di operazioni di correzione o ricostruzione dell'immagine che è stata progressivamente corrotta da livelli incrementali di rumore. In particolare, distrugge i dati di training aggiungendo gradualmente rumore gaussiano[14] per imparare ad invertire il processo di *denoising* per la generazione del campione.

1.3.1 Processo di denoising

Il modello di diffusione (Croitoru et al., 2022) si suddivide in due fasi: il processo di diffusione, o *fixed forward diffusion process,* e il processo di diffusione inverso, o *generative reverse denoising process*. Nel processo di diffusione i modelli prendono il dato in input (generalmente un'immagine) e aggiungono gradualmente rumore gaussiano attraverso una serie di *T steps* utilizzando una catena di Markov fissa, mentre nel processo di diffusione inverso il rumore viene ritrasformato in un campione della distribuzione *target* [Fig. 1.3].

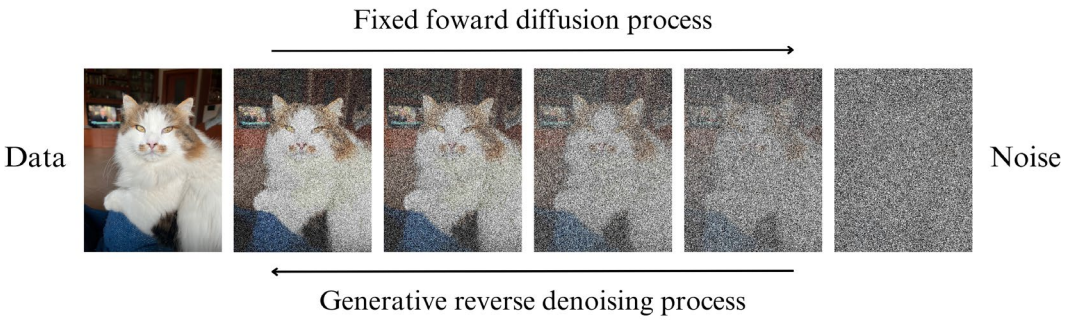

Figura 1.3: Processo di *denoising*

[13] Modello stocastico che descrive una sequenza di possibili eventi in cui la probabilità di ciascun evento dipende solo dallo stato raggiunto nell'evento precedente

[14] Rumore che ha una funzione di densità di probabilità uguale a quella della distribuzione normale

Dopo l'addestramento, ovvero quando l'immagine viene trasformata in modo asintotico[15] in puro rumore gaussiano, possiamo utilizzare il modello di diffusione per la generazione di dati sintetici tramite il rumore campionato in modo casuale attraverso il processo di *denoising*.

La ricerca attuale sui modelli di diffusione si basa principalmente su tre formulazioni predominanti: modelli probabilistici di diffusione *denoising*, modelli generativi basati su punteggio ed equazioni differenziali stocastiche.

Uno dei vantaggi dei modelli diffusi rispetto ad altri modelli generativi è che possono essere più stabili e meno soggetti a problemi come il collasso del modello, in cui il modello produce solo pochi esempi diversi. Inoltre, essi possono essere addestrati su un insieme di dati di bassa qualità e utilizzati per generare dati di alta qualità, come ad esempio la modellizzazione di immagini ad alta risoluzione.

1.4 Implicit Neural Representation

L'Implicit Neural Representation (INR), o rappresentazione neurale implicita, è un approccio utilizzato nella *Computer Graphics*[16] che consente la generazione e la manipolazione di immagini e forme tridimensionali, come la compressione, la ricostruzione di scene 3D da immagini 2D e l'inferenza di informazioni semantiche.

Originariamente proposta nel 2007 (Kenneth, 2007), la rappresentazione neurale implicita viene utilizzata per codificare un segnale *target* continuo attraverso l'utilizzo di una rete neurale artificiale. La rete, parametrizzata da un insieme di pesi ed addestrata su campioni rappresentati in modo discreto[17] dello stesso segnale, rappresenta una mappatura delle coordinate spaziali di input (posizione, direzione, profondità).

Una rete di parametrizzazione di pixel coordinati, o Coordinated Pixel Network (CPN), è un tipo di rete neurale per la generazione di immagini che implementa una

[15] Tende ad avvicinarsi sempre più, senza raggiungerlo

[16] Il termine Computer Graphics, coniato nel 1960 dai ricercatori Hudson e Fetter, faceva riferimento a "quasi tutte le cose sui computer che non fossero testo o suono"

[17] Le grandezze variano in modo discontinuo, ovvero passano da un valore ad un altro senza assumere valori intermedi

rappresentazione neurale implicita per generare una funzione di parametrizzazione che assegna un vettore di coordinate ad ogni pixel dell'immagine, ad esempio da pixel a RGB[18]. [Fig. 1.5]

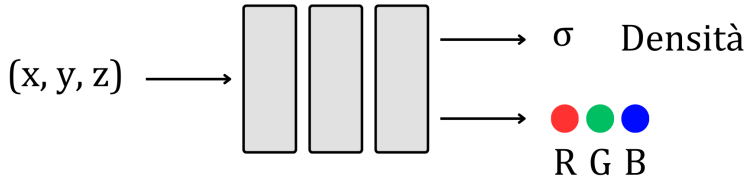

Figura 1.5: Rete di parametrizzazione da pixel a RGB

[18] Ogni pixel di un'immagine digitale è rappresentato da tre colori: R red, G green e B blue, i quali permettono di ottenere un'ampia gamma di colori visibili

Capitolo 2

Applicazioni delle tecniche generative

Le applicazioni delle tecniche generative spiegate nel capitolo precedente sono molteplici. In questo capitolo ci soffermeremo in particolar modo sui deepfake, la *data augmentation* e gli aspetti creativi di queste tecnologie.

2.1 Deepfake

Il termine deepfake, neologismo nato dall'incrocio tra la locuzione *deep learning* e *fake* (falso), si riferisce a contenuti multimediali, quali immagini, video, audio e testo, falsi che sono generati o manipolati utilizzando algoritmi di *deep learning*; l'intento è quello di indurre colui che li osserva a percepirli come una rappresentazione fedele della realtà.

Nascono convenzionalmente alla fine del 2017, quando un anonimo gruppo con lo pseudonimo "*Deepfakes*" pubblica i primi video falsi di natura pornografica sul popolare sito di aggregazione Reddit[19], facendo uso dell'approccio *face-swap* implementato con l'applicazione FaceApp[20]. Dopo questo evento, nonostante la circolazione di deepfake pornografici sia stata proibita e bloccata tramite la rimozione degli stessi sulle piattaforme social, la creazione e divulgazione di deepfake è diventata inarrestabile.

I deepfake vengono impiegati in diversi settori, come il cinema, la medicina, l'arte, le comunicazioni digitali, l'intrattenimento e per scopi commerciali (e-commerce e moda). Ma, tutt'ora, oltre il 90% dei deepfake presenti in rete è di natura

[19] Reddit, fondata nel 2005 da Steve Huffman, Aaron Swartz e Alexis Ohaian, è una piattaforma di *social news* nella quale è possibile condividere contenuti multimediali, messaggi e *hyperlinks* (links)

[20] Applicazione per dispositivi mobili sviluppata dalla società russa Wireless Lab che permette di generare automaticamente delle trasformazioni altamente realistiche del volto delle persone nelle foto. Lanciata nel 2017, ha avuto grande successo nel 2019 quando è stato aggiunto il filtro per l'invecchiamento del volto e alcune celebrità hanno introdotto l'hashtag #FaceAppChallenge, da utilizzare quando si pubblicava una foto del volto invecchiato realizzato tramite l'app

pornografica e una buona parte rimanente riguarda la creazione di deepfake per la diffusione di fake news.

2.1.1 Deepfake visivi

La contraffazione di dati sintetici è ben antecedente alla nascita dei deepfake. La *Computer Graphics* comprende un'area molto ampia di tecniche e strumenti informatici che vengono utilizzati per la generazione o la manipolazione di immagini e video digitali 2D e 3D.

Le tecniche più comuni e utilizzate per la manipolazione di contenuti visivi sono l'aggiunta o la replica di un oggetto, tramite l'inserimento di un nuovo oggetto copiato dalla stessa immagine (*copy – move*) o da un'immagine diversa (*splicing*), e la rimozione di un oggetto, tramite l'estensione dello sfondo (*inpainting*). Inoltre, è possibile effettuare modifiche post-elaborazione, quali il ridimensionamento, la rotazione e la regolamentazione del colore (Verdoliva, 2020).

La manipolazione delle immagini digitali è un mezzo molto efficacie per la diffusione di fake news. Nel 2008, il governo iraniano è stato accusato di aver falsificato una foto dei test missilistici. La foto manipolata con la tecnica *copy-move* è stata pubblicata sul sito web ufficiale delle Guardie Rivoluzionarie Iraniane affermando che quattro missili si stavano dirigendo verso il cielo, quando in realtà ne erano stati lanciati presumibilmente tre (anche se ci sono ulteriori dubbi sul loro numero effettivo). [fig. 2.1]

(a) foto originale (b) foto manipolata

Figura 2.1: Manipolazione di un'immagine tramite le tecniche *copy-move*

Se la manipolazione di immagini è un mezzo efficacie per la diffusione di fake news, la manipolazione dei volti è un mezzo ancora più efficacie e pericoloso, in quanto i volti giocano un ruolo centrale nella comunicazione umana.

I metodi di manipolazione del volto (Busch et al., 2022 & Akhtar, 2023) possono essere suddivisi in quattro categorie principali: (a) *entire face syntesis*, (b) *identity swap*, (c) *face reenactment* e (d) *attribute manipulation.*

(a) L'*entire face syntesis*, ovvero la sintesi dell'intero volto, è una manipolazione che permette di creare intere immagini di volti non-esistenti. Queste tecniche raggiungono risultati sorprendenti, generando immagini facciali di alta qualità con un alto livello di realismo per la persona che le osserva. [Fig. 2.2]

Figura 2.2: Immagini ritraenti persone che non esistono realizzate tramite il sito www.this-person-does-not-exist.com

(b) L'*identity swap*, o *face swap*, è una manipolazione che consiste nel sostituire il volto di una persona (*target*) con il volto di un'altra persona (*source*), con l'obiettivo di generare immagini e video falsi realistici. Alcune delle applicazioni che permettono di generare deepfake con la tecnologia face-swap sono DeepSwap e Faceover. [Fig. 2.3]

Figura 2.3: Immagini manipolate con la tecnica della face-swap (Peng et al., 2022 & Zhu et al., 2021)

(c) La *face reenactment*, o *expression swap,* è una manipolazione che consente di modificare l'espressione del soggetto, sostituendo l'espressione facciale di una persona (*source*) con l'espressione facciale di un'altra persona (*target*), Questa tecnica è utilizzata anche per la generazione di video deepfake e solitamente viene combinata a tecniche di deepfake audio. Di seguito alcuni esempi [Fig. 2.4].

Figura 2.4: Estratti di immagini manipolate con la tecnica *della face reenactment*

(d) L'*attribute manipulation,* nota anche come *face editing* o *face retouching*, è una manipolazione che consiste nel modificare alcuni attributi del viso come il colore dei capelli, della pelle, il genere, l'età, l'aggiunta di occhiali, ecc. Alcuni esempi [Fig. 2.5] realizzati con l'applicazione per dispositivi mobili chiamata FaceApp.

(a) modifica dell'età (b) modifica del genere

Figura 2.5: Immagini manipolate con la tecnica del *face editing*. (a) a sinistra la foto originale ritraente il cantante Harry Styles, a destra la foto modificata con l'effetto invecchiamento. (b) a sinistra la foto originale ritraente l'attrice Angelina Jolie, a destra la foto modificata con l'effetto di cambio del genere

2.1.2 Deepfake audio

Come per la contraffazione di immagini e video, anche quella degli audio si è sviluppata antecedente ai deepfake e alle tecniche di *deep learning*. I primi strumenti utilizzati per la manipolazione dell'audio vocale risalgono alla prima metà del Novecento, quali talk-box e vocoder, ma solamente nel 1997 viene creato il primo *software* che lavora in maniera totalmente autonoma: l'Auto-Tune[21].

I metodi di manipolazione degli audio sono suddivisi in due categorie: *container based* e *content based* (Benvinamarad & Shirldonkar, 2020). Viene definita *container based* la manipolazione della struttura del file audio, dei metadati e della relativa descrizione, mentre *content based* la manipolazione del contenuto.

I deepfake audio, chiamati anche *skin* vocali, consistono nella generazione di frasi vocali che suonano come pronunciate da una determinata persona, associata alla skin vocale selezionata. Tali tracce audio risultano altamente convincenti in quanto spesso la voce sintetica non è distinguibile dalla voce reale. Questo avviene grazie alla replica accurata delle caratteristiche della voce, quali la tonalità, il timbro, l'estensione, gli accenti, le cadenze ed anche lo stato emotivo.

Nel processo di generazione della voce sintetica viene utilizzata la tecnologia Text-To-Speech[22] (TTS), ovvero un *software* che riceve in input un testo e produce in output un discorso udibile. Il testo ricevuto in input viene diviso in più parti e rappresentato in fonemi[23]. In passato il *software* TTS permetteva di creare voci dal suono "naturale" senza cercare di replicare la voce di una persona specifica, ma con

[21] Software creato nel 1997 da Antares Audio Technologies che analizza la forma d'onda del segnale registrato in *input* e apporta automaticamente le correzioni necessarie impostate dall'utente in precedenza. Le manipolazioni audio che possono essere effettuate sono la correzione delle imperfezioni della voce (note stonate o fuori tonalità) e la creazione di effetti sonori unici (*pitch shifting*, *doubling* e *chorusing*)

[22] Uno dei primi sistemi di sintesi vocale Text-To-Speech (da testo a voce) viene realizzato negli Stati Uniti nel 1968 presso il laboratorio di intelligenza artificiale del MIT. Il sistema, chiamato "Pattern Playback", faceva uso di una registrazione audio digitale di una voce umana che veniva in un secondo momento manipolata per creare suoni vocali

[23] Unità logica alla base della produzione di un suono. Ad esempio, pésca e pèsca hanno fonemi caratterizzati dall'accento diverso

il passare degli anni la tecnologia si è sviluppata e ora viene utilizzato per la clonazione di voci realistiche.

All'inizio del 2023, Microsoft ha presentato VALL-E, acronimo di *Voice Agonistic Lifelike Language*, un nuovo sistema di intelligenza artificiale Text-To-Speech in grado di generare discorsi altamente realistici e simili a quelli umani (Wang et al., 2023). Il modello prende in input la registrazione della voce di una persona di appena 3 secondi e il testo da utilizzare per la sintesi e, tramite un sistema di riconoscimento del linguaggio naturale, replica le caratteristiche della voce, trasformando il testo in parlato.

A differenza delle precedenti *pipeline* che prevedevano la generazione del fonema, seguita dallo spettrogramma di M[24] e dalla creazione della forma d'onda, la *pipeline* di VALL-E genera un vettore di features tramite una rete neurale al posto dello spettrogramma di M.

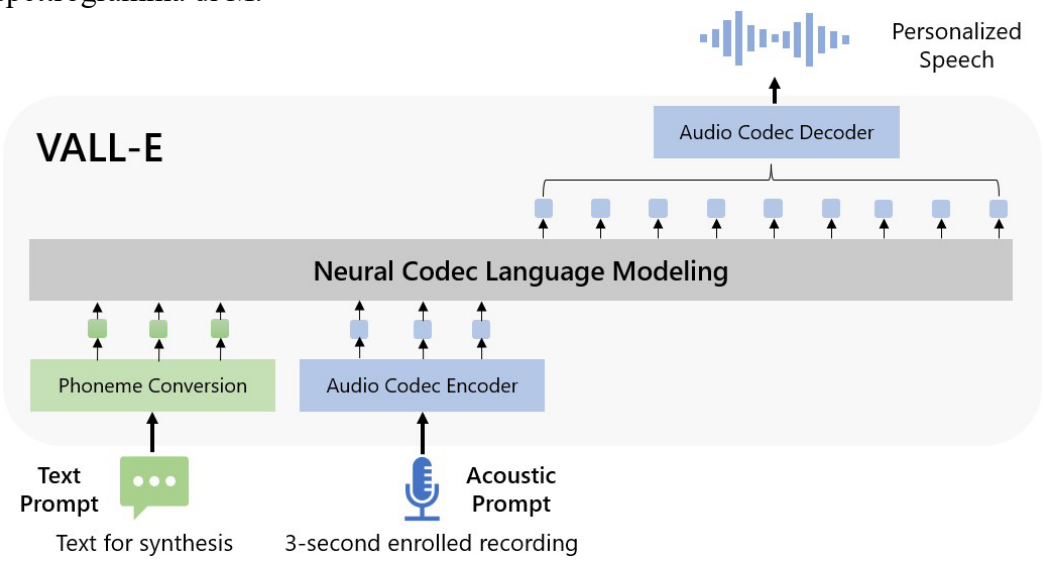

Figura 2.6: Funzionamento di VALL-E

A pochi mesi dalla presentazione di VALL-E, Microsoft ha introdotto un altro sistema, chiamato VALL-E X (*Cross Lingual* VALL-E), che utilizza il riconoscimento del linguaggio naturale per sintetizzare un discorso parlato in una lingua diversa da quella originale. VALL-E X eredita grandi capacità di

[24] Scala di percezione dell'altezza (*pitch*) di un suono. A differenza delle scale lineari delle frequenze, la scala di Mel è costruita in modo tale da aderire di più alla percezione umana del suono

apprendimento in contesto dal suo predecessore e può essere applicato sia per le attività di traduzione vocale che per la sintesi vocale interlinguistica *zero-shot* (Zhang et al., 2023).

2.1.3 Digital forgery detection

La *multimedia forensics* è un'area emergente della *digital forensics*[25] che si occupa di analizzare i contenuti digitali multimediali, quali foto, video e audio, al fine di produrre prove nel dominio forense. In particolare, la *digital forgery detection,* o il rilevamento di manipolazioni digitali, consente di verificare se un contenuto rispetta i criteri di autenticità e integrità; dunque, se è stato manipolato o meno.

Le tecniche di rilevamento delle manipolazioni di immagini (Battiato et al., 2016) sono suddivise in due principali categorie: le metodologie attive, quali firme digitali e *watermarking*[26] digitali, consentono di inserire o allegare informazioni aggiuntive all'immagine stessa, mentre quelle passive utilizzano le informazioni implicite dell'immagine.

A differenza delle firme digitali, facilmente invalidabili da qualsiasi modifica dei dati e quindi più adatte alla protezione del diritto d'autore, il *watermarking* è un sistema che in alcuni casi si rivela un po' più robusto e affidabile, in quanto è difficile rimuoverlo o alterare il contenuto dell'immagine senza modificarlo (Zanardelli et al., 2023).

I metodi passivi tradizionali, ovvero antecedenti all'era del *deep learning*, impiegano diverse tecniche di campi come la statistica, la fisica, la geometria e l'elaborazione di segnali ed hanno la caratteristica di non aver bisogno di *dataset* di training ampi (a volte non hanno neanche bisogno di *dataset* di training). Una delle molteplici tecniche per la rilevazione delle manipolazioni visive è quella del *pixel-based*, che prevede l'analisi dei pixel dell'immagine stessa. Mentre, i metodi passivi

[25] La *digital forensics*, o informatica forense, è un'area della scienza forense che si occupa dell'identificazione, acquisizione, analisi e conservazione delle prove digitali in modo tale che possano essere utilizzate nei procedimenti legali (ISO IEC 27037/2012)

[26] Il *watermarking* è un messaggio che viene aderito a un contenuto multimediale, può essere visibile o non visibile e non modifica la semantica del contenuto

che utilizzano le tecniche di *deep learning* possono essere suddivisi in base al tipo di contraffazione rilevata, le proprietà di localizzazione e il tipo di architettura. Come visto nella sezione 2.2.1, la maggior parte delle manipolazioni dei deepfake sono manipolazioni del volto di diverso tipo; quindi, è opportuno fare uso di metodi che utilizzano algoritmi di rilevamento del volto.

Una delle tecniche proposte per il rilevamento delle manipolazioni audio digitali (Zhao & Malik, 2013) è l'utilizzo della firma dell'ambiente acustico, ma, come per la firma digitale, questa tecnologia presenta poca robustezza. L'analisi dello spettrogramma, delle caratteristiche acustiche e dei *pattern*, anche tramite l'uso di algoritmi di *machine learning*, sono, ad oggi, i metodi più efficaci e utilizzati nel campo della *forgery detection* audio.

2.2 Data augmentation

La *data augmentation* è un insieme di tecniche che vengono utilizzate per aumentare artificialmente la quantità dei dati presenti nel *dataset* di training, senza effettivamente raccogliere nuovi dati, per ovviare al problema dell'*overfitting*. I nuovi dati prodotti possono essere di due tipi: gli *augmented data*, o dati aumentati, e i dati sintetici.

2.2.1 Tecniche per la manipolazione degli augmented data

Gli *augmented data* (Shorten & Khoshgoftaar, 2019) sono copie dei dati già presenti nel *dataset* di training a cui sono state apportate delle piccole modifiche tramite l'applicazione di cambiamenti casuali controllati, quali trasformazioni geometriche (ribaltamento, ridimensionamento, ritaglio e rotazione) e dello spazio-colore (luminosità, contrasto e canali RGB), filtraggio (nitidezza, sfocatura), cancellazione di una parte dell'immagine e *mix* di più immagini tra di loro, per quanto riguarda le immagini [Fig. 2.7]; aggiunta di rumore, *shifting*, modifica della velocità e del *pitch* (tono), per l'audio.

Figura 2.7: Modifiche dell'immagine originale tramite l'applicazione di cambiamenti controllati

2.2.1 Tecniche per la generazione dei dati sintetici

I dati sintetici, invece, sono dati generati artificialmente attraverso l'uso di due principali tecniche generative analizzate nel capitolo 1: le Generative Adversarial Networks e i Vatiational AutoEncoders.

Il framework TANDA [Fig. 2.8], acronimo di Transformation Adversarial Networks for Data Augmentations, è stato proposto nel 2017 per affrontare una delle sfide nel campo della data augmentation generativa. Basandosi sull'architettura delle GANs (Gatner et al., 2017), il generatore riceve in *input* dei dati originali ed una sequenza di trasformazioni desinate generando in output dei dati sintetici simili a quelli reali presenti nel *dataset* di training.

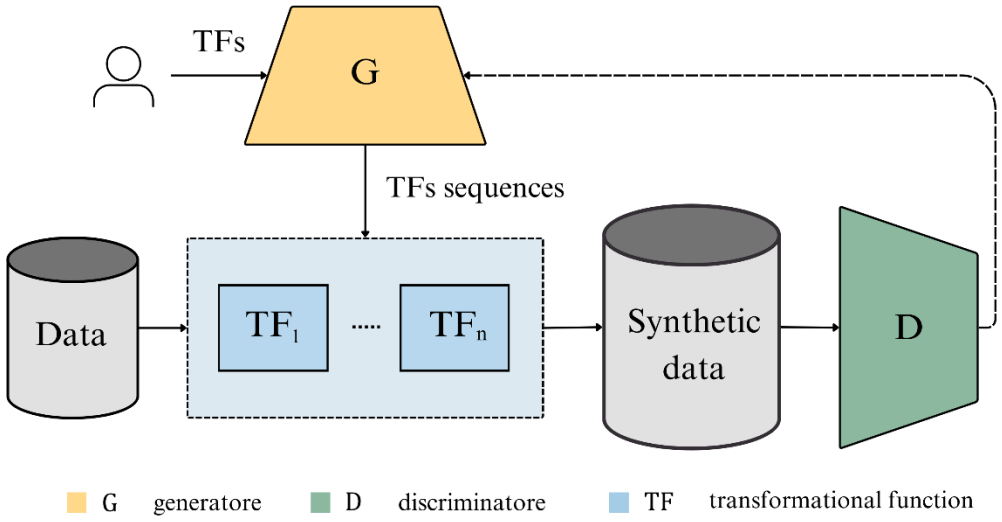

Figura 2.8: Processo di creazione dei dati sintetici automatizzato tramite TANDA

L'Augmented Analysis, termine coniato dalla società Gartmer nel 2017, è un approccio che automatizza gli *insight* utilizzando il *machine learning* e la comprensione ed elaborazione del *Natural Language Process*[27] per aumentare la Business Intelligence, la condivisione e l'analisi dei dati (Minu & Ahmad, 2020).

2.2.2. Applicazioni

La *data augmentation* viene utilizzata per obiettivi svariati e in diversi campi di applicazione: nel settore medico (dove l'acquisizione e l'etichettatura dei dati richiede molto tempo ed è costosa), nei sistemi di riconoscimento vocale (in quanto aggiungere rumore e distorsioni ad una registrazione audio permette di simulare il mondo reale), nel settore della guida autonoma (per il riconoscimento di oggetti, verificare le condizioni meteorologiche e la gestione di situazioni di emergenza) ed infine nel campo del *Natural Language Process*, (ad esempio per aumentare la diversità culturale o ampliare i campi semantici).

2.3 Aspetti creativi

Le tecniche generative vengono spesso utilizzate per la creazione di contenuti creativi, quali la pittura, l'architettura, la musica e il cinema, e i risultati ottenuti sono sempre più sorprendenti.

Nel 2018, è stata venduta all'asta da Christie's per 432.500 dollari, una stampa su tela realizzata interamente da una Generative Adversarial Network (GAN). L'opera, intitolata "ritratto di Edmond Belamy[28]", fa parte di un gruppo di ritratti della famiglia immaginaria Belamy, creato dal collettivo parigino Obvious[29] (McCormack, Gifford & Hutchings, 2019) [Fig. 2.9]. Per la realizzazione di queste tele la GAN è stata allenata utilizzando un *dataset* di training di 15 mila ritratti

[27] Il *Natural Language Processing* (NLP), o elaborazione del linguaggio naturale, comprende gli algoritmi di intelligenza artificiale che sono in grado di analizzare, rappresentare e comprendere il linguaggio naturale, ovvero il linguaggio che si è evoluto in modo naturale negli esseri umani. Il termine viene coniato nel 1950 da Alan Turing nel documento "*Computing Machinery and Intelligence*"

[28] Il nome deriva dall'interpretazione francese di "*Goodfellow*": Bel ami

[29] Obvious è un collettivo di ricercatori parigini che lavora con gli ultimi modelli di deep learning per esplorare le potenzialità creative dell'intelligenza artificiale. https://obvious-art.com

compresi tra il XIV e il XX secolo e, come in ogni dipinto, al termine di ogni opera viene apposta la firma, rappresentante l'equazione delle GAN:

$$\min G \ \max D \ \mathbb{E}_x [\log (D(x))] + \mathbb{E}_z [\log(1 - D(G(z)))]$$

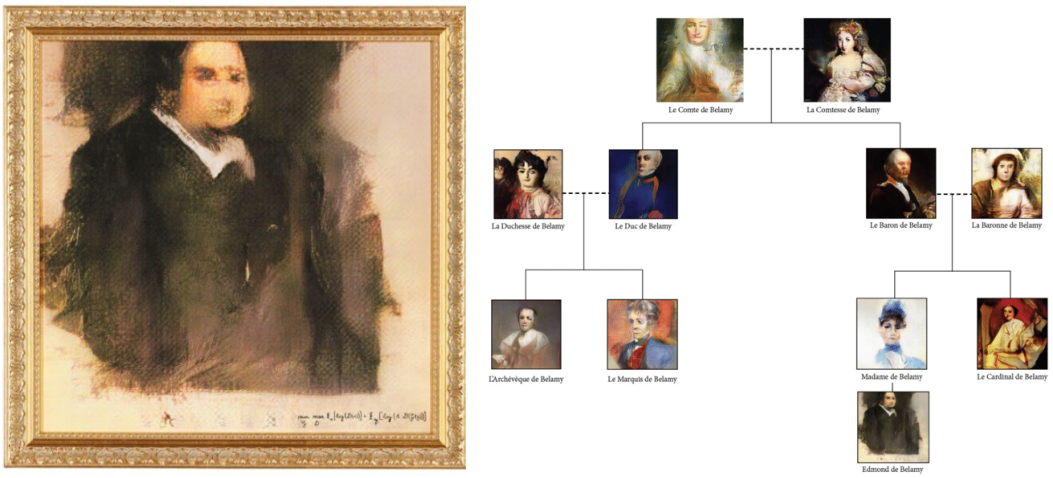

Figura 2.9: A sinistra il ritratto di Edmond Bellamy venduto all'asta da Christie's per 432 mila dollari, a destra tutte le opere della "Famille de Bellamy" realizzate con lo stesso *dataset* di training

Negli ultimi anni, con l'avvento del *deep learning,* i *software* per la generazione di contenuti creativi multimediali, quali foto, video, audio, testo e codice, sono stati affinati sempre di più e nel 2017 è stata introdotta da Google Research una nuova architettura chiamata "*transformer*" (Vaswani et al., 2017), che permette di generare modelli linguistici di alto livello. I *software* più utilizzati e conosciuti nel panorama attuale si suddividono in Text-to-Image, come DALL-E, Stable Diffusion e MidJourney, Text-to-Text, come ChatGPT, Text-to-3DModel, come Shape-E, e Text-to-Audio, come AIVA.

2.3.1 DALL-E

DALL-E (acronimo che riprende "Dalì", da Salvador Dalì, e "Eve", dal personaggio Eve del film Pixar WALL-E) è un *software* rilasciato da OpenAI[30] a gennaio 2021 che, dato in input un testo, genera immagini sintetiche originali e realistiche che rispettano il contenuto di una descrizione testuale fornita come *prompt*. Il modello è stato allenato su un ampio *dataset* di coppie di testo-immagine,

[30] OpenAI è un'organizzazione americana no-profit fondata da Elon Musk e Sam Altman nel dicembre 2015. Si occupa di promuovere e sviluppare una *friendly AI* (intelligenza artificiale amichevole) in modo che possa trarne beneficio tutta l'umanità

in modo tale da comprendere le relazioni tra oggetti e concetti diversi per la produzione di immagini coerenti. Ad esempio, se la descrizione testuale in input è "un gatto che gioca con una pallina blu in giardino" il *software* è in grado di comprendere la relazione tra il gatto, la pallina, il colore blu e il giardino.

2.3.2. Stable Diffusion

Stable Diffusion è un *software open source*[31] rilasciato da StabilityAI[32] ad agosto 2022 che permette di generare immagini a partire da una descrizione testuale. Questo *software* fa uso degli *Stable Diffusion Models* (SDMs), ovvero un modello generativo a diffusione latente, una variante dei modelli di diffusione (Capitolo 1.3), che, tramite l'aggiunta graduale di rumore gaussiano, genera l'immagine richiesta in input dall'utente.

2.3.3 MidJouney

Rilasciato a luglio 2022, MidJourney è un programma disponibile sulla piattaforma server Discord[33]. Come per DALL-E e Stable Diffusion, dato in input un testo, il *software* genera un'immagine che rispetti fedelmente il contenuto della descrizione, realizzando immagini realistiche.

2.3.4 ChatGPT

"ChatGPT è un modello di elaborazione del linguaggio naturale basato sull'architettura GPT (Generative Pre-trained Transformer) sviluppato da OpenAI.

ChatGPT utilizza un'architettura a Transformer, che gli consente di elaborare il testo di input e generare le risposte in modo sequenziale, una parola alla volta. generando risposte coerenti e fluide. Inoltre, il modello utilizza il concetto di pre-training, cioè l'addestramento su un vasto corpus di testo in modo non supervisionato prima di essere addestrato su un task specifico, per acquisire una

[31] Rende disponibile il suo codice sorgente. Il *software* è rilasciato con una licenza per cui chi detiene il copyright cede i diritti gratuitamente per eventuali modifiche

[32] StabilityAI è stata fondata nel 2019 da Emad Mostaque con lo scopo di risolvere alcune delle sfide dell'intelligenza artificiale, utilizzando tecnologie aperte ed accessibili a tutti

[33] https://www.midjourney.com

conoscenza del linguaggio naturale generale che può essere applicata a una vasta gamma di domande e input."

Testo realizzato interamente da ChatGPT[34] (descrizione testuale in input: spiega brevemente che cos'è ChatGPT)

2.3.5 Shap-E

Shap-E, rilasciato a maggio 2023 da OpenAI (Jun & Nichol, 2023), è un *software* disponibile gratuitamente su GitHub che permette la conversione di una parola in un oggetto tridimensionale.

2.3.6. AIVA

AIVA, acronimo di *Artificial Intelligence Virtual Artist*, è un *software* gratuito che, dati in input una serie di parametri, quali stato emotivo, epoca, stile e densità delle note, è in grado di comporre tracce musicali realistiche. Creato a febbraio 2014, AIVA è specializzato in musica classica e sinfonica ed è il primo artista virtuale riconosciuto dalla SACEM, l'associazione francese che rappresenta i diritti d'autore e dei compositori di musica originale.

[34] https://chat.openai.com

Capitolo 3

Aspetti legali delle tecnologie deepfake

Come visto nei capitoli precedenti, le tecniche generative permettono di realizzare contenuti multimediali falsi mai visti prima d'ora. I deepfake possono essere utilizzati per molti scopi benefici in settori come la sanità, l'istruzione e il cinema, ma sono state riscontrate numerose problematiche di tipo legale connesse all'utilizzo malevolo di queste tecnologie (Europol, 2022). In questo capitolo verrà, in primo luogo, analizzato il GDPR per dare una chiara definizione del c.d. dato personale e, successivamente, verranno presentati e analizzati i quattro aspetti legali più importanti nel panorama attuale: la "deepfake *pornography*", la diffusione di fake news, le truffe telefoniche e la questione della proprietà intellettuale legata alla *creative* AI. Infine, viene analizzato l'AI Act, un importante regolamentazione europea per l'intelligenza artificiale.

3.1 GDPR: il volto e la voce come dati personali

Il GDPR, acronimo di *General Data Protection Regulation*, viene definito dal Consiglio dell'Unione Europea e dal Consiglio europeo come la legge sulla privacy e la sicurezza più severa al mondo[35]. Il regolamento, entrato in vigore il 25 maggio 2018, ha sostituito la direttiva 95/46/CE sulla protezione dei dati e, in Italia, ha abrogato gli articoli del decreto legislativo 196/2004 con esso incompatibili. Il GDPR indica quali sono i diritti fondamentali dell'era digitale per tutte le persone appartenenti all'Unione Europea, gli obblighi da parte di coloro che trattano i dati personali, i metodi per garantire la conformità alle norme vigenti e le eventuali sanzioni per chi viola queste regole. Inoltre, è importante sottolineare che il

[35] https://gdpr.eu/what-is-gdpr/

regolamento è applicabile anche al di fuori i territori dell'Unione Europea, se le organizzazioni raccolgono dati relativi alle persone all'interno dell'Unione[36].

Per analizzare in che misura il GDPR offre protezione contro l'uso dannoso dei deepfake, è fondamentale identificare se durante il processo di generazione degli stessi i dati personali vengono elaborati o meno. I dati personali sono definiti come qualsiasi informazione relativa a una persona fisica viva che sia identificata o identificabile, direttamente o indirettamente[37]. Quindi, il GDPR non è applicabile se la persona in questione è deceduta[38], se si tratta di una persona giuridica[39] o se i dati sono anonimi[40].

Come visto nella sezione 2.1.1, alcuni dei meccanismi alla base dei deepfake visivi, quali *face-swap* e *face reenactment*, utilizzano immagini o video di una persona c.d. *target* e immagini o video di una persona c.d. *source* per effettuare la manipolazione. Il volto è la parte identificativa fondamentale di una persona; dunque, un'immagine del volto di una persona che può essere identificata è considerata un dato personale.

Per quanto riguarda i deepfake audio, anch'essi rientrano nella definizione di dato personale espressa dall'articolo 4 comma 1 del GDPR. Dunque, se viene generato artificialmente un audio replicando accuratamente le caratteristiche della voce di una determinata persona, avremmo in quel caso un dato personale.

Come esplicitato nel Considerando 51 del GDPR, l'immagine del volto e la voce rientrano nella definizione di dati biometrici quando sono trattate attraverso un dispositivo tecnico specifico che consente l'identificazione univoca o l'autenticazione della persona fisica[41].

[36] Articolo 3 EU GDPR: https://gdpr-text.com/it/read/article-3/
[37] Articolo 4 comma 1 EU GDPR: https://gdpr-text.com/it/read/article-4/
[38] Considerando 27 EU GDPR: https://gdpr-text.com/it/read/recital-27/
[39] Considerando 14 EU GDPR: https://gdpr-text.com/it/read/recital-14/
[40] Considerando 26 EU GDPR: https://gdpr-text.com/it/read/recital-26/
[41] Considerando 51 EU GDPR: https://gdpr-text.com/it/read/recital-51/

3.2 Deepfake pornography

Il primo utilizzo malevolo della tecnologia dei deepfake risale al 2017, quando vengono diffuse su Reddit immagini e video pornografici ritraenti star di Hollywood, come Gal Gadot, Emma Watson e Taylor Swift, realizzati all'insaputa delle stesse. Ad oggi, il 96% dei deepfake presenti sul web sono di natura pornografica e il 99% di essi riguardano donne. A differenza dell'esordio dei deepfake, i quali ritraevano principalmente personaggi famosi, da qualche anno chiunque può diventare vittima di tali contenuti, dunque, è fondamentale capire e analizzare il fenomeno.

Il termine "deepfake *pornography*", o "deepfake pornografici", fa riferimento all'utilizzo di tecniche generative per l'alterazione di immagini e video con contenuto pornografico. Le vittime del contenuto in questione vengono "spogliate" artificialmente ed appaiono in atteggiamenti sessualmente espliciti, che in realtà sono a loro estranei. Sebbene le immagini siano modificate artificialmente con la tecnica dello *face-swap*, dunque le vittime non siano effettivamente coinvolte nell'atto sessualmente esplicito, la creazione e la successiva divulgazione dei contenuti multimediali falsi lede inevitabilmente la dignità e la privacy delle vittime, poiché collocate forzatamente in un contesto a cui non appartengono.

3.2.1 Il caso DeepNude e l'istruttoria del Garante per la Protezione dei Dati Personali contro Telegram

Nel giugno 2019, viene introdotta un'applicazione chiamata "*DeepNude*" che, tramite l'utilizzo dell'algoritmo *open source* pix2pix[42], permetteva a chiunque di "spogliare" virtualmente le immagini di un soggetto femminile[43], ottenendo risultati realistici. Nonostante la versione gratuita dell'app avesse il c.d. *watermarking* che copriva parte dell'immagine e quella a pagamento una scritta in alto a sinistra con la dicitura "*fake*", con *software* informatici per la manipolazione

[42] Modello GAN per la creazione di immagini realistiche a partire da schizzi (Isola et al., 2016)

[43] Come specificato dal creatore dell'app, era possibile creare deepfake solamente di soggetti femminili, in quanto è più facile trovare in rete immagini in cui appaiono nudi

delle immagini, quali *Photoshop*, era semplice rimuovere queste scritte e far circolare in rete le immagini come se fossero vere[44].

A soli pochi giorni dal lancio, il creatore si è trovato costretto a chiuderle l'app "per motivi etici", in quanto era stata scaricata e utilizzata da oltre novanta mila persone e i risultati erano diventati preoccupanti. Avendo un codice sorgente *open source*, pochi mesi dopo lo stesso strumento viene reso disponibile su Telegram, dove le immagini delle donne vengono manipolate tramite l'uso di un BOT[45] chiamato *DeepNude*, in quanto riprendeva l'utilizzo dell'app precedente.

Visti i potenziali effetti lesivi del software e l'eventuale pericolosa diffusione delle immagini false, nell'ottobre 2020 il Garante per la Protezione dei Dati Personali ha aperto un'istruttoria (GPDP, 2020) nei confronti di Telegram per tutelare la privacy delle vittime, poiché, come emerge nel comunicato stampa, è presente un elevato rischio che le immagini in questione possano essere utilizzate per fini estorsivi o di revenge porn[46].

3.2.2 Possibili tutele penali all'interno dell'ordinamento italiano

In Italia non esiste una legge *ad hoc* che tuteli le vittime dalla diffusione non consensuale di contenuti sessualmente espliciti realizzati artificialmente; dunque, verranno in seguito discusse delle disposizioni che potrebbero essere applicabili alla fattispecie "deepfake pornografici", per la protezione della persona lesa.

Per quanto riguarda la diffusione illecita di immagini o video sessualmente espliciti reali, il Codice penale italiano dispone già una tutela specifica, introdotta nel 2019

---

[44] https://www.wired.it/mobile/app/2019/06/28/app-deepnude-fake-donne/

[45] Abbreviazione di *robot*, è un programma che esegue attività automatizzate e ripetitive che viene progettato per imitare o sostituire le azioni di un essere umano

[46] Il Revenge porn fa riferimento alla diffusione non consensuale di foto intime o sessualmente esplicite, spesso da parte dell'ex partner. È importante specificare che, nonostante sia diventato di utilizzo comune anche a livello istituzionale, spesso è utilizzato impropriamente. La parola "revenge", ovvero vendetta, implicherebbe che la vittima abbia innescato in qualche modo il comportamento dell'autore dell'atto lesivo, quasi colpevolizzandola. Inoltre, parlare di "porn" fraintenderebbe la natura dei materiali diffusi, passando dalla sfera intima alla dimensione pubblica ed esposta del prodotto pornografico, con il mero scopo dell'intrattenimento (Viola & Voto, 2022). Per queste ragioni, è più opportuno utilizzare l'espressione di diffusione non consensuale di immagini o video sessualmente espliciti, permettendo di tenere in considerazione uno spettro più ampio di condotte

con il c.d. "Codice Rosso[47]". L'articolo 612 *ter* del Codice penale è volto a tutelare la riservatezza di contenuti sessualmente espliciti, in quanto potrebbero ledere la reputazione e la dignità della persona offesa, punendo sia chi diffonde le immagini senza il consenso della persona rappresentate, sia coloro che ricevono le immagini, le scaricano dal web o le diffondono a loro volta[48]. Dunque, la norma in questione è apparentemente idonea anche a prevenire i rischi dei deepfake pornografici. Con una sua lettura più rigorosa, i contenuti in questione dovrebbero innanzitutto essere realizzati con il consenso della vittima, non all'insaputa della stessa come nel caso dei deepfake pornografici; tuttavia, la normativa non fa alcun riferimento a contenuti multimediali falsi. Secondo il principio di tassatività[49], il quale implica che la norma penale deve individuare gli estremi del fatto-reato in essa contenuti in modo che si possa desumere con precisione ciò che è lecito e ciò che è vietato, e data la mancanza di giurisprudenza in materia, non è quindi possibile far rientrare i deepfake pornografici nella fattispecie incriminatrice quale l'art. 612 ter c.p.[50].

Una possibile tutela penale per i deepfake pornografici si può ricercare nell'art. 595 comma 3 del Codice penale con il reato di diffamazione aggravata[51]. In particolare, la sentenza n. 41276 del 2015 della Cassazione penale afferma che: "*integra il reato di diffamazione la condivisione sulla rete di filmati riproducenti scene di atti sessuali*"[52]. Dunque, sebbene i deepfake pornografici non siano veri, sono in grado di offendere gravemente la reputazione del soggetto leso e quindi possono rientrare nella fattispecie quale l'art. 595 c.p.[53].

[47] Legge 19 luglio 2019, n. 69 "Modifiche al Codice penale, al codice di procedura penale e altre disposizioni in materia di tutela delle vittime di violenza domestica e di genere"

[48] Art. 612 *ter* c.p. (R.D. 19 ottobre 1930, n. 1398) "Diffusione illecita di immagini o video sessualmente espliciti"

[49] Corollario del principio di legalità, di cui gli artt. 25 Cost., 1 c.p. e 99 c.p.

[50] *Ivi*, p. 30

[51] Art. 595 c.p. (R.D. 19 ottobre 1930, n. 1398) "Diffamazione"

[52] Cassazione penale, sez. V, del 10 marzo 2015, n. 41276

[53] *Ibidem*

Quando i deepfake pornografici vengono generati per fini estorsivi verso la vittima stessa, è possibile ricorrere alla tutela penale per il reato di estorsione, disposto dall'art. 629 c.p.[54].

Altre fattispecie che sono potenzialmente idonee alla tutela della vittima dei deepfake pornografici sono il furto di identità, di cui all'art. 494 c.p.[55], lo stalking, disposto dall'art. 612 *bis* c.p.[56] e il trattamento illecito dei dati personali, art. 167 del Codice della Privacy[57].

3.2.3 Direttiva del Parlamento Europeo e del Consiglio sulla violenza contro le donne e la violenza domestica

L'8 marzo 2022 è stata proposta dalla Commissione Europea una Direttiva riguardante la violenza contro le donne e la violenza domestica[58]. Tale direttiva è un atto legislativo di fondamentale importanza nonostante siano presenti ancora delle lacune (WAVE, 2022), in quanto l'approccio delle disposizioni della stessa è orientato soprattutto verso la criminalizzazione a discapito della prevenzione, della lettura intersezionale e del riconoscimento del lavoro svolto dai centri antiviolenza.

Con il Comunicato Stampa del 9 giugno 2023[59], il Consiglio dell'Unione Europea ha definito la sua posizione a riguardo. L'articolo 7 della Direttiva in questione criminalizza la condivisione non consensuale di materiale intimo o manipolato. Nello specifico, l'art. 7 (b) sancisce come reato il "*produrre, [...] manipolare o alterare e successivamente rendere accessibile al pubblico [...], tramite tecnologie dell'informazione e della comunicazione, immagini, video o analogo materiale in modo da far credere che un'altra persona partecipi ad atti sessualmente espliciti,*

[54] Art. 629 c.p. (R.D. 19 ottobre 1930, n.1398) "Estorsione"
[55] Art. 494 c.p. (R.D. 19 ottobre 1930, n.1398) "Sostituzione di persona"
[56] Art. 612 bis c.p. (R.D. 19 ottobre 1930, n.1398) "Atti persecutori"
[57] Art. 167 del Codice della Privacy (D.lgs. 30 giugno 2003, n.196) "Trattamento illecito dei dati"
[58] https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52022PC0105
[59] https://www.consilium.europa.eu/it/press/press-releases/2023/06/09/violence-against-women-and-domestic-violence-council-agrees-position-on-draft-eu-law/

senza il suo consenso, qualora tali condotte possano arrecare un danno grave a detta persona"[60].

Inoltre, integrerebbe reato anche "*minacciare di assumere le condotte di cui alle lettere a) e b) al fine di costringere un'altra persona a compiere un determinato atto, acconsentirvi o astenersi dallo stesso*"[61].

Dunque, se la proposta di Direttiva venisse approvata, gli Stati Membri sarebbero obbligati a procedere ad una regolamentazione della fattispecie[62] e, dunque, ci sarebbe una legge *ad hoc* che tuteli le vittime dalla diffusione non consensuale di contenuti sessualmente espliciti realizzati artificialmente.

3.2.4 Deepfake pornografici e minori

Come riportato dal *The Korea Herald*, a maggio 2021 la polizia ha arrestato 94 persone, tra teenager e ventenni, tutte sospettate di aver computo reati di diversa natura nei cinque mesi precedenti, mediante l'utilizzo della tecnologia deepfake. In particolare, hanno realizzato molteplici video pornografici ritraenti vittime dai 10 ai 20 anni, tra cui 109 ragazze e 5 ragazzi[63].

In Italia, le vittime di deepfake pornografici minori di anni diciotto sono tutelate dal nostro ordinamento tramite l'art. 600 *quater* 1 c.p., che disciplina il reato di pornografia virtuale. Nel comma 2 si dispone che le immagini sessualmente esplicite possono essere anche "*realizzate con tecniche di elaborazione grafica non associate in tutto o in parte a situazioni reali, la cui qualità di rappresentazione fa apparire come vere situazioni non reali*"[64]. Inoltre, vengono integrati anche ai reati di pornografia minorile e la detenzione di materiale pornografico realizzato utilizzando minori degli anni diciotto.

---

[60] Art. 7 (b) della proposta di direttiva del Consiglio Europeo del 17 marzo 2023 9305/23 (Fascicolo interistituzionale: 2022/0066(COD))

[61] Art. 7 (c) della proposta di direttiva del Consiglio Europeo del 17 marzo 2023 9305/23 (Fascicolo interistituzionale: 2022/0066(COD))

[62] Art. 288 TFUE

[63] https://www.koreaherald.com/view.php?ud=20210502000064

[64] Art. 600 *quarter* 1 c.p. (R.D. 19 ottobre 1930, n.1398) "Pornografia virtuale"

3.2.5 Rimozione dei contenuti sessualmente espliciti falsi: diritto all'oblio

Ogni qualvolta una persona vuole eliminare contenuti ritraenti sé stessa, può farne richiesta. Il GDPR sancisce il diritto all'oblio, ovvero alla cancellazione dei dati personali che riguardano la persona ritraente in foto o video, senza ingiustificato ritardo[65].

Inoltre, per contrastare la diffusione incontrollata di immagini e video sessualmente espliciti falsi, il *National for Missing and Exploited Children* (NCMEC), con il supporto di Meta, PornHub, Only Fans e Yubo, ha realizzato "*Take It Down*" una piattaforma che consente ai minori di anni diciotto di segnalare immagini e video espliciti di sé stessi da Internet[66].

Come riportato dall'app stessa, anche le vittime con età superiore agli anni diciotto possono procedere alla segnalazione dei contenuti in questione, tramite il sito web "*StopNCII.org*"[67].

3.3 Deepfake e disinformazione

Uno studio condotto dal *Massachusetts Institute of Technology* [68] (MIT), analizzando la piattaforma social Twitter, ha dimostrato che le notizie false circolano fino a sei volte più velocemente di quelle vere. Fino a qualche anno fa le notizie false erano principalmente testuali, ma ora lo possono essere anche sotto forma di immagini o video, creando sempre più convinzione che una notizia falsa sia vera. Questo avviene perché le persone sono ancora inclini a credere a quello che vedono, dunque che il falso sia reale.

Il miglioramento progressivo di queste tecniche generative sta permettendo, come nel caso dei deepfake pornografici, di generare immagini e video falsi sempre più realistici e sofisticati, e i *software* sono alla portata di tutti.

---

[65] Art. 17 EU GDPR

[66] Disponibile al link: https://takeitdown.ncmec.org

[67] Disponibile al link: https://stopncii.org

[68] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

3.3.1. Dal caso Obama al caso Zelensky

Per dimostrare la potenzialità dei rischi dei deepfake, nel 2018 il regista di Get Out, Jordan Peele, e BuzzFeed hanno generato una finta dichiarazione pubblica da parte di Barack Obama[69], tramite l'applicazione FakeApp e Adobe After Effects. Nel maggio 2019, la prima vittima di queste tecnologie è Nancy Pelosy, la quale appare ad un convegno come se fosse ubriaca e, tra le numerose persone, anche l'ex sindaco di New York Rudy Giuliani era convinto che il video fosse vero. In risposta al rifiuto da parte di Facebook di rimuovere dal social il video deepfake della Speaker della Camera dei Rappresentanti[70] è la volta di Mark Zuckerberg, il quale si vantava in un video di "avere il controllo sulle nostre vite". Sempre nello stesso periodo anche in Italia vengono creati dei deepfake da parte di Striscia la Notizia, come quello ritraente Matteo Renzi. Nel 2022 è la volta del presidente ucraino Zelensky, il quale chiederebbe ai suoi soldati di deporre le armi e di arrendersi alla Russia.

Negli ultimi anni le tecniche di *lip-syncing* stanno diventando sempre più avanzate e dunque i video falsi sono sempre più difficili da riconoscere, ma non solo, anche le immagini create a scopo di disinformazione sono sempre più sofisticate e possono creare situazioni di momentaneo panico. Ne è stato esempio la falsa l'esplosione del Pentagono il 22 maggio che ha fatto crollare vertiginosamente la borsa di Wall Street per qualche minuto, causando una momentanea perdita di qualche miliardo di dollari di capitalizzazione totale[71].

3.3.2 Il caso FakeYou e l'istruttoria del Garante per la Protezione dei Dati Personali contro la società "The Storyteller"

Nel 2021, viene introdotto l'aggiornamento di "FakeYou", il quale, facendo uso della tecnologia Text-to-Speech[72], consente all'utente di scegliere un determinato personaggio famoso, quale Giorgia Meloni, e successivamente dato in input un testo

[69] https://www.youtube.com/watch?v=cQ54GDm1eL0&t=1s
[70] https://www.buzzfeednews.com/article/davidmack/facebook-nancy-pelosi-doctored-video
[71] https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/
[72] *Ivi*, p. 17

generare un audio falso con la voce del personaggio in questione, facendogli dire frasi che non ha mai detto.

Visti i potenziali rischi "*che potrebbero determinarsi da un uso improprio di un dato personale, quale è appunto la voce*", il Garante per la Protezione dei Dati Personali ha aperto un'istruttoria nei confronti di "*The Storyteller*". La società, come specificato dal Garante, dovrà fornire con urgenza le modalità di generazione della voce dei personaggi famosi, la tipologia di dati che sono trattati e le finalità del trattamento dei dati dei personaggi famosi e dei dati degli utenti che fanno uso dell'applicazione in questione. Inoltre, viene chiesto di "*indicare l'ubicazione dei data center che archiviano i dati personali, sia con riferimento agli utenti registrati dall'Italia, sia ai personaggi noti, e le misure tecniche ed organizzative adottate per garantire un livello di sicurezza adeguato al rischio*" (GPDP, 2022).

Inoltre, applicazioni come FakeYou portano inevitabilmente ad ampliare il fenomeno della disinformazione in quanto consentono a tutti in modo semplice e veloce di clonare la voce di personaggi famosi, quali politici, attori, giornalisti.

3.3.3 Regolamentazione italiana delle fake news

Le fake news sono regolate dall'articolo 656 del Codice penale[73]. È importante specificare che in questa norma non c'è la distinzione tra la pubblicazione della notizia falsa e la diffusione della stessa; dunque, quando viene pubblicato un contenuto realistico di un personaggio famoso dove dice qualcosa ce non ha mai detto, sono punibili tutti coloro che contribuiscono alla diffusione della notizia falsa. Inoltre, nel diritto penale vale il principio del *ignoratia legis non excusa*, ovvero l'ignoranza della legge non scusa, motivo per cui nessuno può invocare l'ignoranza, l'inconsapevolezza o la buona fede dopo aver contribuito alla diffusione di una fake news[74]. La fattispecie descritta dal 656 c.p. non viene applicata nel caso in cui il fatto costituisca più grave reato, come il disfattismo

[73] Art. 653 c.p. (R.D. 19 ottobre 1930, n.1398) "Pubblicazione o diffusione di notizie false, esagerate o tendenziose, atte a turbare l'ordine pubblico"

[74] Art. 5 c.p. (R.D. 19 ottobre 1930, n.1398) "Ignoranza della legge penale"

politico[75] e il rialzo e ribasso fraudolento di prezzi sul pubblico mercato o nelle borse di commercio[76].

La diffusione di contenuti multimediali falsi, inoltre, è integrata con il reato di sostituzione di persona e potrebbe essere integrato dal reato di diffamazione aggravata, descritto dall'articolo 595 comma 3 del Codice penale[77].

3.4 Deepfake audio e truffe telefoniche

Un ulteriore utilizzo malevole dei deepfake sono le truffe telefoniche, eseguite tramite la clonazione della voce di una determinata persona con il fine di estorcere denaro. La clonazione della voce è un concetto nuovo, ma desta già numerose preoccupazioni in quanto sta diventando sempre più accurato, accessibile a tutti e semplice da utilizzare grazie alle nuove applicazioni di intelligenza artificiale.

La *Federal Trade Commission* riporta che solo nel 2022 sono state più di 5 mila le segnalazioni di truffe telefoniche effettuate da persone che pretendevano di essere amici o familiari, per una perdita totale di 11 milioni di dollari[78]. Come, ad esempio, il caso descritto dal *The Washington Post,* nel quale Ruth Card e il marito Greg Grace effettuano, senza esitazione, un bonifico di 3 mila dollari canadesi verso un conto online, dopo essere stati chiamati dal presunto nipote in prigione che chiedeva i soldi per la cauzione[79]. E ancora, come riportato dal *The Wall Street Journal*, il caso in Regno Unito del CEO di un'azienda di energia che effettua un versamento di 220 mila euro sul conto bancario di un fornitore ungherese dopo essere stato contattato telefonicamente dal suo presunto superiore[80].

---

[75] Art. 265 c.p. (R.D. 19 ottobre 1930, n.1398) "Disfattismo politico"

[76] Art. 501 c.p. (R.D. 19 ottobre 1930, n.1398) "Rialzo e ribasso fraudolento di prezzi sul pubblico mercato o nelle borse di commercio"

[77] *Ivi*, p. 33

[78] https://www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022

[79] https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/

[80] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

3.4.1 Tutele penali sul piano interno

In Italia, il reato di truffa telefonica è disciplinato dall'art. 640 del Codice penale: "*chiunque, con artifizi o raggiri, inducendo taluno in errore, procura a sé o ad altri un ingiusto profitto con altrui danno*"[81]. Inoltre, viene anche integrato il reato di sostituzione di persona, di cui all'art. 494 c.p.[82]

3.5 AI ART

L'arte viene definita dall'enciclopedia Treccani come il "*complesso di regole ed esperienze elaborate dall'uomo per produrre oggetti o rappresentare immagini tratte dalla realtà o dalla fantasia*"[83]. Sono compresi nella definizione di arte la pittura, la scultura, la musica, l'architettura, la danza, la poesia e la recitazione.

La creazione di contenuti creativi artistici realizzati tramite l'utilizzo delle tecniche generative hanno portano a due principali problematiche di tipo legale: la possibile violazione dei diritti di proprietà intellettuale[84] delle opere già esistenti e i problemi di attribuzione sulla titolarità dei diritti sulle creazioni dell'intelligenza artificiale.

3.5.1 Il caso di Andersen contro Stability AI et al.

Il 13 gennaio 2023, Sarah Andersen, Kelly McKernan e Karla Ortiz hanno intentato una causa[85] contro le piattaforme Stability AI Ltd., Midjourney Inc. e DeviantArt Inc., definendole "strumenti di collage che violano i diritti di milioni di artisti". I tre artisti sostengono che le piattaforme hanno utilizzato le opere senza consenso o compenso per la costruzione del *dataset* di training, denominato LAION-Aesthetics[86]. Inoltre, Andersen afferma che le immagini generate tramite l'utilizzo illecito del suo stile e quello di altri autori ha danneggiato il valore che la loro arte

[81] Art. 640 c.p. (R.D. 19 ottobre 1930, n.1398) "Truffa"

[82] *Ivi*, p. 31

[83] https://www.treccani.it/enciclopedia/arte

[84] Insieme di diritti legali che mirano alla tutela delle creazioni umane in ambito scientifico, industriale e artistico. Fanno parte dei diritti di proprietà intellettuale i diritti d'autore, i diritti di proprietà industriale e i diritti connessi

[85] https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/

[86] Per un approfondimento sulla vicenda: https://news.artnet.com/art-world/class-action-lawsuit-ai-generators-deviantart-midjourney-stable-diffusion-2246770

aveva in precedenza, in quanto l'arte è stata "diluita" in mezzo a tante immagini dall'aspetto simile[87].

Recentemente, anche Getty Images ha intentato una causa separata contro Stability AI, sostenendo violazioni dei diritti d'autore e di proprietà sui marchi registrati, in quanto ha trovato più di 15 mila immagini dalla sua libreria all'interno del *dataset* di training di Stable Diffusion. Il CEO della società Getty Images afferma in un'intervista che non ricerca un compenso finanziario, ma vuole solamente chiarezza legale sui diritti dei creatori[88].

3.5.2. Tutela dei diritti di proprietà intellettuale delle nuove tecnologie

Da come emerge da questi casi e come visto nel primo capitolo, le tecniche generative creano dati sintetici a partire dai *dataset* di training; dunque, tutte le opere realizzate dall'intelligenza artificiale prendono spunto inevitabilmente da immagini di oggetti già esistenti. Ad oggi, non esiste una normativa *ad hoc* per la protezione dei diritti di proprietà intellettuale derivanti dall'utilizzo di immagini esistenti per la generazione di nuovi contenuti multimediali tramite le tecniche generative, né a livello internazionale, né a livello nazionale; quindi, ogni tribunale può rispondere con un'interpretazione diversa.

In Italia, le opere d'arte generate con l'intelligenza artificiale potrebbero incorrere nel reato di plagio, disciplinato dall'articolo 171 comma 1 lettera a-bis) della legge sul diritto d'autore[89]. Il problema, però, è molto più ampio, poiché non è chiaro se la responsabilità nel caso in cui venga appurato che ci sia plagio sia attribuita al soggetto che fa uso del *software*, lo sviluppatore o entrambi.

Il 7 aprile 2023, l'Harvard Business Review ha pubblicato un articolo dove viene affrontato lo scontro tra la spiccata abilità delle nuove tecnologie alla creazione di

---

87 https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html

88 https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

89 Art. 171 della Legge sulla protezione del diritto d'autore (L. 22 aprile 1941, n. 633)

contenuti creativi e i diritti d'autore, riportando alcune buone pratiche che devono essere adottate da creatori, sviluppatori e aziende per la mitigazione dei rischi[90].

3.5.3 Tutela dei diritti derivanti dalle creazioni dell'intelligenza artificiale

Le leggi esistenti in materia di proprietà intellettuale sono concordi nell'attribuire i diritti d'autore alle opere che sono realizzate dalla mente umana, ma, dal momento in cui ad oggi l'uomo non è l'unico in grado di creare contenuti artistici, sorge spontaneo chiedersi chi è il titolare dei diritti d'autore delle opere che vengono generate dall'intelligenza artificiale.

È importante prima di tutto distinguere tra le opere generate dall'intelligenza artificiale in modo completamente autonomo e le opere generate con l'assistenza dell'uomo. Nel secondo caso, l'intelligenza artificiale sarebbe uno strumento e quindi implicherebbe lo sforzo creativo umano per la creazione dell'opera (ad esempio, nella generazione delle frasi di *prompt*). Quindi il diritto d'autore spetterebbe alla persona fisica che ha dato origine all'opera, anche se dovrebbero essere analizzati più nel dettaglio i contributi individuali di ognuno. Mentre è più complicato definire chi sia il titolare dei diritti d'autore nel caso in cui l'intelligenza artificiale opera in modo completamente autonomo, poiché le macchine non possiedono la capacità giuridica.

Una prima ipotesi è quella di attribuire i diritti d'autore e di sfruttamento economici in capo al soggetto che ha ideato il *software* in questione o colui che ha fornito i mezzi necessari alla sua realizzazione[91]. La seconda ipotesi è quella data dalla proposta del Parlamento Europeo in materia di diritti di proprietà intellettuale.

3.5.4 Proposta di risoluzione del Parlamento europeo sui diritti di proprietà intellettuale per lo sviluppo di tecnologie di intelligenza artificiale

Nel 2020, il Parlamento europeo ha proposto una risoluzione riguardante i diritti di proprietà intellettuale per lo sviluppo delle tecnologie dell'intelligenza artificiale,

[90] https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

[91] https://www.exportiamo.it/aree-tematiche/15001/intelligenza-artificiale-e-proprieta-intellettuale-una-partita-ancora-aperta/

per proteggere l'innovazione, garantire la certezza del diritto e creare la fiducia necessaria per incoraggiare gli investimenti in queste nuove tecnologie (Parlamento europeo, 2020). La risoluzione esplicita che la titolarità dei diritti spetterebbe ancora solamente alle persone giuridiche, in particolare il titolare dei diritti sarebbe colui che ha redatto e pubblicato legittimamente l'opera. L'unico requisito (Franceschelli & Musolesi, 2022) è considerare soddisfatta la condizione di originalità non solo se il processo è creativo, ma anche quando lo è il risultato, poiché il presupposto rimane che le creazioni tradizionali e quelle generate dall'intelligenza artificiale abbiano lo scopo di espandere il patrimonio culturale in comune, nonostante la creazione avvenga attraverso un diverso atto.

3.6 AI ACT

Il 21 aprile 2021 la Commissione europea ha presentato la proposta di un regolamento volto all'armonizzazione delle leggi sull'intelligenza artificiale[92]. Il progetto di legge in questione è il primo tentativo in assoluto di emanare una regolamentazione orizzontale dell'AI con l'obiettivo di garantire il corretto funzionamento del mercato unico europeo, creando le condizioni per lo sviluppo e l'utilizzo di sistemi di AI affidabili nell'Unione Europea.

In particolare: i) garantire che i sistemi di AI immessi sul mercato dell'Unione siano sicuri e rispettino il diritto dell'Unione già esistente, ii) garantire la certezza del diritto per facilitare gli investimenti e l'innovazione nell'AI, iii) migliorare la governance e l'effettiva applicazione del diritto dell'Unione in materia di diritti fondamentali e requisiti di sicurezza applicabili ai sistemi di AI e iv) favorire lo sviluppo di un mercato unico per le applicazioni di AI legali, sicure e affidabili, evitando la frammentazione del mercato.

La bozza del regolamento in questione, come specificato dal Titolo II, segue un approccio basato sul rischio, ovvero l'intervento legale è adattato al livello concreto di rischi. Tutti i sistemi che presentano un rischio limitato (Madiega, 2021), tra i cui

[92] Proposta completa al link: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0006.02/DOC_1&format=PDF

i *software* che creano contenuti deepfake vengono proposti degli obblighi di trasparenza minimi. Come riportato nel titolo IV della proposta di regolamento "*Se un sistema di IA viene utilizzato per generare o manipolare immagini o contenuti audio o video che assomigliano notevolmente a contenuti autentici, dovrebbe essere previsto l'obbligo di rivelare che tali contenuti sono generati ricorrendo a mezzi automatizzati*".

Con l'approvazione dell'AI act da parte del Parlamento europeo il 14 giugno 2023, l'entrata in vigore del regolamento è stimata per l'inizio del 2024 e, come per il GDPR, decorsi i 24 mesi sarà direttamente applicabile in tutti gli Stati membri.

Capitolo 4

Ricerca: detection dei deepfake audio

Nel capitolo precedente sono state analizzate le problematiche legali a cui possono portare le tecniche generative se utilizzate in modo non opportuno. In particolare, come la generazione dei deepfake audio stia diventando sempre più sofisticata, rendendo difficile per le persone riconoscere cosa sia vero da cosa sia falso, incorrendo nella possibilità di, ad esempio, truffe telefoniche.

Per contrastare l'abuso dei deepfake audio, gli esperti stanno sviluppando due approcci: la rilevazione manuale e la rilevazione automatica. La rilevazione automatica viene effettuata tramite algoritmi di *deep learning,* mentre quella manuale consiste nella rilevazione grazie ad agenti umani esperti che analizzano individualmente ogni audio.

4.1 Obiettivo della ricerca

L'obiettivo della ricerca è quello di valutate la capacità umana di rilevare gli audio deepfake, in particolar modo quando l'algoritmo di classificazione non identifica correttamente gli audio deepfake.

4.2 Scelta dei campioni e partecipanti

Al fine della ricerca sono stati eseguiti quattro test diversi. Nel primo e nel secondo test sono stati analizzati 60 audio, per un totale di 9 audio veri e 51 audio falsi (85% falsi). Questi due test vengono raggruppati in quanto viene utilizzato lo stesso metodo per la scelta dei campioni. Per la terza prova sono stati analizzati 31 audio, di cui 12 veri e 19 falsi (61.3% falsi). Mentre per la quarta prova sono stati analizzati 31 audio, di cui 5 veri e 26 falsi (83.9% falsi). Per ottenere risultati il più affidabili possibile, sono stati prelevati 22 campioni per ogni audio, per un totale di 1320 campioni per il test 1 e il test 2, 682 per il test 3 e 682 per il test 4.

Gli studenti che hanno preso parte al test erano iscritti ai corsi di laurea magistrale di Cybersecurity e ICT for Internet and Multimedia, con un'età compresa tra i 21 e i 36 anni, il 57% erano uomini, mentre il 43% donne.

Il *dataset* contenente gli audio utilizzati in questa ricerca è quello dell'Automatic Speaker Verification, Spoofing and Countermeasures Challenge del 2019 (ASVSpoof), il quale era composto per l'80% di audio falsi (Todisco et al., 2019).

4.3 Modalità di esecuzione

Il test è stato effettuato in ambienti isolati e nella stanza erano presenti solamente due persone: chi somministrava il test e chi era tenuto a sostenerlo. Inoltre, per permettere a tutti di trovarsi nelle stesse condizioni, sono state utilizzate delle cuffie *over-ear* con cancellazione del rumore e l'audio del computer era sempre allo stesso volume per tutti gli utenti.

Ad ogni partecipante è stato spiegato brevemente l'esperimento e dopo eventuali domande o chiarimenti poteva svolgerlo in autonomia. Il *software* è stato implementato con MathLab ed era user friendly. Per prima cosa veniva richiesto all'utente di inserire alcune generalità, quali età, sesso biologico e se avesse problemi di udito. [Fig.4.1]

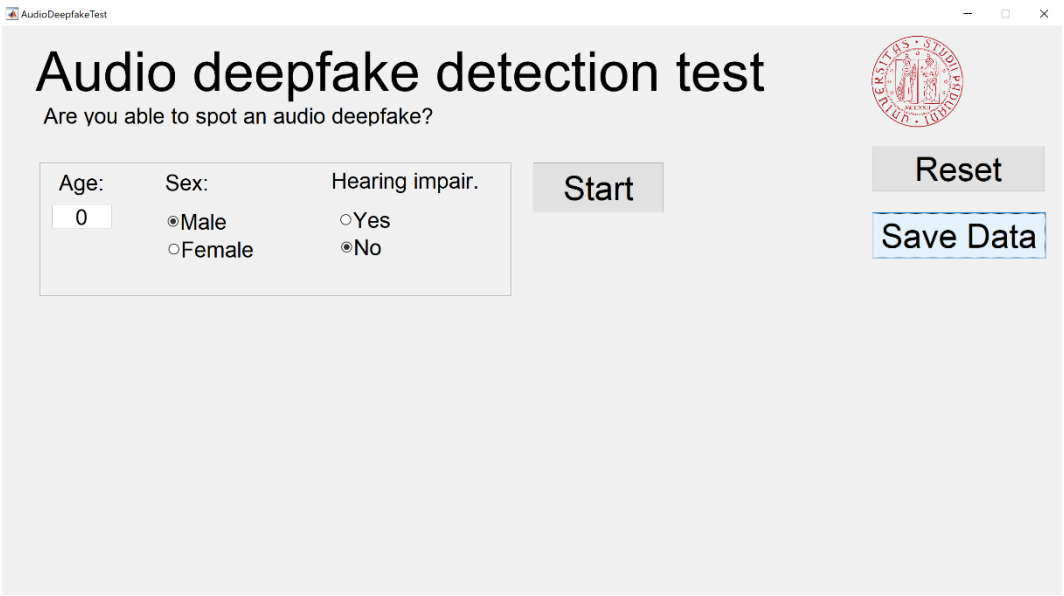

Figura 4.1: Interfaccia grafica della schermata iniziale

In ogni test erano presenti 5 audio di *training* che sono stati successivamente scartati, in modo tale che il partecipante che effettuava il test prendesse confidenza con il *software* in questione. Ogni audio poteva essere ascoltato solamente una volta ed era introdotto da un *countdown* (3, 2, 1) per permettere agli utenti di concentrarsi maggiormente all'ascolto. Dopo ogni ascolto era richiesto di valutare l'audio scegliendo tra le seguenti alternative: reale (codificato successivamente come 1), reale ma non sicuro (2), non so (3), falso ma non sicuro (4) e falso (5). [Fig.4.2]

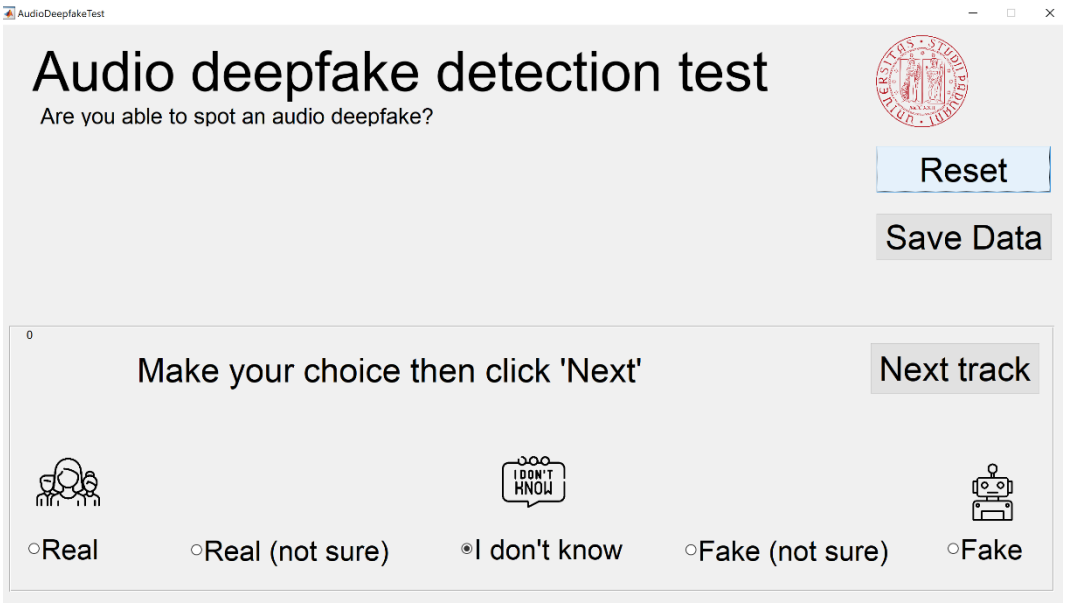

Figura 4.2: Interfaccia grafica del test dopo aver ascoltato un audio

Alla fine del test i dati venivano salvati in formato *csv* e veniva chiesto al partecipante se ci fossero eventuali considerazioni.

4.3 Analisi

Per l'analisi dei dati raccolti è stato implemento un programma con R Studio.

Per ogni test sono stati importati tre file *csv* principali: il *dataset* principale contenente l'audio ID e il corrispettivo valore reale, il *dataset* contenente l'audio ID e i risultati di detection dell'algoritmo di classificazione e il *dataset* contenente l'audio ID e i risultati di detection dei partecipanti del test; tutti aventi il *label* uguale a 1 per gli audio falsi e a 0 per gli audio veri. Inoltre, sono stati importati anche il

file contenente le generalità dei partecipanti e il file contenente la lunghezza in secondi per ogni audio ID.

Per ogni test è stata calcolata la percentuale di uomini e donne, l'età minima e massima ed è stato verificato se qualcuno avesse qualche problema di udito. È emerso che sul totale di 88 persone una persona aveva problemi di udito.

Per ogni test, la metrica utilizzata per l'analisi dei risultati è stata la matrice di confusione: una matrice 2x2 con 4 diverse combinazioni di valori effettivi e previsti che permettono di valutare le prestazioni della detection dei deepfake audio:

- TP (True Positive): predire che è un audio è falso ed è effettivamente falso;
- FP (False Positive): predire che è un audio è falso, ma in realtà è vero;
- TN (True Negative): predire che un audio è vero ed è effettivamente vero;
- FN (False Negative): predire che un audio è vero, ma in realtà è falso.

		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (T**N**)	False Positives (F**P**) **Type I error**
	Positive +	False Negatives (F**N**) **Type II error**	True Positives (T**P**)

Figura 4.3: Matrice di confusione

La prima matrice è stata quella dell'algoritmo di classificazione, mentre la seconda quella con i risultati dati dai partecipanti del test. Successivamente è stata calcolata l'accuratezza per entrambe le matrici, ovvero quante classi sono state previste correttamente, sia prendendo in considerazione gli audio veri e quelli falsi, sia solamente gli audio falsi.

Infine, tramite l'utilizzo di un ggplot, è stato creato un grafico rappresentante tutti gli audio per ogni test che mostra l'accuratezza di risposta sia del classificatore che dei partecipanti per ogni audio, in modo da consentire un'analisi visiva dei risultati e individuare in modo immediato gli audio più significativi.

4.4 Risultati

4.4.1 Test 1 e 2

Su un totale di 60 (51 falsi) audio, quelli individuati dai partecipanti sono un totale di 41 (33) mentre quelli sbagliati sono 19 (18), con una precisione del 68,3% (64,7%). Gli audio individuati dal classificatore sono in totale 29 (21) mentre quelli sbagliati sono 31 (30), con un'accuratezza del 48,3% (41,2%). In questo caso, l'accuratezza del classificatore non risulta elevata in quanto per la creazione del *dataset* dei test 1 e 2 sono stati intenzionalmente selezionati gli audio in cui il classificatore commetteva errori.

Considerando il *dataset* dei partecipanti e il *dataset* del classificatore, su un totale di 51 falsi audio del test 1 e del test 2, il 25,5% degli audio viene individuato da entrambi, il 56,9% da almeno uno di loro (41,2% i partecipanti, 15,7% il classificatore) e il 17,6% nessuno di essi.

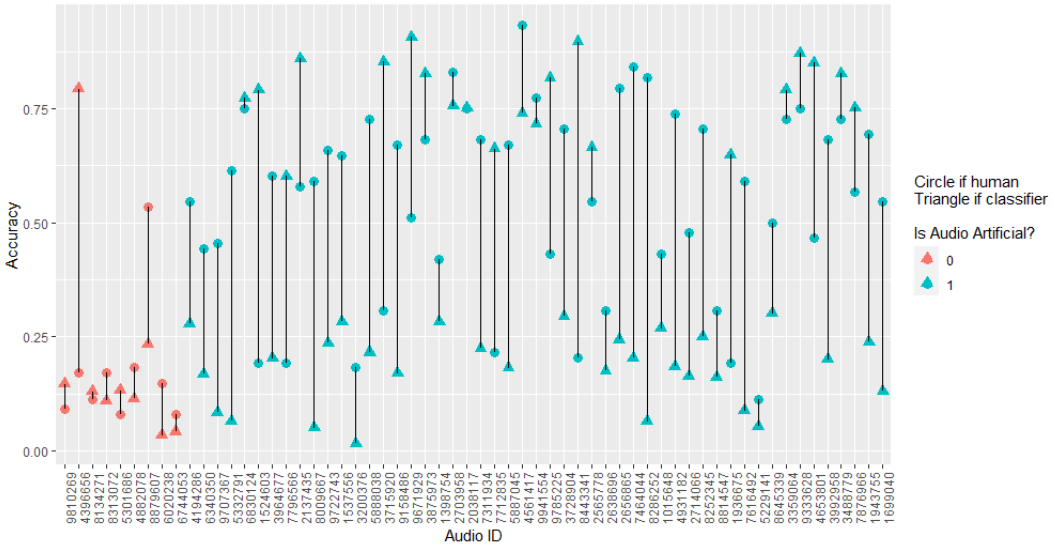

Figura 4.4: comparazione dell'accuratezza delle risposte dei partecipanti (cerchio) e del classificatore (triangolo) per ogni audio vero (rosso) e falso (blu) del test 1 e 2

4.4.2 Test 3

Durante l'analisi dei risultati è emerso che il *dataset* utilizzato per questo test non era ottimale, dunque è stato scartato.

4.4.3 Test 4

Su un totale di 31 (26 falsi) audio, quelli individuati dai partecipanti sono un totale di 25 (20) mentre quelli sbagliati sono 6 (6), con una precisione dell'80,7% (76,9%). Gli audio individuati dal classificatore sono in totale 25 (21) mentre quelli errati sono 6 (5), con un'accuratezza dell'80,6% (80,8%).

Considerando il *dataset* dei partecipanti e il *dataset* del classificatore, su un totale di 26 falsi audio dal test, il 61,5% degli audio viene individuato da entrambi, il 34,6% da almeno uno di essi (15,3% i partecipanti, 19,2% il classificatore) e il 3,9% nessuno di loro.

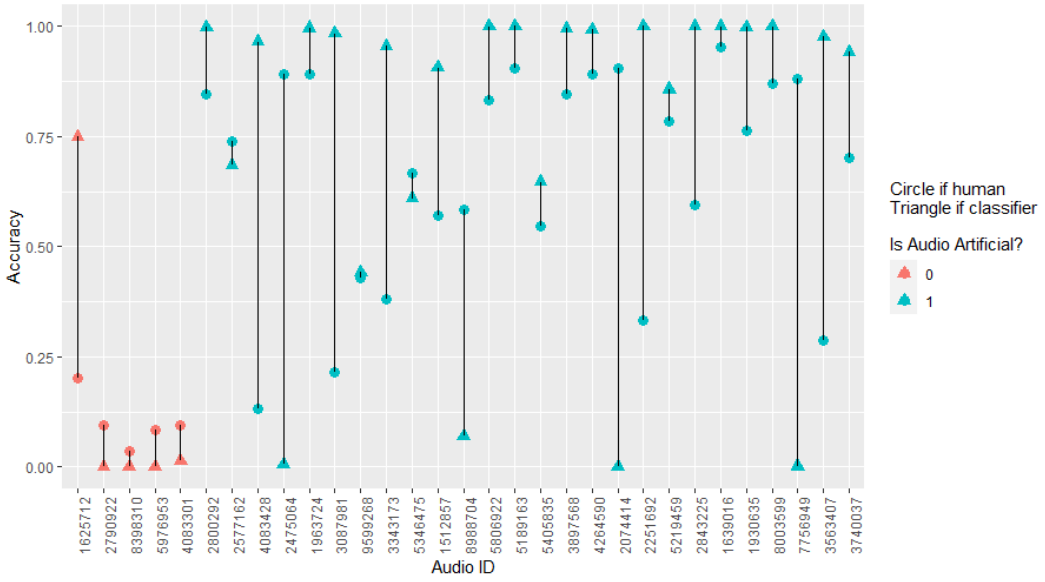

Figura 4.5: comparazione dell'accuratezza delle risposte dei partecipanti (cerchio) e del classificatore (triangolo) per ogni audio vero (rosso) e falso (blu) del test 4

4.5 Considerazioni finali

Come si evince dai grafici [Fig. 4.4 e 4.5], i *dataset* che sono stati utilizzati per valutare la detection degli audio falsi avevano caratteristiche diverse. Nel test 4, il *dataset* era principalmente composto da audio in cui l'algoritmo aveva valori di

classificazione vicino ai limiti (0,1). Nonostante vi fossero alcuni casi in cui l'algoritmo commetteva errori significativi, ad esempio rilevando erroneamente un valore di 0 anziché 1 (con una valutazione molto vicina allo 0), l'accuratezza complessiva dell'algoritmo di classificazione era nettamente superiore (80,8% solo audio falsi). Mentre, nel test 1 e 2 sono stati considerati risultati più intermedi, in cui gli errori dell'algoritmo erano meno significativi, ma l'accuratezza complessiva era inferiore (41,2% solo audio falsi).

Per quanto riguarda la detection effettuata dai partecipanti, anche in questo caso l'accuratezza è nettamente superiore nel test 4, con il 76,9% di corretta rilevazione degli audio falsi, contro il 64,7% del test 1 e 2.

Conclusione

Lo scopo di questo elaborato è quello di esaminare le implicazioni legali che possono derivare dalla creazione di contenuti multimediali falsi o dalla loro manipolazione, in particolar modo i deepfake video e audio. Attraverso un'analisi dettagliata delle fattispecie, emerge l'urgente necessità di nuove regolamentazioni a livello europeo.

Per quanto riguarda la creazione e successiva diffusione dei deepfake pornografici, la direttiva UE sulla lotta alla violenza contro le donne e la violenza domestica mirerebbe ad integrare la criminalizzazione dei deepfake nel quadro giuridico di tutti gli Stati membri e, dunque, consentirebbe di tutelare le vittime dalla diffusione non consensuale di contenuti sessualmente espliciti realizzati artificialmente.

L'entrata in vigore dell'AI Act porterebbe ad una rigorosa regolamentazione dell'intelligenza artificiale in molteplici settori, tra cui l'intelligenza artificiale generativa. Con l'introduzione dei requisiti minimi di trasparenza, le aziende sarebbero obbligate a dichiarare esplicitamente quando un contenuto è stato generato con l'AI, salvaguardando la generazione di contenuti illegali, e sarebbero tenute a pubblicare in sintesi quali sono i *dataset* di training utilizzati per allenare gli algoritmi per la generazione di contenuti multimediali falsi, mitigando così i possibili rischi derivanti dall'uso dell'AI generativa.

La ricerca presentata evidenzia la necessità di sistemi di detection sempre più sofisticati, in quanto quando l'algoritmo di classificazione commette errori, un utente umano non esperto non è sempre in grado di identificare un audio quando è falso. Alla luce dei risultati ottenuti, su un totale di 91 (77 falsi) audio, 10 (10) non sono stati riconosciuti né dai partecipanti né dall'algoritmo, con una precisione totale di detection del 89,0% (87,0% falsi). Questo sottolinea l'importanza di sviluppare sistemi sempre più accurati per rilevare le possibili frodi a cui gli utenti possono incorrere, tutelando sempre di più l'individuo dagli utilizzi malevoli di queste nuove tecnologie.

Bibliografia

Ajder H., Patrini G., Cavalli F., & Cullen L., *The State of Deepfakes. Landscape, Threats and Impact*, Deeptrace, 2019.

Akhtar Z., *Deepfakes Generation and Detection: A Short Survey*, Journal of Imaging, vol. 9 (1), 18, 2023.

Antoniou A., Storkey A.J., & Edwards H., *Data Augmentation Generative Adversarial Networks,* ArXiv abs/1711.04340, 2017.

Bank D., Koenigstein N., & Giryes R., *AutoEncoders*, ArXiv abs/2003.05991, v. 2, 2021.

Battiato S., Giudice O., & Paratore A., *Multimedia Forensics: discovering the history of multimedia contents,* in Proceedings of the 17th International Conference on Computer Systems and Technologies (CompSysTech'16), Palermo, Italy, 2016, pp. 5 – 16.

Benvinamarad P. R., & Shirldonkar M. S., Audio Forgery Detection Techniques: Present and Past Review, in Proceedings of the 4th International Conference on Trends in Electronics and Informatics (ICOEI), vol. 48184, Tirunelveli, India, 2020, pp. 613 – 618.

Busch C., Vera-Rodriguez R., Tolosana R., & Rathgeb C., *Handbook of Digital Face Manipulation and Detection. From DeepFakes to Morphing Attacks,* in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Springer, 2022.

Commissione europea, *PROPOSTA di Direttiva del Parlamento europeo e del Consiglio sulla lotta alla violenza contro le donne e alla violenza domestica,* 2022/0066 (COD), Strasburgo, 8 marzo 2022.

Commissione europea, *PROPOSTA di Regolamento del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge*

sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'Unione, 2021/0106 (COD), Bruxelles, 21 aprile 2021.

Croitoru F. A., Hondru V., Ionescu R. T., & Shah M., *Diffusion Models in Vision: A Survey*, in Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14 (8), 2022.

Dang H., Liu F., Stehouwer J., Liu X., & Jain A.K., *On the Detection of Digital Face Manipulations,* in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5781 – 5790.

Franceschelli G., & Musolesi M*., Copyright in generative deep learning*, Data & Policy, 4, E17, 2022.

Garante per la Protezione dei Dati Personali (GPDP), *Deep fake: Garante avvia istruttoria su app che falsifica le voci* [Comunicato stampa], Roma, 12 ottobre 2022.

Garante per la Protezione dei Dati Personali (GPDP), *Il Garante privacy apre un'istruttoria nei confronti di Telegram per il software che "spoglia" le donne* [Comunicato stampa], Roma, 23 ottobre 2020.

Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., & Bengio Y., *Generative Adversarial Nets*, in Proceedings of the 27th International Conference in Neural Information Processing Systems (NIPS), vol. 2, 2014.

Isola P., Jun-Yan Z., Tinghui Z., & Alexei A.E., *Image-to-Image Translation with Conditional Adversarial Networks*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5967 – 5976.

Jun, H., & Nichol, A., *Shap-E: Generating Conditional 3D Implicit Functions,* ArXiv, abs/2305.02463*,* 2023.

Kenneth O. S., *Compositional pattern producing networks: A novel abstraction of development,* in Genetic Programming and Evolvable Machines, vol. 8 (2), 2007, pp. 131 – 162.

Kingma D.P., & Welling M., *Auto-Encoding Variational Bayes*, ArXiv abs/1312.6114, v. 11, 2022.

Korshunova I., Shi W., Dambre, J., & Theis L., *Fast Face-Swap Using Convolutional Neural Networks,* in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–27 October 2017, pp. 3697 – 3705.

Langr J., & Box V., *GANs in Action. Deep learning with Generative Adversarial Networks,* Manning Publications, 2019.

Larson E. J., *Il mito dell'intelligenza artificiale. Perché i computer non possono pensare come noi,* in Micalizzi P., Franco Angeli, 2022.

Madiega T. A., *Artificial intelligence act*, European Parliament: European Parliamentary Research Service, 2021.

McCormack J., Gifford T., & Hutchings P., *Autonomy, Authenticity, Authorship and Intention in Computer Generated Art*, in Proceedings of the 8th International Conference in Computational Intelligence in Music, Sound, Art and Design, vol. 11453, Springer Cham, 2019, pp. 35 – 50.

Millière R., *Deep Learning and Synthetic Media,* Synthese, 2022.

Minu M. S., & Ahmad Z., *Augmented Analytics: The Future of Business Intelligence*, in Recent Trends in Computer Science and Software Technology, Mantech Publications, vol. 5 (1), Mantech Publications, 2020, pp. 7 – 13.

Mirsky Y., & Lee W., *The Creation and Detection of Deepfakes: A Survey,* ACM Computing Surveys, vol. 54 (1), Article 7, 2022.

Parlamento europeo, *RELAZIONE sui diritti di proprietà intellettuale per lo sviluppo di tecnologie di intelligenza artificiale*, [Documento di seduta], A9-0176/2020, 2020.

Peng B., Fan H., Wang W., Dong J., & Lyu S., *A Unified Framework for High Fidelity Face Swap and Expression Reenactment,* in Proceedings of the IEEE Transactions on Circuits and Systems for Video Technology, vol. 32 (6), 2022.

Radford A., Metz L., & Chintala S., *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, ArXiv abs/1511.06434, v. 2, 2016.

Ramesh A., Dhariwal P., Nichol A., Chu C., & Chen M., *Hierarchical Text-Conditional Image Generation with CLIP Latents,* ArXiv, abs/2204.06125, 2022.

Ratner A. J., Ehrenberg H., Hussain Z., Dunnmon, J., & Ré, C., *Learning to Compose Domain-Specific Transformations for Data Augmentation*, Advances of the 31st Conference on Neural Information Processing Systems (NIPS), vol. 30, 2017.

Regolamento UE n. 2016/679 (GDPR) e D.lgs. 30.06.2003, n. 196 (Codice in materia in protezione dei dati personali), come modificato dal D.lgs. 10.08.2018, n. 101.

Seow J. W., Lim M. K., Phan R. C.W., & Liu J. K., *A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities*, Neurocomputing, vol. 513, 2022, pp. 351 – 371.

Shorten, C., & Khoshgoftaar, T. M., *A survey on image data augmentation for deep learning,* Journal of big data, vol. 6 (1), 2019, pp. 1 – 48.

Todisco M., Wang X., Sahidullah M, Delgado H., Nautsch A., Yamagishi J., Evans N., Kinnunen T., & Lee K. A*., ASVSpoof 2019: Future horizons in spoofed and fake audio detection*, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2019.

Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., & Polosukhin I., *Attention is All you Need*, NIPS, 2017.

Verdoliva L., *Media Forensics and Deepfakes: An Overview,* in Proceedings of the IEEE Journal of Selected Topics in Signal Processing, vol. 14 (5), 2020, pp. 910 – 932.

Viola M., & Voto C., *La diffusione non consensuale di contenuti intimi ai tempi dei deepfake: una controprofezia ottimista*, in Rivista Italiana di Filosofia del linguaggio, 2022, pp. 65 – 72.

Wang C., Chen S., Wu Y., Zhang Z., Zhou L., Liu S., Chen Z., Liu Y., Wang H., Li J., He L., Zhao S., & Wei F., *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,* ArXiv, 2301.02111, 2023.

Wang X., Wang K., & Lian S., *A survey on face data augmentation for the training of deep neural networks*, Neural Computing & Applications 32, 2020, pp. 15503 – 15531.

Westerlund M., *The Emergence of Deepfake Technology: A Review*, Technology Innovation Management Review, vol. 9 (11), 2019, pp. 40 – 53.

Women Against Violence Europe (WAVE), *Advocacy Update: Proposal for EU Directive Combating Violence Against Women and Domestic Violence*, [Dichiarazione pubblica], novembre 2022.

Yu Y., Gong Z., Zhong P., & Shan J., *Unsupervised Representation Learning with Deep Convolutional Neural Network for Remote Sensing Images*, in Image and Graphics, ICIG, vol. 10667, Springer Cham, 2017, pp. 97 – 108.

Zanardelli M., Guerrini F., Leonardi R., & Adami N., *Image forgery detection: a survey of recent deep-learning approaches*, Multimedia Tools and Applications, vol. 82, 2023, pp. 17521 – 17566.

Zhang Z., Zhou L., Wang C., Chen S., Wu Y., Liu S., Chen Z., Liu Y., Wang H., Li J., He L., Zhao S., & Wei F., *Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling*, arXiv, abs/2303.03926, 2023.

Zhao H., & Malik H., *Audio recording location identification using acoustic environment signature*, in Information Forensics and Security, in Proceedings of the IEEE Transactions, vol. 8, 2013, pp. 1746 – 1759.

Zhu Y., Li Q., Wang J., Xu C., & Sun, Z., *One Shot Face Swapping on Megapixels,* in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4832 – 4842.

Sitografia

Andersen S., *The Alt-Right Manipulated My Comic. Then A.I. Claimed it*: https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html, consultato il 10 giugno 2023.

Appel G., Neelbauer J., & Schweidel D. A., *Generative AI Has an Intellectual Property Problem*: https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem, consultato il 10 giugno 2023.

BuzzFeedVideo, *You Won't Believe What Obama Says In This Video!*: https://www.youtube.com/watch?v=cQ54GDm1eL0&t=1s, visionato il 16 maggio 2023.

Castelli M., *Intelligenza Artificiale e Proprietà Intellettuale: una Partita ancora Aperta:* https://www.exportiamo.it/aree-tematiche/15001/intelligenza-artificiale-e-proprieta-intellettuale-una-partita-ancora-aperta/, consultato il 2 luglio 2023.

Chen M., *Artists and Illustrators Are Suing Three A.I. Art Generators for Scraping and 'Collaging' Their Work Without Consent*: https://news.artnet.com/art-world/class-action-lawsuit-ai-generators-deviantart-midjourney-stable-diffusion-2246770, consultato il 10 giugno 2023.

Court Listened: https://www.courtlistener.com, consultato il 12 giugno 2023.

De Jure, *Banche dati editoriali GFL*: https://dejure.it/, consultato il 25 maggio 2023.

Dizikes P., *Study: On Twitter, false news travels faster than true stories*: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308/, consultato il 21 maggio 2023.

Federal Trade Commision, *New FTC Data Show Consumers Reported Losing Nearly $8.8 Billion to Scams in 2022. Reported fraud losses increase more than 30 percent over 2021:* https://www.ftc.gov/news-events/news/press-

releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022/, consultato il 12 marzo 2023.

Giacobini G., *Storia dell'app che genera(va) false foto di nudo femminile*: https://www.wired.it/mobile/app/2019/06/28/app-deepnude-fake-donne/, consultato il 21 maggio 2023.

https://eur-lex.europa.eu/homepage.html, consultato il 30 maggio 2023

Mack D., *Facebook Said It Won't Take Down A Doctored Video Of Nancy Pelosi That It Knows Is Fake. The video makes the House speaker appear to slur her words*: https://www.buzzfeednews.com/article/davidmack/facebook-nancy-pelosi-doctored-video, consultato il 9 giugno 2023.

Oremus W., Harwell D., & Armus T., *A tweet about a Pentagon explosion was fake. It still went viral. Apparently AI-generated image sparked a brief dip in stock market. It could have been worse*: https://www.washingtonpost.com/technology/2023/05/22/pentagon-explosion-ai-image-hoax/, consultato il 19 maggio 2023.

Stupp C., *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. Scams using artificial intelligence are a new challenge for companies*: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, consultato il 15 maggio 2023.

Verma P., *They thought loved ones were calling for help. It was an AI scam. Scammers are using artificial intelligence to sound more like family members in distress. People are falling for it and losing thousands of dollars:* https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/, consultato il 6 maggio 2023.

Vincent J., *Getty Images is suing the creators of AI tool Stable Diffusion for scraping its content*: https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit, consultato il 10 giugno 2023.

Wolford B., *What is GDPR, the EU's new data protection law*?: https://gdpr.eu/what-is-gdpr/, consultato il 19 maggio 2023.

Yonhap, *Police arrest 94 suspects over deepfake crimes in 5 months, with 70 percent teenagers*: https://www.koreaherald.com/view.php?ud=20210502000064, consultato il 21 maggio 2023.