Università degli studi di Padova

Dipartimento di Diritto Pubblico, Internazionale e Comunitario

Corso di Laurea Triennale in

DIRITTO E TECNOLOGIA

Dalle filter bubbles ai deepfake: nuove forme di manipolazione della realtà. Un'analisi degli impatti e delle misure per il contrasto

Relatore: Prof. Claudio Sarra

Laureanda: Sara Baldo

Matricola: 2037673

Anno Accademico 2023/2024

ABSTRACT

Ad oggi quasi ogni aspetto della nostra esistenza è governato da sistemi di machine learning in grado di apprendere autonomamente dai dati e dagli esempi forniti per produrre previsioni. La criticità emerge nel momento in cui i dati dell'utente profilato vengono impiegati per filtrare sulla base dei suoi interessi i contenuti da mostrare. L'utente si ritrova confinato all'interno di una bolla di informazioni coerenti con la propria visione della realtà in grado di incidere sulla formazione del dibattito pubblico e sulla propagazione di disinformazione. Il deep learning rappresenta il ramo più avanzato del machine learning. I contenuti deepfake, prodotti mediante deep learning, hanno la caratteristica di apparire estremamente realistici e dunque idonei ad ingannare il destinatario che non è in grado di riconoscere l'artificialità del contenuto. Il presente elaborato intende analizzare gli impatti e il legame tra bolle filtro, disinformazione e deepfake, oltre che analizzare i principali approcci adottati da piattaforme e istituzioni per il contrasto del fenomeno nel suo complesso.

INDICE

INTRODUZIONE **5**

CAPITOLO I **7**

L'EFFETTO FILTER BUBBLE DEGLI ALGORITMI DI PROFILAZIONE **7**

1.1 Il machine learning e la profilazione 7

1.2 Dalla profilazione alla filter bubble 11

1.3 Impatto sociale: conseguenze e rischi della bubble democracy 15

CAPITOLO II **19**

IL DEEP LEARNING: UN NUOVO MODO DI PRODURRE DISINFORMAZIONE **19**

2.1 La disinformazione online: definizioni, origine ed esempi 19

2.2 Disinformazione iperrealistica: deep learning e deepfake 24

2.3 I principali modelli generativi di IA 29

CAPITOLO III **33**

MISURE E TENTATIVI DI REGOLAMENTAZIONE PER IL CONTRASTO ALLA DISINFORMAZIONE E AI DEEPFAKE **33**

3.1 Fact-checking e altre misure pratiche per combattere la disinformazione 33

3.2 Contrasto ai deepfake e Regolamento europeo sull'IA 39

CONCLUSIONE **43**

INTRODUZIONE

I sistemi di *machine learning* governano oggigiorno quasi ogni aspetto della nostra esistenza. Introdotta negli anni ‘50, l’espressione “machine learning” descrive quelle tecnologie in grado di apprendere autonomamente da dati, esempi ed esperienze maturate per generare previsioni. Se da una parte gli algoritmi di machine learning agevolano le nostre vite fornendoci ad esempio raccomandazioni personalizzate sulla base dei nostri interessi, dall'altra vi è il rischio di un isolamento intellettuale. Le informazioni fornite dagli algoritmi sono infatti frutto di un filtraggio. Nel 2011 Eli Pariser conia l’espressione “*filter bubble”* per descrivere il fenomeno: l’utente si ritrova confinato all'interno di una bolla di informazioni coerenti con la sua personale visione della realtà. Queste bolle rappresentano il terreno fertile per la propagazione di disinformazione e di teorie estremiste che potrebbero talvolta sfociare in episodi violenti quale è stato ad esempio il *pizzagate*. Il presente elaborato intende focalizzarsi sulla questione della disinformazione, ed in particolare sulla generazione di contenuti ingannevoli o falsi estremamente convincenti mediante impiego di sistemi di Intelligenza Artificiale. Il *deepfake* rappresenta l’ultima frontiera del fenomeno; grazie alla tecnologia *deep learning*, una branca del machine learning fondata su reti neurali, è oggi possibile produrre artificialmente contenuti multimediali estremamente realistici, in cui un individuo può risultare fare o dire cose mai fatte o dette, con conseguenze drammatiche. Giornalisti, fornitori dei servizi web e istituzioni hanno tentato di implementare delle soluzioni per contrastare la diffusione incontrollata di disinformazione online e di deepfake, compresa una regolamentazione dei due fenomeni, fino a giungere nel 2024 all'approvazione del primo regolamento al mondo sull’Intelligenza Artificiale da parte del Parlamento europeo. Il primo capitolo del presente elaborato cerca di descrivere i sistemi di machine

learning e le conseguenze derivanti dall'impiego di dati di natura personale per l'addestramento dell'algoritmo. Il secondo capitolo approfondisce il fenomeno della disinformazione online, diretta conseguenza delle filter bubbles ed estremamente dannosa quando prodotta mediante deep learning, si parla a tal proposito di deepfake. All'interno del capitolo vengono anche presentati alcuni esempi dei più noti modelli generativi di IA disponibili, tra cui GPT, DALL-E e Sora. Il terzo capitolo infine illustra le migliori strategie e tentativi di regolamentazione adottati per combattere la disinformazione online e di conseguenza i deepfake, in particolare si analizza il Regolamento europeo sull'IA approvato nel 2024.

CAPITOLO I

L'EFFETTO FILTER BUBBLE DEGLI ALGORITMI DI PROFILAZIONE

1.1 Il machine learning e la profilazione

Per essere in grado di comprendere a pieno la portata e le conseguenze sociali della *filter bubble,* risulta necessario fare un passo indietro ed analizzare la tecnologia che costituisce le fondamenta di suddetto fenomeno. Il riferimento è al *machine learning*, una branca dell'intelligenza artificiale che si occupa di implementare sistemi in grado di apprendere autonomamente dai dati e dagli esempi forniti, oltre che dalle esperienze maturate[1]. Affinché la macchina sia in grado di eseguire un determinato compito, la programmazione, e dunque l'intervento umano, oggi non sono più indispensabili[2]. In uno studio pubblicato dal Parlamento Europeo sull'impatto del GDPR (Regolamento Generale sulla Protezione dei Dati) sull'Intelligenza Artificiale, il machine learning viene descritto come un sistema che individua le affinità tra dati e produce corrispondenti modelli che permettono di predire risposte corrette a possibili input[3]. In altri termini, i sistemi che adottano tale tecnologia vengono istruiti a produrre previsioni sulla base di una moltitudine di dati ed esempi forniti, l'algoritmo in completa autonomia riconosce patterns e con il tempo è in grado di provvedere output ai problemi proposti[4].

Se oggi il machine learning rappresenta una delle tecnologie più diffuse nella quotidianità di ciascun individuo, è grazie alla disponibilità sempre crescente di dati e all'incremento del potere computazionale[5]. Ogni giorno ciascun soggetto, inconsapevolmente, contribuisce alla immissione di molteplici dati nel cosiddetto "mercato dei dati" [6]. Una semplice ricerca su Google o richiesta vocale ad un assistente virtuale come Siri o Alexa, rappresentano una risorsa di dati prelevati dalla macchina per auto-apprendere il modo in cui comunicano oralmente e per iscritto gli esseri umani nelle varie parti del mondo. Maggiori saranno i dati immessi, migliori saranno le prestazioni

---

[1] Montaldo, 2020, p. 225
[2] Royal Society, 2017, p.5
[3] *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* Study Panel for the Future of Science and Technology. European Parliamentary Research Service. PE 641.530 June 2020.
[4] O'Neil, 2016, p.69
[5] Royal Society, 2017, p. 16 ss.
[6] Bianca, 2019, p.42

dell'algoritmo. Non sorprende dunque scoprire che l'algoritmo di Google Translate produca risultati più accurati quando la lingua in cui un testo viene tradotto è l'inglese, Google ha difatti accesso ad una vastità di documenti, siti web e ricerche degli utenti da cui apprendere prettamente in lingua inglese[7]. Per quanto concerne l'evoluzione del potere computazionale, risulta fondamentale citare la legge di Moore[8].

La nascita del concetto di machine learning si deve ad Arthur Samuel, ingegnere di IBM che alla fine degli anni '50 creò un software simulatore del gioco della dama, programmato per apprendere dalle sue esperienze[9]. Da allora tale tecnologia si è evoluta e ad oggi chiunque ha la possibilità di interagire quotidianamente con sistemi che la adottano. Si pensi alle raccomandazioni di Spotify, Netflix, Amazon o piattaforme simili; nel momento in cui viene effettuato l'accesso al profilo personale dell'utente, impiegando tecniche di machine learning l'algoritmo analizza una serie di dati tra cui i click, la cronologia e la geolocalizzazione e mostra contenuti che prevede possano essere rilevanti per l'utente[10]. Ma il machine learning non si limita ai soli sistemi di raccomandazione, esso viene applicato anche ai motori di ricerca come Google[11], ai sistemi di riconoscimento vocale, agli assistenti virtuali[12] e ai filtri antispam delle caselle di posta elettronica[13]. Negli ultimi anni la diffusione di tale tecnologia ha coinvolto i più svariati campi, la salute[14], la finanza, i servizi pubblici e il settore legale ne sono solo alcuni esempi[15]. Il futuro del machine learning è senza dubbio rappresentato dalla tecnologia di *deep learning*, un ramo del primo che impiega reti neurali per effettuare previsioni[16]. Il prossimo capitolo si focalizza sull'analisi di tale tecnologia e il suo impiego in rete. In particolare viene trattato il fenomeno del *deepfake*, che attraverso la tecnologia di deep

---

[7] Crepaldi, 2023

[8] La legge di Moore, teorizzata da Gordon Moore nel 1965, prevedeva che il numero di transistor all'interno di un circuito integrato sarebbe raddoppiato ogni due anni circa, con un incremento della velocità, affidabilità e contemporanea diminuzione del costo. Treccani, *Legge di Moore*

[9] Franklin, 2014, p. 7

[10] Laddove i dati raccolti siano insufficienti, l'algoritmo estrae dati di utenti terzi che possiedono caratteristiche simili all'utente principale. Si parla in questo caso di *collaborative filtering*. Longo, 2019, p.39.

[11] L'algoritmo ritorna in output i link alle pagine web che egli considera rilevanti per l'utente, in base alla query immessa.

[12] Come Alexa o Siri, che sono in grado di percepire la parola pronunciata dall'utente e di trascriverla, tradurla o eseguire dei comandi. Royal Society, 2017, p. 22 ss.

[13] Pitruzzella, 2018, p.26. Altri settori che coinvolgono il machine learning sono la *computer vision*, la traduzione automatica e la *fraud detection*. Royal Society, 2017, p.22

[14] Attraverso la produzione di diagnosi o l'assistenza ai medici. Il machine learning può estrarre dati da cartelle mediche dei pazienti, ma anche da dispositivi smart come gli smart watch e restituire informazioni.

[15] Royal Society, 2017, p.34 ss

[16] Paris, 2019, p.12

learning è in grado di riprodurre artificialmente contenuti estremamente realistici e quindi idonei ad ingannare un pubblico non consapevole.

Quando il machine learning impiega dati personali per produrre output accurati il termine da utilizzare per descrivere il fenomeno è "profilazione", definito dal GDPR come "qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica"[17]. L'idea di una profilazione degli individui nasce dalla psicometria, scienza che analizza la personalità di un individuo quantificandola[18]. Le piattaforme, a partire dai dati, plasmano delle identità digitali degli utenti impiegate poi dagli inserzionisti per cogliere i gusti, desideri e comportamenti dei consumatori[19] e dunque innescare il comportamento desiderato[20]. Un esempio interessante proviene da Netflix. La nota piattaforma di streaming, analizzando i dati dei propri utenti, ha previsto che una serie diretta da David Fincher e con attore protagonista Kevin Spacey ispirata ad un prodotto della BBC degli anni '90 avrebbe avuto successo, è nata così la popolare serie House of Cards[21]. Se per le aziende la profilazione rappresenta una efficace tecnica di marketing mirato, per l'utente essa consiste in una diretta irruzione della privacy[22] e, come si vedrà nel paragrafo seguente, in pericolose manipolazioni della coscienza individuale e collettiva.

Il GDPR pone particolare attenzione alla questione della profilazione, *ex* art 22 si considera la profilazione una tipologia di processo decisionale automatizzato relativo alle persone fisiche, pertanto *ex* artt. 13 e 14 l'interessato deve esserne informato per garantire che il trattamento sia corretto e trasparente. Nello specifico, non è possibile sottoporre l'interessato a decisioni fondate unicamente sulla profilazione laddove esse producano effetti giuridici che lo riguardano o che in ogni caso incidano significativamente sulla sua persona. Costituiscono deroghe a tale diposizione i casi in cui la profilazione sia

[17] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati), *ex* art. 4.4.
[18] Astone, 2020, p. 167
[19] Bianca, 2019, p.46
[20] Lagioia, 2020, p.91
[21] Italiano, 2018, p.220
[22] Bianca, 2019, p. 42

necessaria per concludere o eseguire un contratto tra interessato e titolare del trattamento, nel caso in cui il titolare del trattamento sia uno Stato membro dell'Unione Europea o vi sia espressa autorizzazione del diritto dell'UE, e infine laddove l'interessato abbia manifestato esplicito consenso. In ogni caso l'interessato ha diritto ad "ottenere l'intervento umano da parte del titolare del trattamento, di esprimere la propria opinione e di contestare la decisione"[23]. È possibile osservare come il GDPR non rappresenti in realtà un rimedio efficace contro la profilazione dell'interessato, la fruizione dei servizi forniti da social networks e portali presuppone che l'interessato abbia accettato le condizioni poste dal titolare del trattamento, si tratta di fatto di un contratto *ex* art. 22 par. 2, lett. *a)* che dunque impedisce all'interessato di avere una tutela contro la profilazione [24]. Anche il consenso rappresenta un ostacolo lieve per il titolare del trattamento, l'interessato tende in effetti ad accettare trattamenti dei dati con eccessiva facilità, senza possedere una consapevolezza adeguata delle conseguenze di un consenso, ovvero la perdita del pieno controllo sulla propria identità personale in rete[25].

Il paragrafo seguente illustrerà come la profilazione effettuata dalle piattaforme abbia un ruolo del tutto centrale nella formazione del dibattito pubblico e nella propagazione della disinformazione. I contenuti che vengono mostrati all'utente in rete, essendo frutto di profilazione, rispecchiano le sue preferenze, vi sono quindi una serie di informazioni con cui egli non entrerà mai in contatto. L'utente si ritrova confinato in una bolla di informazioni coerenti con la propria visione della realtà[26], le conseguenze, come si vedrà, possono essere drammatiche.

---

[23] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati), *ex* art 22.

[24] Le Condizioni d'uso di Meta ad esempio chiariscono che, affinché l'utente possa usufruire dei prodotti gratuitamente, l'azienda necessita degli introiti degli inserzionisti. Meta utilizza perciò le informazioni personali dell'utente per mostrargli le inserzioni ritenute rilevanti. Per maggiori dettagli: https://it-it.facebook.com/legal/terms

[25] Astone, 2020, p.1, Bianca, 2019, p.48

[26] Fasan, 2019, p. 109

1.2 Dalla profilazione alla filter bubble

Le conseguenze della profilazione in rete, se introdotte da esempi tratti dalla quotidianità di ciascun individuo possono essere più agevolmente comprese. Si prendano pertanto in esame gli algoritmi di Google, il motore di ricerca più utilizzato. La funzione primaria dei motori di ricerca è fornire le informazioni richieste dall'utente. Si tratta di veri e propri intermediari tra la fonte e il destinatario dell'informazione; essi individuano, selezionano, ordinano e solo infine mostrano i contenuti relativi alla query digitata. Questo processo viene realizzato da alcuni algoritmi, è dunque l'algoritmo a stabilire l'ordine e il tipo di informazioni che verranno presentate all'utente. Se un'informazione appare al primo posto o nella terza pagina di Google, ciò può influenzare notevolmente il destinatario del contenuto[27]. Il potere informativo è dunque nelle mani di chi adotta l'algoritmo e non di chi realizza le notizie.

Tra gli algoritmi impiegati da Google figurano *Page Rank* e *RankBrain*. Nonostante sia piuttosto complesso trovare informazioni ufficiali su come essi operino, visti i diritti di proprietà intellettuale in mano al colosso e la generale opacità degli algoritmi[28], è comunque possibile sintetizzare il loro funzionamento nel seguente modo: *Page Rank* analizza dati oggettivi come il numero di link contenuti in una pagina[29], le sue prestazioni tecniche e la corrispondenza tra i termini contenuti in essa e quelli contenuti nella query, per individuare una prima iniziale gerarchia dei siti basata sulla presunta qualità. Successivamente l'algoritmo analizza le ricerche passate di tutti gli utenti Google e la specifica query inserita dall'utente per perfezionare l'ordine dei contenuti da presentare. *RankBrain* invece è in grado di associare le parole digitate dall'utente all'interno della barra di ricerca al probabile significato inteso, senza che vi debba essere corrispondenza tra i termini utilizzati dall'utente e i termini contenuti nelle pagine web da indicizzare. Ciò è reso possibile dalla capacità dell'algoritmo di individuare similitudini tra molteplici query di ricerca, superando la problematicità di un input ambiguo fornito dall'utente[30].

---

[27] Pitruzzella, 2018, p. 25
[28] *Ivi*, 2018, p. 27
[29] Una pagina con più link è infatti considerata di alta qualità. Wang, 2003, p.1
[30] Se l'utente digita nella barra di ricerca "best flower shop in Los Angeles" l'algoritmo è in grado di percepire la somiglianza con un'altra ricerca più popolare ad esempio "best LA flower shops", dunque il

Facebook rappresenta un ulteriore esempio di come il contenuto mostrato all'utente sia strettamente legato all'operato degli algoritmi. All'interno della sezione *News Feed* del popolare social network, i post, messaggi, e notizie visibili vengono selezionati e ordinati dall'algoritmo *EdgeRank*, che utilizzando tecniche di machine learning prevede quali contenuti rispecchiano le preferenze dell'utente e dunque sono idonei a generare maggiori interazioni[31]. Una ricerca pubblicata dal Washington Post rivela che circa il 60% dei post degli amici e delle pagine seguite da ciascun utente non sono in realtà visibili nel proprio *News Feed*[32]. L'obiettivo primario dei social network e più in generale delle piattaforme è monetizzare attraverso i compensi pubblicitari. Dunque maggiore è l'engagement dell'utente, ovvero i click, i commenti, i like, il tempo trascorso all'interno della piattaforma, maggiore è il profitto di inserzionisti e conseguentemente delle piattaforme, titolari degli spazi pubblicitari[33].

Attraverso queste lunghe ma doverose premesse, il lettore avrà compreso che, se da un lato ricevere informazioni che rispecchiano le richieste dell'utente e i suoi personali interessi e gusti rappresenta certamente un vantaggio in termini di tempo, costo della ricerca, soddisfazione nell'interazione social, dall'altro lato vi è il rischio di un isolamento intellettuale. La personalizzazione dei contenuti in rete, frutto di algoritmi profilatori in grado di conoscere ogni aspetto della personalità dell'utente, impedisce di entrare in contatto con tutta la restante parte di informazioni in circolazione[34]. Questo fenomeno prende il nome di *filter bubbling*, la fattispecie patologica della profilazione[35].

La *filter bubble*, o bolla di filtraggio, è la condizione nella quale un soggetto nella rete riceve informazioni filtrate sulla base dei propri interessi attraverso algoritmi che, come si è già discusso precedentemente, sono in grado di apprendere dai dati condivisi consensualmente[36]. La conseguenza è che l'utente incontra solo contenuti in grado di confermare le proprie convinzioni e pregiudizi[37], perdendo di fatto l'accesso al pluralismo

---

ranking di pagine fornite da Google Search sarà relativo alla seconda ricerca, di cui l'algoritmo ha più dati degli utenti a disposizione. Sullivan, 2016; Pitruzzella, 2018, p. 25; Fasan, 2019, p.108

[31] Pitruzzella, 2018, p. 27

[32] Costa, 2016, p. 257 ss.

[33] Astone, 2020, p. 167, Berkman Klein Center for Internet & society…, p.1

[34] Pitruzzella, 2018, p. 29

[35] Bianca, 2019, p.48

[36] Astone, 2020, p. 164

[37] Si parla a tal proposito di *confirmation biases*, la psicologia insegna che l'uomo per natura tende ad assimilare solo informazioni che coincidono o sono in grado di rafforzare le proprie opinioni. Hooke Pearson, G., Knobloch-Westerwick, S., 2019.

informativo e la capacità di esercitare consapevolmente l'autodeterminazione[38]. Il fenomeno fin qui descritto non nasce in realtà con la rete, la così definita "polarizzazione di gruppo" è un processo naturale appartenente ad ogni società. Tuttavia attraverso la rete gli effetti si estendono, raggiungendo più soggetti e agevolando la diffusione di notizie false o distorte[39].

La bolla rappresenta un terreno fertile per la creazione di gruppi sociali isolati composti da membri che condividono le stesse idee, ritenute da questi uniche verità, e che tendono a rifiutare l'esistenza di opinioni differenti[40]. Quando un soggetto, accedendo alla rete, si ritrova nella sua personale bolla, difficilmente realizza di ricevere contenuti personalizzati[41], e ancora più difficilmente è consapevole dell'esistenza di contenuti divergenti rispetto alle proprie idee[42]. Quando si parla di filter bubbling, si fa riferimento anche al fenomeno delle *echo chambers*, o camere d'eco, spazi virtuali all'interno dei quali il soggetto incontra esclusivamente opinioni che confermano le proprie, manifestate durante accessi precedenti alla rete. Bolle filtro e camere d'eco congiuntamente contribuiscono a formare gruppi sociali con ideologie spesso tendenti all'estremismo, alla radicalizzazione e alla cospirazione[43].

L'espressione '*filter bubble'* venne introdotta dall'attivista del web Eli Pariser nel 2011, con la pubblicazione del saggio "*The filter bubble. What Internet is hiding from you*". L'opera seguì di un paio d'anni la decisione di Google e Facebook di adottare sistemi di personalizzazione, Pariser voleva mettere in guardia il mondo dal concreto pericolo dell'utilizzo di tecniche di profilazione ed algoritmi nella rete[44]. Egli descrive il fenomeno come "quel personale ecosistema di informazioni che viene soddisfatto da alcuni algoritmi"[45]. Altre definizioni descrivono la bolla come un ambiente virtuale in cui regnano l'autoreferenzialità e un basso livello di accettazione delle novità[46]. I contenuti che difficilmente l'utente approverebbe in quanto non coerenti con i personali interessi culturali, politici, ideologici non vengono mostrati[47], e questo non perché

[38] Astone, 2020, p 164
[39] Vernice, 2021, p. 96
[40] Fasan, 2019, p. 112
[41] Longo, 2019, p.41
[42] *Ivi*, 2019, p.40
[43] Montaldo, 2020, p. 227
[44] Bianca, 2019, p. 44
[45] Pariser, 2011
[46] Treccani, *Filter bubble*
[47] Pariser, 2011

l'utilizzatore dell'algoritmo abbia il desiderio di manipolare le coscienze collettive, ma piuttosto perché segregare gli internauti in gruppi chiusi agevola l'operato degli inserzionisti che riescono a raggiungere i target più facilmente[48]. Alcuni studiosi ipotizzano che la bolla sia frutto delle scelte consapevoli dell'utente, e dunque non vi sarebbe il pericolo di conseguenze negative. Gli algoritmi interverrebbero solo per incrementare meccanismi già innestati dall'utente, è il singolo individuo infatti a decidere di interagire prettamente con soggetti che condividono le proprie opinioni[49]. Nonostante vi sia un fondo di verità in tale teoria, va ricordato che l'utente è scarsamente consapevole dei meccanismi che regolano il web, e potrebbe perciò ignorare il fatto che vi siano ulteriori idee e opinioni al di fuori di quelle mostrate dall'algoritmo.

Nonostante le piattaforme sembrerebbero operare in buona fede, spinte unicamente dal desiderio di incrementare i profitti, gli effetti delle bolle di filtraggio potrebbero essere devastanti per la democrazia laddove i contenuti diffusi siano di tipo politico o ideologico. Il prossimo paragrafo ha l'obiettivo di definire le conseguenze della cosiddetta "*bubble democracy*"[50], anche attraverso la presentazione di alcuni esempi che hanno coinvolto la politica mondiale negli ultimi anni.

[48] Montaldo, 2020, p. 227
[49] Palano, 2019, p. 78
[50] Bianca, 2019, p. 45

1.3 Impatto sociale: conseguenze e rischi della bubble democracy

L'isolamento dell'individuo all'interno di una filter bubble contribuisce all'incremento della frammentazione e polarizzazione dei gruppi sociali, specialmente quando i contenuti diffusi sono di tipo politico. In un contesto virtuale in cui, a causa della personalizzazione esercitata dall'algoritmo, l'individuo con difficoltà incontra contenuti di ideologia opposta, si fa estremamente limitata la possibilità di confronto, ponendo dunque in serio pericolo il futuro della democrazia [51]. Si parla a tal proposito di *bubble democracy*.

Uno stato democratico si basa su una cittadinanza consapevole, il pluralismo informativo, ovvero la circolazione di una pluralità di idee ed opinioni, è un principio fondamentale sancito dal nostro ordinamento in via interpretativa *ex* art 21 Cost.[52]. Tale diritto tuttavia risulta fortemente leso dagli effetti generati dalla profilazione e dalle conseguenti bolle di filtraggio. Il soggetto in rete ha a disposizione una importante mole di informazioni proveniente da una pluralità di fonti, tuttavia come è stato appurato nel corso di questa trattazione, è l'algoritmo a stabilire i contenuti da mostrare. La bolla è concepita perciò come un attentato alla democrazia[53].

Vi sono alcuni elementi fondamentali che contribuiscono a danneggiare l'assetto democratico di un Paese. In primo luogo l'utente medio non è mai pienamente cosciente di ricevere informazioni veicolate dall'algoritmo, egli è dunque convinto che ciò che Internet mostra sia l'unica e neutra verità[54]. Inoltre vi è il grande problema della diffusione di fake news e disinformazione[55], sempre più persone si affidano a blog, forum e social network anziché ai tradizionali canali informativi, quali radio, televisione, stampa per informarsi, forse dimenticando che le normative stringenti applicate a questi ultimi non riguardano invece i primi che sono dunque liberi di pubblicare qualsiasi notizia senza controllo delle fonti[56]. Il rapporto Censis sulla comunicazione 2022 illustra che, a fronte

---

[51] Fasan, 2019, p.113
[52] "Tutti hanno diritto di manifestare liberamente il proprio pensiero con la parola, lo scritto e ogni altro mezzo di diffusione". La libertà di manifestazione del pensiero secondo gli interpreti ha due accezioni, una esplicita di diritto a esprimere il proprio pensiero e una implicita di libertà di ricevere informazioni. L'interpretazione del principio nella sua accezione passiva è sostenuto dalla logica, infatti qualora non vi fosse libertà di ricevere informazioni, verrebbe meno il diritto a manifestare e diffondere il pensiero, dunque il principio originario dell'articolo. Vernice, 2021, p. 90 , Borrello, 2017, p.70
[53] Bianca, 2019, p. 45
[54] Vernice, 2021, p.94
[55] *Ivi*, 2021, p. 96
[56] *Ivi*, 2021, p. 93

di un 51.2% di utenti che dichiara di informarsi guardando il telegiornale, vi è un 35.2% che si informa su Facebook, un 23.4% su motori di ricerca e un 19.3% su siti web d'informazione[57]. Si tratta di dati estremamente allarmanti considerate le premesse sopra fatte.

Andando oltreoceano vi sono numerosi esempi del danno subito dalla democrazia a causa dei social media e, nello specifico, delle bolle filtro. Le elezioni statunitensi del 2016 rappresentano un ottimo punto di partenza per comprendere l'impatto nel mondo fisico delle filter bubble. Di rilievo è il caso *Pizzagate*, poco tempo prima delle elezioni presidenziali statunitensi, tra le fila dell'estrema destra iniziò a diffondersi una teoria secondo cui vari democratici, inclusa la candidata alle presidenziali Hillary Clinton, erano coinvolti in un traffico di minori, e la presunta base operativa del complotto era una pizzeria della capitale. Il 4 dicembre 2016 un ventottenne venne arrestato per aver assaltato a mano armata tale pizzeria, alla polizia dichiarò che il suo obiettivo era investigare sul *Pizzagate* e salvare i bambini presunte vittime del progetto segreto dei democratici[58]. L'anno seguente diventò virale una teoria diffusa sul sito 4chan[59] dall'anonimo Q, secondo la quale il neoeletto presidente Trump stesse lavorando ad una missione segreta per smantellare una élite di miliardari, tra i cui membri svettava la famiglia Clinton, che clandestinamente controllava il paese, il cosiddetto *deep state*. La teoria, nota come *QAnon*, iniziò a diffondersi capillarmente in tutto il web raggiungendo un pubblico più vasto rispetto alla sola estrema destra, riunendo tutti i fautori di quelle teorie che si fondano su razzismo, supremazia bianca, omofobia, e cospirazionismo. Nel 2020 tutti i profili e gruppi legati al movimento vennero rimossi dai principali social, perché considerati potenziale minaccia terroristica in grado di istigare alla violenza per la loro abilità di fomentare panico e terrore tra i propri seguaci[60]. Tuttavia questo non fermò il movimento, che continuò ad acquisire consensi all'interno di social come Reddit e Parler[61], caratterizzati da linee guide meno rigide rispetto a Twitter o Facebook. Il 6 gennaio 2021, un numeroso gruppo di repubblicani e di supporter di Trump e delle varie teorie sopra menzionate fece irruzione all'interno del Campidoglio statunitense, convinto

[57] Censis, 2022
[58] Metropolitan Police Department, 2016
[59] Un sito web dove ognuno può postare e condividere immagini. Poi evolutosi in 8chan, chiuso a sua volta nel 2019 dopo aver presumibilmente ospitato un post che annunciava la sparatoria avvenuta ad El Paso con 23 vittime https://edition.cnn.com/2019/08/04/business/el-paso-shooting-8chan-biz/index.html. Il sito è stato poi riaperto con il nome di 8kun.
[60] Treccani, *QAnon*
[61] Cuono, 2021

che la vittoria di Joe Biden fosse una frode. Fu Trump stesso attraverso vari tweet ad incoraggiare la partecipazione al suo noto raduno del 6 gennaio, occasione nella quale invitò la folla a combattere per la patria aggiungendo che in caso contrario "*you're not going to have a country anymore*". La folla si sentì dunque giustificata a compiere gli atti noti a tutti[62]. Limitandosi a questo breve resoconto di fatti interessanti, va ora compreso il nesso che lega questi al concetto di *bubble democracy*.

La creazione di bolle di filtraggio ed echo chamber implica che, soggetti particolarmente vulnerabili, una volta entrati in contatto con teorie estremiste e cospiratorie come *QAnon* o il *pizzagate*, difficilmente saranno in grado di uscirne, considerata l'inondazione di contenuti generati dall'algoritmo che confermano i loro pensieri. Ogni utente è tenuto a credere che la propria verità sia l'unica, tuttavia non è in grado di spiegarla a chi la pensa diversamente. La tendenza crescente è la perdita di tolleranza nei confronti di soggetti che la pensano diversamente, ciascuna ideologia difficilmente riesce a trovare possibilità di confronto e dialogo. Qualora nel web accada che due utenti con ideologie contrapposte vengano in contatto, ecco che invece che l'instaurazione di un dialogo costruttivo si manifestano scontri e offese da parte di entrambe le parti[63]. L'odio e la violenza verbale vincono sul confronto di idee.

Non è solo l'estrema polarizzazione indotta dagli algoritmi il pericolo di una democrazia sempre più debole, la grande mole di dati a disposizione potrebbe essere utilizzata dai politici stessi per influenzare il voto dei cittadini quando si è in prossimità di elezioni[64]. Si pensi allo scandalo di Cambridge Analytica, società che offriva servizi di consulenza alle campagne elettorali. Cambridge Analytica era in grado di effettuare analisi predittive e comportamentali basate su big data e data mining[65] di migliaia di utenti. L'algoritmo utilizzato permetteva di individuare la personalità di ciascun utente attraverso un processo di microtargeting. La società utilizzò i servizi di Amazon Mechanical Turk, che offriva agli utenti Facebook la possibilità di partecipare a sondaggi retribuiti online. Al termine del sondaggio l'utente era tenuto a connettere il proprio profilo Facebook al sito. Tutti i dati del profilo social dell'utente, inclusi i dati degli amici, furono a disposizione della società, che li utilizzò per inviare contenuti mirati agli elettori

[62] Britannica, *January 6 U.S. Capitol attack*
[63] Costanzo, 2019, p. 76
[64] *Ivi*, 2019, p.77
[65] Tecnica di elaborazione di grandi quantità di dati per estrarre informazioni.

in grado di influenzare il loro voto[66]. Cambridge Analytica è stata società di consulenza per la campagna presidenziale di Trump del 2016, e probabilmente interferì anche nel referendum Brexit a favore del '*leave*'[67]. Considerati gli effetti della profilazione e delle bolle filtro sulla sicurezza pubblica e sul voto degli elettori, è necessario che vi sia una regolamentazione più efficace rispetto al GDPR, che tuteli gli individui contro la profilazione, e la disinformazione. È altresì indispensabile un controllo dei gruppi estremisti e radicali che diffondono teorie in rete. La consapevolezza dell'effettiva esistenza del fenomeno delle filter bubble nelle nostre vite può essere un punto di partenza per ridurne gli effetti, si parla a tal proposito di *media literacy*.

Nel prossimo capitolo verrà affrontato nel dettaglio il fenomeno della disinformazione online. In particolare si analizzeranno il funzionamento e gli effetti della nuova tecnologia di IA, il *deep learning*, che è in grado di generare artificialmente immagini, video, testi, audio, estremamente realistici che, se diffusi in rete senza regolamentazione sono idonei a creare disinformazione e fomentare tutto l'universo di violenza, odio e cospirazione presentato in questo capitolo.

[66] Federal Trade Commission, 2019, Boldyreva, 2018
[67] Le indagini sono ancora in corso. Risso, 2017

CAPITOLO II

IL DEEP LEARNING: UN NUOVO MODO DI PRODURRE DISINFORMAZIONE

2.1 La disinformazione online: definizioni, origine ed esempi

La disinformazione costituisce oggi una tra le principali minacce alla democrazia e alla sicurezza di una società. Questo in ragione della sua idoneità a minare la fiducia degli individui nei confronti dei media e delle istituzioni, ad influenzare i risultati elettorali, come accaduto negli Stati Uniti nel 2016, ad ostacolare l'esercizio della libertà di espressione e l'assunzione da parte dei cittadini di decisioni informate[68]. L'aspetto definitorio del fenomeno risulta in questo contesto fondamentale; spesso termini come disinformazione, misinformazione, malainformazione vengono adottati in maniera intercambiabile, quando al contrario ogni espressione possiede la propria definizione[69].

La letteratura predominante definisce la misinformazione un'informazione falsa, diffusa senza l'intento di ingannare[70]; si tratta di quel comportamento eseguito da coloro che condividono con leggerezza in rete un contenuto senza verificarne fonte e veridicità[71]. La malainformazione, dall'altro lato, consiste in informazioni corrette, ma rappresentate in modo tale da veicolare messaggi falsi o manipolatori[72]. Un esempio di malainformazione riguarda la pratica del *greenwashing*, strategia di marketing adottata da molte aziende, che facendo leva sui presunti, non documentati, benefici e bassi impatti ambientali dei loro prodotti, ingannano i consumatori[73]. Infine la disinformazione consiste nella divulgazione di contenuti falsi con l'intento volontario di ingannare il

---

[68] Commissione Europea, *Un codice di condotta dell'UE più rigoroso sulla disinformazione*

[69] Tuttavia i tre termini sono strettamente legati e congiuntamente rientrano nella definizione più ampia di "disordine informativo"

[70] Shu, 2020, p. 2

[71] Matucci, 2018, p. 11

[72] Khan, 2022, p. 130

[73] Un esempio di questa pratica ha coinvolto la compagnia aerea KLM, accusata dalla ONG Fossilfree di aver utilizzato alcuni slogan come "Fly Sustainable" all'interno delle campagne pubblicitarie per far sembrare sostenibili i propri voli e dunque ingannare il consumatore. La Corte di Amsterdam nel marzo 2023 ha accolto le ragioni della ONG sostenendo che le asserzioni degli slogan erano troppo vaghe e generali in relazione ai benefici ambientali e dunque idonee ad ingannare il consumatore che potrebbe credere che volare con KLM sia sostenibile. Per una traduzione della sentenza si veda: https://www.clientearth.org/media/cx4po41h/klm-judgment-20-march-2024.pdf; Parlamento UE, *Fermare il greenwashing: come l'UE regola le asserzioni ambientali*

destinatario[74], screditare opponenti politici, influenzare elezioni, creare confusione e generare conflitti sociali[75]. La presenza nella nostra quotidianità di simili fenomeni è estremamente diffusa e per questo dannosa, si consideri che il World Economic Forum ha inserito misinformazione e disinformazione al primo posto nel ranking 2024 dei rischi globali per il breve termine. L'accesso a tecnologie avanzate e nuovi sistemi di IA, oltre che la decrescente fiducia nei confronti delle istituzioni e dei media tradizionali, permetteranno difatti, a chiunque ne abbia un interesse, di contribuire alla polarizzazione sociale e alla violenza ideologica[76], provocando tutte le possibili conseguenze di cui si è già discusso nel capitolo precedente.

Per quanto concerne il termine *fake news*, espressione divenuta *buzzword*[77]dei media e dunque onnipresente quando si parla di contenuti falsi e ingannevoli, la letteratura si è divisa nel fornire una definizione. Vi è chi considera il termine fake news un sinonimo di disinformazione[78] e chi una sottocategoria di essa[79]. Accogliendo qui quest'ultima idea, si definisce la fake news come una forma di disinformazione online caratterizzata da un contenuto falso o ingannevole, rappresentato con un design credibile, assimilabile a una notizia, diffuso con l'obiettivo di manipolare e ingannare una collettività[80]. Più in generale "articoli recanti notizie che sono intenzionalmente false e verificabili e che potrebbero trarre in inganno il lettore"[81]. All'interno di questo capitolo si andrà a inquadrare il fenomeno della disinformazione in generale; le fake news verranno citate solo come esempi di notizie a contenuto disinformativo.

Le ragioni che spingono un soggetto a credere ad una disinformazione sono prettamente legate all'emotività dell'individuo stesso. Uno stato di incertezza[82], pressione emotiva, ansia[83] sono situazioni nelle quali il soggetto non ha la capacità di controllare le

---

[74] Shu, 2020, p. 2
[75] Bennett, 2021, p. 3
[76] World Economic Forum, 2024, p.18
[77] Termine ambiguo che muta significato in base ai contesti di utilizzo. Baptista, 2022, p.634
[78] Shu, 2020, p. 2
[79] Baptista, 2022, p. 634
[80] *Ivi*, 2022, p. 640
[81] Vernice, 2021, p. 96
[82] Gli studi dimostrano che nelle situazioni di incertezza provocate da disastri naturali, eventi politici e crisi la disinformazione è molto frequente. L'utente che necessita di informazioni che le fonti ufficiali non sono in grado di provvedere tempestivamente, tende a rivolgersi al web, ai social media, ai blog, luoghi dove chiunque ha la possibilità di riempire il vuoto informativo con il proprio giudizio personale. Shu, 2020 p. 7
[83] La disinformazione viene spesso generata e diffusa dagli individui per dare un significato o giustificazione a determinate circostanze idonee a procurare ansia o tensione nella collettività. Nel 2017 a seguito dell'uragano Harvey, più di un milione di utenti su Facebook vennero a contatto e condivisero notizie che accusavano il movimento Black Lives Matter della lentezza dei soccorsi. Shu, 2020, p. 7

proprie azioni e spinto dalla vulnerabilità del momento è indotto a credere che quello che legge sia vero. Altro fattore è, come trattato nel capitolo precedente, il *confirmation bias*. Il soggetto che si trova davanti due contenuti relativi allo stesso tema, di cui il primo veritiero e il secondo disinformativo, tende a credere alla notizia che presenta una visione più affine alle sue attitudini. In aggiunta a ciò, l'individuo, come essere sociale che necessita di approvazione e di sentirsi parte di un gruppo, tende a condividere la notizia, che si diffonderà fino a raggiungere un ampio pubblico e a consolidarsi all'interno delle filter bubbles. È chiaro che, se la notizia affine all'ideologia dell'utente è quella falsa, l'impatto che può generarsi può avere gravi conseguenze[84].

A differenza di quanto si possa credere, la disinformazione non ha avuto origine con l'ascesa di Internet, la rete e in particolar modo i social media hanno certamente incrementato gli effetti del fenomeno, ma di esempi di informazioni false create intenzionalmente allo scopo di ingannare determinati soggetti ne è ricca la storia. Nel 1053 papa Leone IX utilizzò un presunto editto di Costantino del 315, la "Donazione di Costantino", per legittimare la sovranità del Papato su tutte le chiese del mondo e sull'Impero, ma l'editto nel XV sec. risultò essere un falso[85]. Nel 1898 la nave militare *USS Maine* affondò a Cuba a causa di una esplosione, l'articolo del *New York Journal* che accusava la Spagna dell'attacco fu una delle cause scatenanti la guerra ispano-americana che poi si concluse con l'indipendenza di Cuba[86]. Nonostante ciò, è evidente che l'intermediazione dei social networks e delle piattaforme nella veicolazione di informazioni e notizie ha assunto oggi il ruolo di cassa di risonanza per la disinformazione[87], in particolare grazie alla rapidità e precisione nel raggiungere i soggetti target, coloro che tenderanno a ricondividere il contenuto[88]. Si è già citato il caso *QAnon* e *pizzagate*, ma vi sono innumerevoli esempi di disinformazione online. Nel 2017, in seguito all'attentato di fronte al Parlamento di Londra, divenne virale una foto che riprendeva una ragazza con *l'hijab* intenta a passare accanto alle vittime, fissando il proprio smartphone, apparentemente indifferente all'orrore accaduto. In seguito si è scoperto che il primo profilo Twitter a diffondere la foto faceva parte di una serie di

[84] Shu, 2020, p.8
[85] Treccani, *Donazione di Costantino*
[86] Britannica, *Destruction of the Maine*
[87] Monti, 2017, p. 83
[88] Comunicazione della Commissione al Parlamento Europeo, al Consiglio, al Comitato Economico e Sociale Europeo e al Comitato delle Regioni. *Contrastare la disinformazione online: un approccio europeo*, 2018, p. 1

account falsi, attenzionati dal Congresso statunitense nell'inchiesta sulle interferenze della propaganda russa durante le elezioni americane. La foto era reale, ma l'account l'ha diffusa in maniera decontestualizzata, accostandole una descrizione in grado di creare polarizzazione e diffondere odio nei confronti dei soggetti di fede islamica. La foto venne poi utilizzata nei social e all'interno di siti di *fake news*[89] per dare maggior credibilità a teorie xenofobe[90].

L'IA rappresenta in questo momento il maggior pericolo per le potenziali vittime di disinformazione, può essere infatti utilizzata per produrre contenuti ingannevoli o falsi estremamente realistici e convincenti. Vi sono molteplici siti gestiti completamente o quasi da IA, senza supervisione umana sui contenuti pubblicati, si parla di *Unreliable AI-Generated News and information websites (*UAINS*)*[91]. L'IA viene impiegata anche per la creazione di *social bots*, dei programmi in grado di imitare il comportamento degli utenti all'interno dei social networks, e dunque capaci di postare, interagire con altri utenti e altri bots, mettere like o condividere contenuti. Il problema di questo tipo di programma è che ad un costo piuttosto contenuto chiunque è in grado di mettere in circolazione account falsi con il fine di accelerare la diffusione di disordini informativi[92]. Vi è poi il fenomeno del *deep fake*, che per la sua rilevanza tratteremo singolarmente nel successivo paragrafo.

La disinformazione rappresenta dunque un serio pericolo per la democrazia di un paese, e lo sarà sempre di più nei prossimi anni quando l'IA sarà a disposizione delle masse. Nel 2022, aziende, organizzazioni e associazioni del settore tra cui Adobe, Meta, Microsoft e TikTok[93] si sono riunite e, sulla base delle linee guida fornite dalla Commissione Europea, hanno redatto un Codice di condotta rafforzato sulla disinformazione, revisione del precedente Codice del 2018. Si tratta di una auto-

[89] I principali siti di fact checking contengono al loro interno una sezione chiamata black list che include un elenco di siti web particolarmente attivi nella produzione di disinformazione ed in particolare di fake news, alcuni esempi sono: Sa Defenza, Voxnews, Open Your Eyes, Informare senza censure. Vi sono poi svariati siti che non producono notizie ma copiano e incollano articoli senza verificare l'autenticità delle fonti: TGNewsItalia, Maurizioblondet.

[90] Macagnone, 2017

[91] Si tratta di siti i cui contenuti vengono pubblicati dall'IA senza una significativa supervisione umana e senza informare il lettore di ciò. Il modo in cui il sito si presenta potrebbe indurre il lettore medio a credere che i suoi contenuti siano prodotti da autori o giornalisti umani.

[92] Shu, 2020, p. 8

[93] Commissione Europea. *Signatories of the 2022 Strengthened Code of Practice on Disinformation*.

regolamentazione degli operatori del settore; il Codice contiene 44 impegni con relative misure da rispettare per combattere la disinformazione. Tra questi figurano ad esempio l'impegno a demonetizzare i contenuti pubblicitari inseriti accanto a prodotti disinformativi, ma anche a garantire trasparenza nella pubblicità politica, etichettando gli annunci politici in modo da informare l'utente del tipo di contenuto incontrato. Vi è poi l'impegno a garantire l'integrità dei servizi per ridurre comportamenti manipolativi quali creazione di profili falsi, bot, *deepfake* dannosi, pubblicità non trasparenti da parte di influencers o disinformazione. Altro impegno prevede che gli operatori che sviluppano, utilizzano, diffondono contenuti generati o manipolati da IA prendano in considerazione gli obblighi di trasparenza e la lista di pratiche manipolative proibite dal nuovo Regolamento sull'IA *ex*. art.5[94].

Sembra dunque che vi sia un interesse da parte delle istituzioni, fornitori di servizi online e piattaforme di ridurre la disinformazione. Tuttavia, nonostante le misure e strategie adottate, il fenomeno è lontano dall'essere arginato. La diffusione di *deepfake* appare essere l'ultima frontiera della disinformazione, basata su modelli di IA in continua evoluzione e perfezionamento e sempre più nel futuro prossimo a disposizione della collettività. Il successivo paragrafo si occupa di inquadrare questo nuovo tipo di contenuti e la tecnologia su cui essi si fondano.

[94] Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2024; Codice di Condotta rafforzato sulla disinformazione, 2022.

2.2 Disinformazione iperrealistica: deep learning e deepfake

I contenuti *deepfake* sono la nuova forma di manipolazione audiovisiva in circolazione, caratterizzata da un maggior livello di realisticità del prodotto e, per il momento, da una minor accessibilità al grande pubblico; per ottenere un contenuto estremamente realistico sono necessarie risorse tecnologiche e competenze. Accanto ai deepfake vi sono i *cheap fake*, contenuti generati da software differenti, meno sofisticati, spesso gratuiti e facili da utilizzare e dunque a disposizione di chiunque, Adobe Premiere o FakeApp ne sono esempi. I cheap fake vanno dai filtri di Snapchat ai video o foto pornografici editati[95], dal photoshopping ai video decelerati o accelerati[96]. Per il momento rimane alta la barriera tecnica e computazionale necessaria per la produzione di video falsi estremamente realistici, tuttavia considerando la costante e rapida evoluzione delle tecnologie di IA, vi è un'urgente bisogno di controllo e moderazione da parte delle aziende del settore e delle istituzioni, per evitare che il fenomeno si sviluppi in maniera incontrastata.

Il termine deepfake è divenuto popolare negli ultimi cinque anni per descrivere quei prodotti digitali creati dall'IA attraverso tecniche di deep learning. Il neologismo deepfake risulta infatti dalla fusione dei termini deep learning e fake[97]. Si potrebbe definire il deepfake il risultato di processi di machine learning in grado di "unire, combinare, sostituire e imporre immagini e video clip per creare video falsi che appaiono autentici"[98], oltre che manipolare i movimenti delle labbra di un soggetto, oppure clonare voci. In generale un individuo può risultare dire o fare cose in video, immagini o audio che egli non ha mai fatto. In contesti creativi e artistici come la cinematografia o le pubblicità, essa rappresenta una tecnologia particolarmente utile per ringiovanire, invecchiare o sostituire attori o per rappresentare attori defunti. Bruce Willis, a titolo di esempio, è divenuto protagonista di uno spot pubblicitario senza doversi presentare alle

---

[95] La giornalista Rana Ayyub, dopo aver per anni criticato l'abuso di potere del partito nazionalista Bharatiya Janata (BJP) indiano, nel 2018 scoprì che un account supporter del partito aveva messo in circolazione su Whatsapp un video pornografico in cui lei appariva protagonista, con il fine di screditare la sua immagine e credibilità. Il volto della protagonista del video era stato editato, tuttavia risultava sufficientemente credibile da convincere chi lo riceveva a ricondividerlo. Fu una vera e propria campagna di misinformazione. Ayyub, 2017

[96] Nel 2019 diventò virale un video raffigurante Nancy Pelosi, in apparenza sotto l'effetto di alcool, criticare Donald Trump. Il video originale è stato in realtà decelerato del 75% per generare l'effetto vocale desiderato. Sadiq, 2019.

[97] GPDP, *Deepfake Vademecum: Deepfake Il falso che ti «ruba» la faccia (e la privacy),* 2020. p.1

[98] Cover, 2022, p. 611

riprese. Deepcake, azienda che fornisce servizi di machine learning e big data, ha creato un deepfake dell'attore che è stato poi utilizzato in uno spot pubblicitario per la compagnia telefonica russa Megafon. Un attore russo ha recitato nello spot e in una fase successiva il suo volto è stato sostituito con il volto di Bruce Willis[99].

Tuttavia non sempre vi è consenso da parte degli interessati, l'attore Tom Hanks ha scoperto che la sua immagine era stata utilizzata per generare un suo deepfake, inserito poi in uno spot pubblicitario relativo ad una offerta per un piano dentale[100]. Episodio simile ha coinvolto l'attrice Scarlett Johansson, convinta che l'azienda OpenAI avesse impiegato la sua voce per la voce virtuale di ChatGpT[101]. Ma non è solo questo il rischio di questo tipo di contenuti. Vi è difatti la possibilità che essi vengano utilizzati a scopo disinformativo o misinformativo. L'origine del termine deepfake si deve ad un utente di Reddit, cui username era per l'appunto "*deepfakes*", nel 2017 pubblicò un video a contenuto pornografico il cui volto della protagonista era stato sostituito con il volto di Gal Gadot, l'attrice che interpreta Wonder Woman nell'omonimo film. Questo fu il primo di una lunga serie di video virali dal contenuto simile che hanno visto come vittime altre celebrità. Una delle conseguenze più gravi della creazione di deepfake è dunque legato all'abuso non consensuale dell'immagine delle donne, si parla dei cosiddetti *deepnude*[102].

Come anticipato, il deepfake è frutto di una tecnica di apprendimento chiamata *deep learning*, si tratta di una branca del machine learning fondata, oltre che su matematica e scienze informatiche, anche sulla neuroscienza[103]. I contenuti creati mediante deep learning hanno la caratteristica di apparire estremamente realistici e dunque idonei ad ingannare il destinatario che non è in grado di realizzare di avere a che fare con un contenuto generato artificialmente. Il sito *thispersondoesntexist.com* è un esempio interessante di impiego di tecnologia deep learning, l'algoritmo utilizzato, basato su reti generative avversarie (GANs), genera ad ogni accesso un volto estremamente realistico ma non appartenente a persone reali[104]. Il deep learning è un modello di

---

[99] Derico, 2022

[100] Turrini, D., 2023

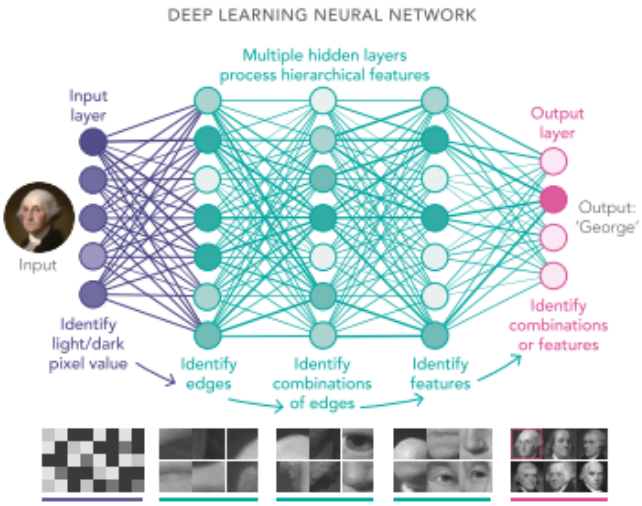
[101] Per maggiori dettagli: https://www.repubblica.it/tecnologia/2024/05/23/news/openai_non_ha_rubato_voce_scarlett_johansson-423091593/. Per la dichiarazione di OpenAI: https://openai.com/index/how-the-voices-for-chatgpt-were-chosen/

[102] Paris, 2019, p.5

[103] Sejnowski, 2018, p.3

[104] Disponibile su: https://www.thispersondoesnotexist.com

apprendimento “profondo” impiegato in molteplici programmi. Google adotta questa tecnologia in un centinaio di suo servizi tra cui Street View e Google Translate. Ma anche il settore delle Self driven cars si basa su questo modello, come anche i servizi di riconoscimento vocale e il settore della computer vision. Il deep learning impiega reti neurali profonde; si tratta di una rete di numerosi layer (strati) connessi tra loro, di cui solo il primo e l’ultimo visibili, ecco perché deep, che performano ognuno analisi differenti. Si parla di reti neurali proprio per la similitudine concettuale che vi è con il funzionamento del cervello umano. Ogni layer può stimolare o inibire i layer collegati, e ognuno possiede peso e influenza differenti.



Un esempio di rete neurale profonda
(Immagine di Lucy Reading-Ikkanda, da Waldrop, 2019, p. 1075)

La caratteristica del deep learning è la capacità di elaborare grandi quantità di dati senza intervento del programmatore, la rete neurale infatti viene istruita attraverso l’esposizione a molteplici esempi. Essa autonomamente regola i pesi di ciascun layer, fondamentali per inibire o stimolare gli altri ad essi collegati, per perfezionare i risultati prodotti in output. Dunque quando il sistema produce un segnale errato, ad esempio riconoscendo una lince dall’immagine di un gatto, significa che almeno un layer è basato su parametri non corretti, accettando segnali che andrebbero rifiutati o viceversa. Il sistema individua il layer e lo istruisce a invertire i responsi delle sue classificazioni diminuendo il peso di quella connessione[105]. Nel riconoscimento facciale, ad esempio, gli strati più bassi prendono e valutano i dati grezzi immessi in input, come i pixel di

[105] Hayes, 2014, p. 186 ss

un'immagine. I layer successivi si occupano poi di riconoscere angoli, curve, orientamenti. Gli strati superiori riescono ad analizzare elementi più complessi, riconoscendo ad esempio nasi, occhi e relazioni spaziali tra essi. Infine l'ultimo strato fornisce l'output richiesto dall'utente.

Questa tecnologia non è perfetta, si tratta di un sistema innanzitutto opaco, ciò che accade nei layer nascosti non è chiaro, e dunque, potrebbe essere impossibile concepire le ragioni della generazione di un determinato contenuto. Inoltre va ricordato che la macchina non è in grado di comprendere il significato di ciò che genera[106], come vedremo nel prossimo paragrafo, se si chiede alla macchina di produrre un video che raffigura una signora soffiare sulle candeline della sua torta di compleanno, il sistema, che non è in grado di comprendere i processi causa-effetto del mondo reale, mostrerà la candelina sempre accesa, anche dopo averci soffiato sopra.

Nonostante i difetti, questa tecnologia rimane in grado di produrre ingenti danni, soprattutto se impiegata per generare contenuti disinformativi, per distorcere una memoria collettiva e soprattutto per danneggiare la riservatezza, dignità e immagine degli individui[107]. La preoccupazione sociale invita ad una regolamentazione rapida, sono sufficienti un migliaio di nostre foto, salvate sul cloud, su qualche dispositivo o pubblicate sul proprio profilo social per permettere a soggetti con cattive intenzioni di produrre contenuti dannosi. Su Youtube è stato caricato un video molto popolare che riprende l'ex Presidente USA Barack Obama pronunciare un discorso sulla pericolosità dei deepfake, "*we're entering an era in which our enemies can make it look like anyone is saying anything at any point in time. Even if they would never say those things.*" il video sembra reale, in realtà le parole sono state pronunciate dall'attore e comico Jordan Peele che ha ceduto i suoi movimenti facciali alle caratteristiche del volto e della voce di Obama mediante deep learning, si tratta dunque di un video deepfake utilizzato per sensibilizzare la collettività sui pericoli di questa tecnologia[108]. La rapidità di diffusione dei deepfake, che la rete per le sue caratteristiche intrinseche agevola, impedisce interventi immediati di fact checking o moderazione. Inoltre bisogna considerare che in alcune applicazioni come Whatsapp, dove i messaggi sono criptati e dunque circolano lungo reti private, contenuti virali potrebbero rimanere nascosti, impedendo ancora di più moderazione o

[106] Waldrop, 2019, p. 1074 ss
[107] Cover, 2022, p. 610
[108] Disponibile su: https://www.youtube.com/watch?v=cQ54GDm1eL0

copertura mediatica[109]. Il prossimo paragrafo andrà a presentare alcuni esempi di popolari modelli generativi di testi, immagini o video recentemente implementati o ancora in fase di sviluppo.

[109] Paris, 2019, p. 8

2.3 I principali modelli generativi di IA

Il campo dell'IA generativa comprende tutti quei modelli che attraverso tecniche di deep learning e addestramento su grandi quantità di dati, producono output in risposta ai prompts[110] immessi dall'utente. Questo paragrafo si occuperà di presentare alcuni esempi concreti di modelli di IA generativa attualmente diffusi o in fase di sviluppo.

Tra le aziende leader nella ricerca nel campo dell'IA vi è OpenAI, fondata nel 2015 con l'obiettivo di sviluppare nuove tecnologie di intelligenza artificiale a beneficio dell'umanità. Il principale investitore dell'azienda è Microsoft che ha potuto usufruire in maniera esclusiva delle tecnologie di OpenAI implementandole all'interno del pacchetto Microsoft365. OpenAI è divenuta particolarmente nota nel 2022 per aver reso disponibile alla collettività ChatGPT, un modello di linguaggio naturale in grado di produrre testi coerenti in risposta agli input dell'utente. L'architettura del modello è basata su un *Generative Pre-trained Transformer* (GPT per l'appunto), un tipo di rete neurale profonda che durante la fase di addestramento, mediante grandi quantità di dati, viene allenata a comprendere e generare il linguaggio naturale dell'uomo. ChatGPT può essere utilizzata per i più diversi scopi, dalla traduzione di testi o pianificazione di itinerari di viaggio, alla stesura di codici sorgente o di piani di allenamento[111].

L'azienda californiana sta attualmente implementando un nuovo modello, evoluzione di GPT e delle successive versioni, chiamato GPT-4. Si tratta di un sistema più avanzato, che genera, modifica, interagisce con l'utente ed è in grado di comporre canzoni, scrivere sceneggiature o fare riassunti di libri, ad esempio. Particolarità di questa versione è la possibilità di inserire come input immagini, da queste esso può restituire responsi, descrizioni, analisi e classificazioni. Da una fotografia di alcuni ingredienti può produrre ad esempio una lista di ricette realizzabili. GPT-4 è stato implementato all'interno dell'app Duolingo per l'apprendimento delle lingue straniere, nella sua versione in sottoscrizione: l'utente ha la possibilità di conversare con il *chatbot* nella lingua che sta apprendendo in maniera immersiva[112]. Funzionalità simili al modello linguistico di OpenAI le possiede *Gemini* di Google DeepMind, azienda di ricerca nel

[110] Input sotto forma di comando prodotto dall'utente
[111] Disponibile su: https://openai.com/chatgpt
[112] Disponibile su: https://openai.com/gpt-4

campo IA fondata nel 2023 dalla fusione tra Google Brain[113] e DeepMind[114]. Il vantaggio principale di Gemini rispetto a GPT è la possibilità di fornire output aggiornati, si potrebbe ad esempio ottenere in output un elenco dei programmi che trasmetteranno stasera in televisione[115], cosa invece non possibile per GPT che può esaminare solamente i dati ricevuti nelle fasi di apprendimento.

Per quanto riguarda la generazione di immagini, OpenAI ha creato DALL-E, un sistema che produce immagini estremamente fedeli rispetto alla didascalia testuale immessa dall'utente. L'ultima versione, DALL-E 3, è capace di apportare modifiche a immagini preesistenti, rispondere in maniera più accurata ai prompts e in generale tradurre le idee dell'utente in immagini. Il modello è al momento disponibile gratuitamente all'interno del chatbot Copilot di Microsoft. Considerati i rischi di un utilizzo dannoso di questi modelli, OpenAI sta lavorando per limitare la generazione di contenuto violento, esplicito o d'odio e per ridurre tentativi di misinformazione o propaganda e l'impiego di immagini relative a personaggi famosi, non solo all'interno di DALL-E ma in tutti i suoi modelli[116].

Per quanto concerne la generazione di video, Sora è il nuovo e più recente strumento realizzato. A partire da un prompt testuale realizza video di massimo un minuto realistici o immaginari. Sora è ad oggi ancora in fase di sperimentazione, tuttavia OpenAI ha condiviso in anteprima alcuni esempi di creazioni prodotte dal modello. Al prompt: *"Tour of an art gallery with many beautiful works of art in different styles*", il modello è riuscito a generare un video estremamente credibile, girato all'interno di una galleria d'arte non esistente, con opere esposte alle pareti interamente frutto dell'IA. I principali difetti di Sora vengono sottolineati dall'azienda stessa e riguardano il non rispetto delle leggi della fisica, problemi di traiettoria della camera, e mancata capacità di rispettare i rapporti causa-effetto, come nel video della signora anziana che soffia sulle candeline della sua torta di compleanno, senza spegnerle. Attualmente OpenAI sta collaborando con red teams (gruppi indipendenti che studiano le debolezze di un programma per

[113] Nel 2011 Google ha avviato un progetto di ricerca, denominato Google Brain, implementando una rete con milioni di neuroni e miliardi di connessioni.
[114] DeepMind è un azienda di IA fondata nel 2010 e improntata sulla ricerca in campo di neuroscienze, ingegneria, matematica, machine learning. Nel 2015 diffonde un nuovo sistema di IA, AlphaGo, in grado di battere il campione mondiale del complicato gioco da tavolo Go. Disponibile su: https://deepmind.google/technologies/alphago/
[115] Disponibile su: https://gemini.google.com/faq
[116] Disponibile su: https://openai.com/safety

migliorarne l'efficacia) che testeranno il modello contro attacchi di misinformazione, contenuti d'odio, e bias e implementeranno strumenti per individuare contenuti ingannevoli ed etichettare gli output di Sora, affinché l'utente riconosca subito dall'etichetta che un determinato video non è reale[117].

La possibilità di generare mediante modelli di deep learning e reti neurali contenuti completamente artificiali, comporta il pericolo di diffusione di deepfake sempre più realistici e capaci di manipolare il pensiero e la memoria collettiva. Il Garante per la Protezione dei Dati Personali (GPDP) ha pubblicato nel dicembre 2020 una informativa di sensibilizzazione sul tema deepfake, ponendo in risalto le tipologie di illeciti che possono derivare dal loro impiego e concludendo con alcuni suggerimenti per la collettività. I *deepnude* rappresentano la minaccia più grave, si tratta di contenuti video o immagine deepfake che rappresentano individui svestiti o in posizioni e situazioni compromettenti. Mentre come già accennato, nei primi post diffusi in rete le vittime del fenomeno erano personaggi noti, oggi chiunque può divenirne vittima soprattutto a scopo di *revenge porn*[118]. Altro potenziale impiego è per commettere atti di cyberbullismo, o per diffondere fake news, specialmente a tema politico. Infine il GPDP cita lo *spoofing*[119], il *pishing* e il *ramsoware*, considerando che i sistemi di sicurezza che si basano su dati biometrici come volti o voce possono essere facilmente ingannati da volti e voci artefatte.

Sarà importante sempre più nel prossimo futuro tutelarsi dal fenomeno, la tecnologia è in costante evoluzione e non ci vorrà molto prima che i difetti dell'IA generativa spariscano e gli output divengano completamente indistinguibili da un contenuto reale. Nonostante le precauzioni adottate da aziende del settore come OpenAI e Google DeepMind, il rischio che questi tipi di illeciti divengano un problema sostanziale della società moderna è elevato. Il GPDP presenta alcuni suggerimenti pratici che ognuno dovrebbe adottare per tutelarsi. Innanzitutto vi è la responsabilizzazione dei soggetti, è necessario avere cautela nel momento in cui si pubblica una immagine online e nei social networks, l'immagine o video potrebbero infatti essere utilizzati da terzi per la generazione di deepfake. Fondamentale tentare di apprendere le tecniche per riconoscere un deepfake, analizzare la sfocatura, l'innaturalità nei movimenti, le deformazioni, le luci,

---

[117] Disponibile su: https://openai.com/sora

[118] Il revenge porn è il reato previsto *ex* art. 612 ter cp, riguarda la diffusione di immagini o video a contenuto sessualmente esplicito senza il consenso delle persone rappresentate.

[119] Attacco informatico che consiste nella falsificazione di diversi tipi di informazione, ad esempio il mittente di un messaggio, così da trarre in inganno il destinatario. Sicurezza Nazionale, *Spoofing*.

le ombre[120] oltre che l'autorevolezza della fonte o la presenza di una pluralità di fonti[121]. È necessario che chiunque abbia qualche dubbio sulla autenticità di un contenuto non lo condivida anzi lo segnali se la piattaforma utilizzata prevede la possibilità di farlo[122]. Per i casi estremi in cui si ritiene vi sia un reato o violazione della privacy è invece fondamentale contattare le autorità o il GPDP[123].

In questo secondo capitolo è stato approfondito il fenomeno della disinformazione online, particolarmente esteso, preoccupante e strettamente legato alle filter bubbles e ad alcuni eventi drammatici già accaduti come l'assalto al Campidoglio o il pizzagate. Uno degli elementi che agevolerà sempre più la diffusione di disinformazione online è il progresso che si sta attuando nel campo del deep learning, e di conseguenza nella possibilità di produrre deepfake ingannevoli. Sono stati citati alcuni esempi di modelli in fase di sviluppo o già a disposizione della collettività come ChatGPT, e si sono viste le misure adottate come precauzione dalle varie aziende del settore per limitare la generazione mediante IA di contenuti a sfondo d'odio, violento, razzista, sessuale, oltre che alcuni suggerimenti pratici per gli utenti forniti dal GPDP. Il prossimo e conclusivo capitolo sarà dedicato agli strumenti tecnici e normativi adottati negli ultimi anni per contrastare il fenomeno deepfake e più in generale la disinformazione.

[120] Vi sono alcuni siti come https://detectfakes.kellogg.northwestern.edu/ dove chiunque può testare le proprie capacità di riconoscere se un'immagine è generata da IA o reale.
[121] Commissariato di Polizia di Stato, 2024
[122] Si vedano gli Standard della Community di Meta ad esempio, https://transparency.fb.com/it-it/policies/community-standards/?source=https%3A%2F%2Fit-it.facebook.com%2Fcommunitystandards
[123] GPDP, 2020.

CAPITOLO III

MISURE E TENTATIVI DI REGOLAMENTAZIONE PER IL CONTRASTO ALLA DISINFORMAZIONE E AI DEEPFAKE

3.1 Fact-checking e altre misure pratiche per combattere la disinformazione

In questo ultimo capitolo verranno illustrate le principali soluzioni adottate da istituzioni, governi, piattaforme, aziende per combattere la diffusione incontrollata di disinformazione online e di deepfake. Premesso che la disinformazione rappresenta una delle principali minacce ad una società democratica e forse la conseguenza più grave della generazione di filter bubbles, all'interno delle quali contenuti violenti, d'odio, propagandistici, cospirazionisti trovano terreno fertile, risulta necessario che delle misure vengano messe in atto, non solo dalle istituzioni, ma anche da giornalisti, ricercatori e dai fornitori delle piattaforme dove queste hanno maggiori possibilità di diffusione. Infine considerando i deepfake come una delle forme più gravi di disinformazione, si vedrà come le regolamentazioni a livello internazionale si stiano adoperando per reprimere il fenomeno, seppur in maniera non del tutto efficace.

Una soluzione impiegata da giornalisti, istituzioni e piattaforme per combattere la disinformazione online è lo strumento del fact-checking. I fact-checkers sono in genere organizzazioni o gruppi di giornalisti che analizzano le principali notizie di pubblico interesse in circolazione e ne verificano la fondatezza, confutando eventuali fake news. Il fact-checking tuttavia si rivela uno strumento utile esclusivamente per quegli individui che hanno interesse a verificare l'autenticità di un contenuto e dunque a cercare conferme o confutazioni all'interno di questi siti. Tutti i soggetti caratterizzati da basse conoscenze digitali o da bias cognitivi, spesso le vittime principali di disinformazione, difficilmente entreranno in contatto con questo tipo di strumento[124]. Un'ulteriore questione critica da considerare è la neutralità del fact-checker, che non sempre è garantita all'utente, è per questa ragione che l'International Fact-Checking Network (IFCN) riconosce come suoi membri solo le organizzazioni che sottoscrivono e adottano il Codice dei principi. In primo luogo le organizzazioni firmatarie non possono essere controllate da Stato, partiti politici o politici stessi. Inoltre esse devono garantire di analizzare notizie provenienti da

[124] Vasu *et al*., 2018, p. 18

schieramenti politici differenti allo stesso modo, di non apportare giudizi e di fornire le risorse ufficiali utilizzate affinché l'utente possa verificare autonomamente che l'operato del fact-checker sia fondato. Oltre a ciò, i sottoscrittori devono assicurare una trasparenza in relazione alle metodologie impiegate[125]. Nonostante i due difetti sopra citati, il fact-checking può rappresentare uno strumento utile contro la disinformazione, come dimostrato in occasione delle recenti elezioni politiche portoghesi di febbraio 2024. Il Polígrafo, primo e principale sito di fact-checking in Portogallo, durante le prime settimane di campagna elettorale ha pubblicato sul social X post di fact-checking relativi alle affermazioni pronunciate nei dibattiti dai candidati, tali post hanno generato il maggior numero di interazioni nella piattaforma rispetto agli altri account appartenenti a testate giornalistiche come CNN Portugal e partiti politici come Chega dell'estrema destra[126]. Anche l'UE ha dimostrato interesse verso questo tipo di strumento. In occasione delle elezioni Europee di giugno 2024 è stato avviato dallo European Fact-Checking Standards Network un progetto congiunto con oltre 40 organizzazioni europee di Fact-checking e Google News per combattere la disinformazione all'interno dell'UE e favorire la presa di decisioni informate da parte dei cittadini europei durante le elezioni[127]. Il progetto è stato denominato "*Elections 24 Check*" e si occupa oltre che di fact-checking, dunque di verificare l'accuratezza delle dichiarazioni fatte dai politici, di *debunking,* ovvero l'analisi della veridicità delle informazioni in circolazione in internet e nei social media, e di *prebunking*, la condivisione di informazioni volte a prevenire una disinformazione futura[128].

Le piattaforme, per il ruolo centrale che occupano nella società come intermediarie dell'informazione, dovrebbero essere le prime a responsabilizzarsi e adottare misure adeguate "per prevenire, individuare, rimuovere e disabilitare l'accesso a contenuti illegali" e/o dannosi quali possono essere i deepnude, oltre che per tutelare i diritti fondamentali degli utenti della rete[129]. Alcune piattaforme hanno implementato delle tecnologie di auto regolamentazione basate sulle segnalazioni degli utenti o degli

---

[125] IFCN, *The commitments of the code of principles*

[126] Lusa Agência de Notícias de Portugal*, PONTOS ESSENCIAIS Eleições: Desporto ganha no Facebook, política domina Twitter/X – MediaLab*

[127] La fondazione della EFCSN oltre che la redazione del relativo Codice è stata supportata e finanziata dalla Commissione Europea. Il progetto è stato portato a termine nel 2023, dunque da quel momento la rete si finanzia autonomamente. Disponibile su: https://efcsn.com/

[128] Disponibile su: https://elections24.efcsn.com

[129] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions. *Tackling Illegal Content Online. Towards an enhanced responsibility of online platform*, 2017, p. 3

algoritmi. L'esempio classico è la bandierina di Facebook che permette all'utente di segnalare i contenuti da lui ritenuti inappropriati, offensivi o pericolosi, che poi verranno analizzati da alcuni fact-checkers appartenenti alla IFCN. Nel 2021 Facebook collaborava con oltre 90 organizzazioni terze di fact-checking in grado di coprire segnalazioni in circa 60 lingue[130]. All'interno dell'UE, il Regolamento 2022/2065, noto come Digital Services Act (DSA), disciplina intermediari e piattaforme al fine di contrastare la diffusione di contenuti illegali e disinformazione online. *Ex*. art. 16 si fa riferimento ai meccanismi di segnalazione e azione, che devono essere garantiti affinché ciascun utente o ente possa segnalare contenuti da lui ritenuti illegali[131]. Va sottolineato tuttavia che il contenuto disinformativo non necessariamente costituisce un contenuto illegale, rimane perciò a discrezione della piattaforma la decisione di applicare restrizioni sottoforma di banner o alert[132], riduzione della visibilità o rimozione del contenuto[133].

Come già citato quando si è esaminato il Codice rafforzato sulla disinformazione, un'ulteriore metodo contro la diffusione di disinformazione è la riduzione dei compensi agli annunci a contenuto politico. La pubblicità personalizzata di tipo politico viene spesso utilizzata per campagne disinformative o misinformative spesso in grado di condividere idee dannose; disincentivando finanziariamente questo tipo di annunci, certamente ne diminuisce il volume.

La legislazione rappresenta un'ulteriore strumento a disposizione dei governi per guidare l'operato delle piattaforme nel controllo del fenomeno. Un esempio è rappresentato dal Network Enforcement Act, adottato in Germania nel 2017 con il fine di controllare la diffusione di hate speech, radicalizzazione, contenuti illegali e disinformazione online[134]. In Italia si è tentato un approccio simile con il Disegno di

---

[130] Meta, *How Meta's third-party fact-checking program works.* Un altro strumento impiegato da Facebook per riconoscere contenuti d'odio, violenti, politici disinformativi, bots e provvedere alla opportuna rimozione è l'algoritmo di machine learning.

[131] Regolamento (UE) 2022/2065 del Parlamento Europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE (regolamento sui servizi digitali), ex. art. 16

[132] Un esempio di alert proviene dai servizi forniti da Meta, come i social Instagram e Facebook che durante la pandemia di COVID-19 mostravano dei banner informativi sopra ai contenuti che utilizzavano termini legati al tema, con link ai siti dei ministeri della salute locali o dell'Organizzazione Mondiale della Sanità. Disponibile su: https://help.instagram.com/234606571236360

[133] Birritteri, 2023, p. 61

[134] La normativa prevede che le piattaforme di social media con oltre 2 milioni di utenti, se ricevono una segnalazione di un potenziale contenuto lesivo, devono valutare se il contenuto sia illegale per il Codice penale (*Strafgesetzbuch*), e in caso positivo entro 24 ore devono provvedere alla sua rimozione. Se questo non avviene le piattaforme possono incorrere in sanzioni fino a 5 milioni di euro. *Gesetz zur Verbesserung*

Legge (DDL) Gambaro, che prevede in particolare l'introduzione di una nuova fattispecie di reato per "chiunque pubblichi o diffonda notizie false, esagerate o tendenziose che riguardino dati o fatti manifestamente infondati o non veritieri, attraverso social media o altri siti che non siano espressione di giornalismo online". La proposta di legge prevede anche l'introduzione di nuovi reati per chiunque diffonda disinformazione o campagne d'odio online che possano provocare allarme pubblico, fuorviare l'opinione pubblica o incidere sui processi democratici. Inoltre il DDL introduce una responsabilità per i gestori delle piattaforme, che devono dunque monitorare ed eventualmente rimuovere tempestivamente i contenuti sopra descritti diffusi all'interno di queste. Il DDL Gambaro non è stato tuttavia ancora discusso[135].

Una riflessione che viene da porsi è se vi sia una reale efficacia di una regolamentazione in materia, essa potrebbe infatti rivelarsi controproducente. La rimozione da parte delle piattaforme di contenuti fuorvianti e tutte le relative conseguenze quali campagne d'odio, discriminazione, polarizzazione, in grado di danneggiare e influenzare la salute pubblica, la sicurezza pubblica, il dibattito civico, la partecipazione politica e l'uguaglianza[136] potrebbe finire per incrementare l'attenzione dell'opinione pubblica sul contenuto rimosso stesso, il fenomeno prende il nome di "*Streisand effect*"[137]. La censura potrebbe produrre l'effetto opposto rispetto a quello inteso di tutelare gli utenti, divenendo anzi più popolare.

Il punto di partenza rimane in ogni caso la capacità dell'utente di distinguere ciò che è vero, o reale nel caso dei deepfake, da ciò che è falso, disinformativo o generato dall'IA. Programmi di media literacy sul tema fake news dovrebbero essere introdotti all'interno delle scuole ma anche diffusi a quelle fasce di popolazione più deboli, potenziali vittime di disinformazione[138]. L'UE all'interno della sua Direttiva 1808/2018 nella versione in lingua italiana impiega l'espressione "alfabetizzazione mediatica",

---

der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG), Bundesamt für Justiz, 2017, §1, §3 (2)(2), §4 (2)

[135] Disegno Di Legge Gambaro. Disposizioni per prevenire la manipolazione dell'informazione online, garantire la trasparenza sul web e incentivare l'alfabetizzazione mediatica, Atto Senato n. 2688. XVII Legislatura, 28 febbraio 2017.

[136] Regolamento (UE) 2022/2065, Considerando n. 95

[137] La cantante e attrice Barbara Streisand nel 2003 avviò una campagna legale contro un fotografo per una foto pubblicata online che mostrava dall'alto la villa dell'attrice, sostenendo che si trattasse di una violazione della sua privacy. Al momento della denuncia la foto era stata visualizzata solo 6 volte, di queste 2 da parte degli avvocati. Circa un mese dopo, quando il caso ormai era divenuto noto al pubblico, la foto era ormai stata vista più di 400'000 volte e ripostata in maniera incontrollata in rete. Ancora oggi la foto è facilmente reperibile online. Britannica, *Streisand Effect*

[138] Vasu, 2018, p. 22 ss.

l'insieme cioè delle competenze e conoscenze necessarie per utilizzare efficacemente e in sicurezza i media. Tra queste competenze e conoscenze deve rientrare oltre che l' "apprendimento in materia di strumenti e tecnologie" anche "la capacità di riflessione critica necessaria per elaborare giudizi, analizzare realtà complesse e riconoscere la differenza tra opinioni e fatti"[139]. A questo fine l'UE richiede a piattaforme e fornitori di servizi di media di collaborare con gli Stati membri nella promozione dell'alfabetizzazione mediatica, con particolare attenzione nei confronti dei minori che potrebbero incorrere in contenuti lesivi[140], inoltre ogni tre anni gli Stati membri devono presentare alla Commissione una relazione per dimostrare i loro sviluppi sulla promozione dell'alfabetizzazione mediatica[141]. Un recente esempio di media literacy coinvolge la Commissione Europea in collaborazione con il Gruppo europeo dei regolatori per i servizi audiovisivi (ERGA) che ha avviato una campagna di sensibilizzazione per il contrasto alla disinformazione in occasione delle Elezioni Europee 2024[142]. Ma la media literacy non può essere solo uno strumento utilizzato per incaricare gli utenti ad auto responsabilizzarsi, devono comunque proseguire le strategie di regolamentazione sia da parte delle istituzioni che delle piattaforme[143]. Uno dei sistemi educativi di media literacy più fruttuosi proviene dalla Finlandia. La *National media education policy* del Ministero dell'educazione e cultura finlandese del 2019 prevede investimenti in diversi programmi volti a tutelare attraverso l'educazione tutti i gruppi sociali contro disinformazione, messaggi antidemocratici, discorsi d'odio, e abusi sessuali in rete, violazioni della privacy e della sicurezza dei dati[144]. I progetti di implementazione della policy riguardano ad esempio la produzione di materiali informativi circa i diritti e responsabilità degli utenti da mettere a disposizione delle scuole e in rete da parte dei vari ministeri e agenzie, ma anche progetti come il *Facts Against Hate*, volto a migliorare gli sforzi contro crimini d'odio e hate speech attraverso raccolte di dati, monitoraggi, report

[139] Direttiva (UE) 2018/1808 del Parlamento Europeo e del Consiglio del 14 novembre 2018 recante modifica della direttiva 2010/13/UE, relativa al coordinamento di determinate disposizioni legislative, regolamentari e amministrative degli Stati membri concernenti la fornitura di servizi di media audiovisivi (direttiva sui servizi di media audiovisivi), in considerazione dell'evoluzione delle realtà del mercato, Considerando 59
[140] *Ivi*, art. 28 ter, par. 3, lett. j)
[141] *Ivi*, art. 33 bis
[142] Campagna di comunicazione ERGA contro la disinformazione. Disponibile su: https://www.youtube.com/watch?v=I8QwjqOd8II
[143] Buckingham *et al.,* 2021, p. 22
[144] National Audiovisual Institute, 2019, p. 5

e cooperazione locale[145]. Inoltre sono stati introdotti all'interno delle scuole insegnamenti obbligatori di pensiero computazionale, programmazione, ICT e media literacy[146].

[145] Ministry of Justice Finland, 2019
[146] National Audiovisual Institute, 2019, p. 48 ss. Bocconi *et al*, 2018

3.2 Contrasto ai deepfake e Regolamento europeo sull'IA

Una volta esaminati i principali strumenti impiegati per il contrasto alla disinformazione online, è possibile entrare nel merito della lotta ai deepfake, che nel corso di questa trattazione abbiamo considerato come forma di disinformazione particolarmente preoccupante a livello sociale. Un primo approccio tecnico riguarda le piattaforme e i fornitori di servizi, i quali si stanno adoperando per progettare tecnologie in grado di identificare prodotti deepfake. La necessità di avere metodi di individuazione automatica di deepfake è la nuova sfida di oggi, non è infatti pensabile che migliaia di video ogni giorno caricati online possano essere analizzati uno per uno. A questo scopo Meta ha promosso nel 2019, in collaborazione con alcune Università e aziende tech, la *Deepfake Detection Challenge*, un'iniziativa aperta ad esperti di tutto il mondo per l'implementazione e il testing, attraverso un dataset di video deepfake creati appositamente per questo scopo, di nuovi modelli in grado di identificare deepfake e video manipolati[147]. A questa iniziativa si aggiunge il progetto *Coalition for Content Provenance and Authenticity* (C2PA) tra i cui membri figurano Adobe, Intel, BBC, Microsoft e Google. Si tratta di una iniziativa volta a garantire l'integrità e la provenienza delle informazioni digitali trasmesse in rete. La missione principale è lo sviluppo di standards tecnici globali per tracciare l'origine dei contenuti pubblicati[148]. Mentre questi approcci sono sicuramente fondamentali, non possono essere considerati le uniche risposte al problema. Uno studio del Dipartimento della difesa statunitense infatti sostiene che laddove un deepfake venga generato mediante reti generative avversarie, vi è una grande possibilità che nessun software di rilevamento sia in grado di riconoscerlo come tale, dunque provvedere a rimuoverlo prima della diffusione o etichettarlo come prodotto dell'IA sarebbe impossibile[149].

Al di là delle misure meramente tecniche, adottabili a posteriori per individuare contenuti deepfake in capo ai fornitori di servizi e piattaforme, l'intervento regolatorio resta uno strumento fondamentale per guidare le nuove tecnologie oltre che favorire una autodisciplina da parte delle piattaforme[150]. I primi ad intervenire con l'obiettivo di regolamentare il fenomeno furono gli Stati Uniti. Nel 2018 il Congresso presentò il

[147] Ferrer *et al*., 2020
[148] Si veda: https://c2pa.org/
[149] Mezzanotte, 2022, p. 60
[150] Liu *et al.*, 2023, p. 1312

Malicious Deep Fake Prohibition Act volto ad introdurre una nuova fattispecie di reato per coloro che creano o distribuiscono contenuti multimediali finti che appaiono realistici[151]. La proposta di legge non è mai stata tuttavia approvata. Restando negli USA, merita una citazione l'Assembly bill dello Stato della California, AB n. 730, in vigore fino al 1° gennaio 2023 che rendeva penalmente punibile l'atto di distribuzione di contenuti audio o video manipolati raffiguranti candidati politici con il fine di ingannare gli elettori durante i 60 giorni precedenti un elezione[152].

L'UE al contrario non hai mai elaborato una strategia per disciplinare il fenomeno attraverso norme specifiche, soprattutto per la difficoltà riscontrata nell'inquadrare il fenomeno sotto un profilo giuridico[153], questo almeno fino al 2024 con l'approvazione del primo regolamento sull'IA al mondo. Fino a quel momento si è sempre scelto di regolamentare macrofenomeni che includessero i deepfake. Tre sono i macro-temi considerati e che si relazionano strettamente al fenomeno deepfake: I) gestione della disinformazione, II) protezione dei dati personali e III) regolamentazione dell'IA. Sul primo punto si è ampiamente discusso, va ricordato in generale il Codice pratico rafforzato sulla disinformazione del 2022 volto a guidare le piattaforme verso una autodisciplina e, in aggiunta a questo, il Digital Service Act, al cui interno è stata introdotta una norma volta a istituire un meccanismo sanzionatorio per le piattaforme che non agiscono nell'arginare la diffusione di contenuti disinformativi all'interno delle stesse[154]. Sempre il DSA *ex* art. 35 in relazione all'attenuazione dei rischi richiede alle piattaforme online e motori di ricerca di utilizzare contrassegni visibili affinché " un elemento di un'informazione, sia esso un'immagine, un contenuto audio o video, generati o manipolati, che assomigli notevolmente a persone, oggetti, luoghi o altre entità o eventi esistenti e che a una persona appaia falsamente autentico o veritiero, sia distinguibile*".*

Il secondo punto coinvolge il GDPR, di cui si è accennato nel primo capitolo. Il trattamento dei dati personali per la produzione di deepfake può verificarsi sia nel momento dell'addestramento dell'algoritmo del sistema IA, sia per la creazione in sé del contenuto che potrebbe quindi impiegare volti o voci dei soggetti interessati dal trattamento. Per la base normativa del GDPR si rimanda al capitolo I.

---

[151] US Congress, S.3805 - 115th Congress (2017-2018): Malicious Deep Fake Prohibition Act of 2018
[152] State of California, AB-730 Elections: deceptive audio or visual media
[153] Mezzanotte, 2022, p. 55
[154] *Ivi*, 2022, p. 52

Infine sul terzo punto relativo alla regolamentazione dell'IA ci soffermiamo, di recente approvazione è infatti il nuovo regolamento sull'IA che definisce e per la prima volta cita il fenomeno deepfake. Il Regolamento sull'Intelligenza artificiale (meglio noto come AI Act) include una regolamentazione del fenomeno deepfake. I deepfake vengono definiti *ex* art. 3 (60) come contenuti immagine, audio o video generati dall'IA o manipolati che rappresentano persone esistenti, oggetti, luoghi, o altre entità o eventi e che potrebbero sembrare autentiche o veritiere ad una persona[155]. Il Regolamento adotta un approccio basato sul rischio per classificare i vari sistemi di IA. Con riferimento ai sistemi IA che generano deepfake non vi è un inquadramento esplicito come sistemi a rischio minimo, limitato, elevato o inaccettabile. In generale essi non rientrano nella lista di sistemi a rischio inaccettabile e dunque vietati, vanno in ogni caso rispettati determinati requisiti minimi e di trasparenza. Tuttavia quando si parla di sistemi di IA impiegati per influenzare elezioni, referendum o comportamenti degli elettori, essi dovrebbero essere trattati come sistemi ad alto rischio[156], in ogni caso *ex* art. 7 la Commissione si riserva la possibilità di ammendare l'annesso III, elenco dei tipi di sistemi ad alto rischio[157]. Ad ogni modo è considerato proibito l'impiego e l'introduzione nel mercato di sistemi di IA che attraverso tecniche subliminali, manipolative o ingannevoli, ostacolano l'assunzione di decisioni informate da parte degli individui, influenzando il comportamento degli stessi e provocandone un danno significativo[158]. Sotto il profilo della trasparenza, considerando la capacità dei sistemi di IA di generare contenuti difficili da distinguere come artificiali, e dunque incrementando il rischio di misinformazione, manipolazione e inganni, i fornitori di tali servizi dovrebbero introdurre tecniche adeguate a dimostrare l'origine del contenuto, come ad esempio watermarks, identificazione dei metadata e metodi crittografici[159], questo a tutela dei soggetti destinatari, affinché essi possano essere informati dell'artificialità del contenuto[160]. *Ex* art. 50 (4) si cita nuovamente il deepfake con la richiesta per gli sviluppatori di sistemi di IA di mostrare che il contenuto deepfake è stato generato o manipolato in conformità con i requisiti di trasparenza richiesti, anche

[155] Parlamento Europeo, RETTIFICA alla posizione del Parlamento europeo definita in prima lettura il 13 marzo 2024 in vista dell'adozione del regolamento (UE) 2024/… del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n.167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale), *ex*. art. 3 (60)
[156] *Ivi*, recital 62
[157] *Ivi*, art. 7
[158] *Ivi*, art. 5 (1)
[159] *Ivi*, recital 133
[160] *Ivi*, recital 132

quando il contenuto generato è di tipo testuale ed è stato diffuso per ragioni informative di pubblico interesse[161]. Questi obblighi non sono richiesti laddove siano state le Autorità ad aver generato tali contenuti per motivi di investigazione, prevenzione, indagine o processo. Mentre nel caso in cui i contenuti siano parte di lavori artistici, creativi, satirici, i requisiti di trasparenza possono limitarsi ad una dichiarazione dell'esistenza della generazione o manipolazione. Nel caso dei testi generati e pubblicati a scopo informativo, i requisiti suddetti non sono necessari se vi è stata revisione umana del contenuto, controllo editoriale e quando vi è la presenza di un soggetto che si assume la responsabilità editoriale della pubblicazione[162].

L'efficacia di tale Regolamento in relazione al fenomeno resta tuttavia dubbia, in primo luogo per la mancanza di una classificazione del tipo di rischio dei sistemi IA che li generano[163], in secondo luogo per l'ambiguità della deroga agli obblighi di etichettatura concessa ai contenuti creativi o artistici con il fine di garantire il diritto alla libertà di espressione e alla libertà delle arti e delle scienze. Vi è difatti la possibilità che contenuti dannosi possano essere ad esempio considerati satira e dunque deroga agli obblighi di trasparenza, impedendo di fatto una reale ed efficace lotta contro il fenomeno[164].

In conclusione, la soluzione regolamentativa non appare da sola sufficiente ad arginare gli effetti dei deepfake, vista l'ambiguità nel definire il fenomeno[165]. L'autoregolamentazione dei fornitori di sistemi IA, insieme a collaborazioni tra giornalisti, coalizioni tra aziende del settore, progetti di media literacy e definizione di standards tecnici per l'individuazione di deepfake restano strumenti di supporto ottimali per la lotta contro i deepfake. Permane perciò fondamentale un intervento collaborativo e coordinato da parte di tutti i soggetti ed enti coinvolti nel fenomeno deepfake.

[161] *Ivi*, recital 134
[162] *Ivi*, art. 50 (4)
[163] Moreno, 2024, p. 6
[164] *Ivi*, p. 5
[165] Mezzanotte, 2022, p. 64

CONCLUSIONE

Il presente elaborato ha cercato di delineare i differenti profili della disinformazione online, dalle sue fondamenta basate sulla profilazione degli individui, proseguendo con il principale luogo di propagazione, le bolle filtro, fino agli effetti drammatici che questi elementi possono comportare per la democrazia e sicurezza di un paese. Lo sviluppo esponenziale dell'Intelligenza artificiale ed in particolare della tecnologia deep learning ha permesso la nascita di un nuovo fenomeno disinformativo, il deepfake, più realistico e ingannevole rispetto alla tradizionale disinformazione online. Considerati i danni che questo tipo di contenuto è in grado di affliggere, risulta necessario un intervento coordinato e cooperativo da parte di giornalisti, aziende che si occupano di sistemi di IA e istituzioni. Un ruolo essenziale nella lotta alla disinformazione online è svolto dai cittadini stessi, destinatari dei contenuti. È necessario che i paesi ricorrano a programmi di media literacy seguendo l'esempio della Finlandia, che coinvolgano tutte le fasce più vulnerabili della popolazione, affinché i cittadini possano acquisire le competenze necessarie per riconoscere una disinformazione e fermarne la diffusione. I primi tentativi di costituire organi di fact-checking e di implementare sistemi di segnalazione e algoritmi in grado di individuare e rimuovere contenuti disinformativi all'interno delle piattaforme hanno mostrato i loro limiti, come anche le tecnologie progettate per riconoscere un contenuto deepfake e provvedere ad etichettarlo prima di una sua diffusione. La graduale evoluzione dell'IA non garantisce infatti che questi approcci siano efficaci, deepfake evoluti potrebbero essere impossibili da riconoscere. Il nuovo regolamento del Parlamento europeo sull'IA rappresenta il primo inquadramento giuridico del fenomeno deepfake anche se rimane ambigua la classificazione del tipo di rischio da esso rappresentato e dubbia l'effettività dei requisiti di trasparenza richiesti. Nel futuro prossimo, per contrastare gli effetti negativi sulla società e sulla democrazia, è perciò necessario investire risorse e tempo nello sviluppo di sistemi più evoluti di identificazione di contenuti generati dall'IA oltre che migliorare l'apparato normativo, sempre in un'ottica cooperativa e soprattutto lungimirante. La sfida è complessa ma fondamentale.

BIBLIOGRAFIA

Astone, A., *Capitalism of digital surveillance and digital disintermediation in the era of the pandemic*, in European Journal of Privacy Law & Technologies, 2, 2020. https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1215/546

Ayyub, R., *I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me*, in Huffington Post, 2018. Disponibile su: https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316 [Data di accesso: 30/03/2024]

Baptista, J.P.; Gradim, A.A., *Working Definition of Fake News*. Encyclopedia 2022, 2, 632–645. https://doi.org/10.3390/encyclopedia2010043

Bennett, W.L, Livingston, S. (editato da), *The disinformation age. Politics, Technology, and Disruptive Communication in the United States,* in SSRC Anxieties of Democracy, Cambridge University Press, 2021. doi.org/10.1017/9781108914628

Berkman Klein Center for Internet & Society at Harvard University, MIT Media Lab at the Massachusetts Institute of Technology, *Media and Information Quality,* Ethics and Governance of Artificial Intelligence Initiative, 2018. https://cyber.harvard.edu/sites/default/files/2018-07/2018-02-12_AIMediaIQ.pdf

Bianca, M., *La filter bubble e il problema dell'identità digitale*, in Media laws, 2, 2019. https://www.medialaws.eu/wp-content/uploads/2019/03/2-2019-Bianca.pdf

Birritteri, E*., Contrasto alla disinformazione, Digital Services Act e attività di private enforcement: fondamento, contenuti e limiti degli obblighi di compliance e dei poteri di autonormazione degli operatori,* in mediaLaws , 2/2023. https://www.medialaws.eu/wp-content/uploads/2023/10/2-23-Birritteri.pdf

Bocconi, S., Chioccariello, A. and Earp, J., *The Nordic approach to introducing Computational Thinking and programming in compulsory education*. Report prepared for the Nordic@BETT2018 Steering Group, 2018, https://www.itd.cnr.it/doc/CompuThinkNordic.pdf

Boldyreva, E., *Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process*, in 18th Professional Culture of the Specialist of the Future, dicembre 2018. https://www.researchgate.net/publication/330032180_Cambridge_Analytica_Ethics_And_Online_Manipulation_With_Decision-Making_Process

Borrello, P., *Alcune riflessioni preliminari (e provvisorie) sui rapporti tra i motori di ricerca ed il pluralismo informativo*, in Media laws, 1, 2017. https://www.medialaws.eu/wp-content/uploads/2019/05/7-.-Borrello.pdf

Boyd, D., *You Think You Want Media Literacy… Do You?*, Data & Society: Points, 2018, https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2

Britannica, *Destruction of the Maine*. Disponibile su: https://www.britannica.com/event/destruction-of-the-Maine [Data di accesso: 29/03/2024]

Britannica, *January 6 U.S. Capitol attack*. Disponibile su: https://www.britannica.com/event/January-6-U-S-Capitol-attack [Data di accesso: 01/03/2024]

Britannica, *Streisand Effect*. Disponibile su: https://www.britannica.com/topic/Streisand-effect [Data di accesso: 13/04/2024]

Buckingham, D., Farinacci, E., Manzoli, G., *Media Education in the Digital Age: An Interview with David Buckingham*, Media and/or literacy? Didattica dei media, didattica coi media, Sociologia della comunicazione 62(XXXII), Franco Angeli, Milano, 2021.

Bundesamt für Justiz *, Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken* (Netzwerkdurchsetzungsgesetz - NetzDG), 2017, https://www.gesetze-im-internet.de/netzdg/NetzDG.pdf

Censis, 18° *Rapporto Censis sulla Comunicazione,* dicembre 2022. https://group.intesasanpaolo.com/content/dam/portalgroup/repository-documenti/ricerche-comportamentali/Sintesi.pdf

Commissariato di Polizia di Stato, *Come riconoscere un deepfake,* 2024. Disponibile su: https://www.commissariatodips.it/notizie/articolo/come-riconoscere-un-deepfake/index.html [Data di accesso: 02/04/2024]

Commissione europea*, Codice di Condotta rafforzato sulla disinformazione*, 2022. Disponibile su: https://digital-strategy.ec.europa.eu/it/policies/code-practice-disinformation [Data di accesso: 29/03/2024]

Commissione Europea, *Un codice di condotta dell'UE più rigoroso sulla disinformazione*. Disponibile su: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/protecting-democracy/strengthened-eu-code-practice-disinformation_it [Data di accesso: 29/03/2024]

Commissione Europea. *Signatories of the 2022 Strengthened Code of Practice on Disinformation.* Disponibile su: https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation [Data di accesso: 29/03/2024]

Commissione Europea, *Comunicazione della Commissione al Parlamento europeo, al Consiglio, al Comitato Economico e Sociale europeo e al Comitato delle regioni. Contrastare la disinformazione online: un approccio europeo*, 2018. https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52018DC0236&from=it

Commissione Europea, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions. Tackling Illegal Content Online*. *Towards an enhanced responsibility of online platform*, Bruxelles, 2017, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017DC05555

Costa, P., *Motori di ricerca e social media: i nuovi filtri nell'ecosistema dell'informazione online e il potere occulto degli algoritmi*, in Avanzini, G., Matucci, G., (a cura di), *L'informazione e le sue regole. Libertà, pluralismo e trasparenza*, Napoli, 2016.

Costanzo, P., *La democrazia digitale (precauzioni per l'uso)*, in Diritto pubblico, Rivista fondata da Andrea Orsi Battaglini, 1, 2019, pp. 71-88. https://www.rivisteweb.it/doi/10.1438/93720

Cover, R., *Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation*, Continuum, 36:4, 609-621, 2022, DOI: 10.1080/10304312.2022.2084039

Crepaldi, M., (a cura di), *Quanto è affidabile Google Traduttore?*, ottobre 2023. Disponibile su: https://preply.com/it/blog/google-traduttore-quanto-e-affidabile/#:~:text=Google%20Traduttore%20traduce%20meglio%20verso,59%20errori%20su%2010%20parole) [Data di accesso: 08/03/2024]

Cuono, M., *Gli americani credono ai loro miti?*, in Teoria politica, 11, 2021, p. 123-131. Disponibile su: https://journals.openedition.org/tp/1808 [Data di accesso: 01/03/2024]

Deepmind, *Alphago*. Disponibile su: https://deepmind.google/technologies/alphago/ [Data di accesso: 02/04/2024]

Derico, B., Clayton, J., *Bruce Willis denies selling rights to his face*, in BBC News, 2022. Disponibile su: https://www.bbc.com/news/technology-63106024 [Data di accesso: 03/04/2024]

Fasan, M., *Intelligenza artificiale e pluralismo: uso delle tecniche di profilazione nello spazio pubblico democratico*, in Rivista di BioDiritto, 1, 2019. https://doi.org/10.15168/2284-4503-354

Federal Trade Commission, *FTC Sues Cambridge Analytica, Settles with Former CEO and App Developer,* luglio 2019. Disponibile su: https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer [Data di accesso: 02/03/2024]

Ferrer, C.C., Dolhansky, B., Pflaum, B., Bitton, J., Pan, J., Lu, J., *Deepfake Detection Challenge Results: An open initiative to advance AI*, in Meta blog, 2020. Disponibile su: https://ai.meta.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ [Data di accesso: 19/04/2024]

Franklin, S., *History, motivations and core themes of AI*, University of Memphis, 2014. https://digitalcommons.memphis.edu/cgi/viewcontent.cgi?article=1029&context=ccrg_papers

Gemini, *Gemini*. Disponibile su: https://gemini.google.com/app/4b3f868f369a4290 [Data di accesso: 02/04/2024]

GPDP, *Deepfake Vademecum: Deepfake Il falso che ti «ruba» la faccia (e la privacy)*, 2020. https://www.garanteprivacy.it/documents/10160/0/Deepfake+-+Vademecum.pdf/478612c7-475b-2719-417f-869e5e66604e?version=2.0

Hayes, B., *Computing Science: Delving into Deep Learning*, American Scientist, May–June 2014, Vol. 102, No. 3, pp. 186- 189, Sigma Xi, The Scientific Research Honor Society. https://www.jstor.org/stable/43707183

Hooke Pearson, G. D., Knobloch-Westerwick, S., Is *the Confirmation Bias Bubble Larger Online? Pre-Election Confirmation Bias in Selective Exposure to Online versus Print Political Information*, in Mass Communication and Society, 4, 2019, 467.

IFCN, *The commitments of the code of principles*, Disponibile su: https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles [Data di accesso: 09/04/2024]

Italiano, G. F., *Intelligenza artificiale: passato, presente, futuro*, in Pizzetti, F., (a cura di), *Intelligenza artificiale, protezione dei dati personali e regolazione*, 2018, Giappichelli editore.

Khan, A., Brohman, K., Addas, S., *The anatomy of "fake news": Studying false messages as digital objects.* In Journal of Information Technology 2022, Vol. 37(2) 122–143. https://journals.sagepub.com/doi/10.1177/02683962211037693

Lagioia, F., Sartor, G., *Profilazione e decisione algoritmica: dal mercato alla sfera pubblica*, in Federalismi.it, 11, 2020. https://www.federalismi.it/ApplOpenFilePDF.cfm?artid=42114&dpath=document&dfile=23042020224508.pdf&content=Profilazione%2Be%2Bdecisione%2Balgoritmica%3A%2Bdal%2Bmercato%2Balla%2Bsfera%2Bpubblica%2B%2D%2Bstato%2B%2D%2Bdottrina%2B%2D%2B

Liu, M., Zhang, X., *Deepfake technology and current legal status of it*, in B. Fox et al. (Eds.): ICICAIE 2022, AHCS 9, pp. 1308-1314, 2023, https://doi.org/10.2991/978-94-6463-040-4_194

Longo, E., *Dai big data alle "bolle filtro": nuovi rischi per i sistemi democratici,* in Percorsi costituzionali, 1, 2019, Jovene editore. https://www.academia.edu/44266744/Dai_big_data_alle_bolle_filtro_nuovi_rischi_per_i_sistemi_democratici

Lusa Agência de Notícias de Portugal, *PONTOS ESSENCIAIS Eleições: Desporto ganha no Facebook, política domina Twitter/X – MediaLab,* 2024. Disponibile su: https://www.lusa.pt/article/42356358/pontos-essenciais-elei%C3%A7%C3%B5es-desporto-ganha-no-facebook-pol%C3%ADtica-domina-twitter-x-medialab [Data di accesso: 11/04/24]

Macagnone, F, *Ragazza musulmana "indifferente" sul Westminster Bridge dopo l'attentato: la fake news opera di un troll russo,* in Il messaggero, 2017, disponibile su: https://www.ilmessaggero.it/primopiano/esteri/ragazza_musulmana_indifferente_westminster_bridge_dopo_attentato_fake_news_opera_troll_russo-3366447.html?refresh_ce [Data di accesso: 29/03/2024]

Matucci, G., *Informazione online e dovere di solidarietà. Le fake news fra educazione e responsabilità,* in Rivista AIC, 1/2018. https://www.rivistaaic.it/images/rivista/pdf/1_2018_Matucci.pdf

Meta, *How Meta's third-party fact-checking program works*, 2021. Disponibile su: https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works [Data di accesso: 14/04/2024]

Metropolitan Police Department, *Arrest Made in an Assault with a Dangerous Weapon (Gun): 5000 Block of Connecticut Avenue, Northwest,* dicembre 2016. Disponibile su: https://web.archive.org/web/20170222203449/http://mpdc.dc.gov/release/arrest-made-assault-dangerous-weapon-gun-5000-block-connecticut-avenue-northwest [Data accesso: 01/03/2024]

Mezzanotte, M., *Fake news, deepfake e sovranità digitale nei periodi bellici*, in Federalismi.it, 2022, https://www.federalismi.it/ApplOpenFilePDF.cfm?artid=48112&dpath=document&dfile=12122022191452.pdf&content=Fake%2Bnews%2C%2Bdeepfake%2Be%2Bsovranit%C3%A0%2Bdigitale%2Bnei%2Bperiodi%2Bbellici%2B%2D%2Bstato%2B%2D%2Bdottrina%2B%2D%2B

Ministry of Justice Finland, *Facts Against Hate*, 2019. Disponibile su: https://oikeusministerio.fi/en/project?tunnus=OM043:00/2019 [Data di accesso: 14/04/2024]

Montaldo, R., *La tutela del pluralismo informativo nelle piattaforme online*, in Media Laws, 1, 2020. https://www.medialaws.eu/wp-content/uploads/2020/03/1-2020-Montaldo.pdf

Monti, M., *Fake news e social network: la verità ai tempi di Facebook*. In Saggi- Sezione Monografica "fake news, pluralismo informativo e responsabilità in rete". MediaLaws, 1/2017. https://www.medialaws.eu/wp-content/uploads/2019/05/8.-Monti.pdf

Moreno, F.R, *Generative AI and deepfakes: a human rights approach to tackling harmful content*, International Review of Law, Computers & Technology, 2024, https://www.tandfonline.com/doi/full/10.1080/13600869.2024.2324540

National Audiovisual Institute*, National media education policy*, Ministry of Education and Culture, 2019, https://medialukutaitosuomessa.fi/mediaeducationpolicy.pdf

O'Neil, C., *Weapons of math destruction: how big data increases inequality and threatens democracy*, prima edizione, New York: Crown Publishers, 2016, p. 63-74.

Openai. *Homepage*. Disponibile su: https://openai.com/ [Data di accesso: 02/04/2024]

Palano, D., *La democrazia alla fine del "pubblico". Sfiducia, frammentazione, polarizzazione: verso una "bubble democracy"?*, in Governare la paura, aprile 2019, p. 35-92.

Paris, B., Donovan, J., *Deepfakes and cheap fakes. The manipulation of audio and visual evidence*, in Data & Society's Media Manipulation, 2019, p.5-18.

Pariser, E., *The filter bubble: what the Internet is hiding from you*, Penguin Books, 2011.

Parlamento Europeo, *RETTIFICA alla posizione del Parlamento europeo definita in prima lettura il 13 marzo 2024 in vista dell'adozione del regolamento (UE) 2024/... del Parlamento europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n.167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828* (regolamento sull'intelligenza artificiale), 2024, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_IT.pdf

Parlamento europeo, *Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence* (Artificial Intelligence Act) and amending certain Union Legislative Acts, 2024. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

Parlamento europeo, *Fermare il greenwashing: come l'UE regola le asserzioni ambientali*. Disponibile su: https://www.europarl.europa.eu/topics/it/article/20240111STO16722/fermare-il-greenwashing-come-l-ue-regola-le-asserzioni-ambientali [Data di accesso: 29/03/2024]

Parlamento Europeo*, Regolamento (UE) 2022/2065 del Parlamento Europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE* (regolamento sui servizi digitali). https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32022R2065

Parlamento europeo, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* STUDY Panel for the Future of Science and Technology. European Parliamentary Research Service. Scientific Foresight Unit (STOA) PE 641.530 June 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

Parlamento Europeo, *Direttiva (UE) 2018/1808 del Parlamento Europeo e del Consiglio del 14 novembre 2018 recante modifica della direttiva 2010/13/UE, relativa al coordinamento di determinate disposizioni legislative, regolamentari e amministrative degli Stati membri concernenti la fornitura di servizi di media audiovisivi* (direttiva sui servizi di media audiovisivi)*, in considerazione dell'evoluzione delle realtà del mercato.* https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32018L1808&from=pl

Parlamento europeo, *Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE* (regolamento generale sulla protezione dei dati). https://www.garanteprivacy.it/documents/10160/0/Regolamento+UE+2016+679.+Arricchito+con+riferimenti+ai+Considerando+Aggiornato+alle+rettifiche+pubblicate+sulla+Gazzetta+Ufficiale++dell%27Unione+europea+127+del+23+maggio+2018

Pitruzzella, G., *La libertà di informazione nell'era di Internet*, in Media laws, 1, 2018. https://www.medialaws.eu/wp-content/uploads/2019/05/1.-Pitruzzella.pdf

Regio decreto 19 ottobre 1930, n. 1398, art. 612 ter.

Risso, L., *Harvesting Your Soul? Cambridge Analytica and Brexit*, in Akademie der Wissenschaften und der Literatur, Jansohn, C., (a cura di), *Brexit Means Brexit?,* Mainz, dicembre 2017, p.75- 87. https://www.adwmainz.de/fileadmin/user_upload/Brexit-Symposium_Online-Version.pdf#page=75

Royal Society, *Machine learning: the power and promise of computers that learn by example*, 2017. https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Sadiq, M., *Real v fake: debunking the 'drunk' Nancy Pelosi footage*, in The Guardian, 2019. Disponibile su: https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video [Data di accesso: 30/03/2024]

Sejnowski, T. J., *The deep learning revolution*, Cambridge, MA: The MIT Press, 2018. https://lccn.loc.gov/2017044863

Senato della Repubblica, *DDL Gambaro, Disposizioni per prevenire la manipolazione dell'informazione online, garantire la trasparenza sul web e incentivare l'alfabetizzazione mediatica*, Atto Senato n. 2688, XVII Legislatura, 2017, https://www.senato.it/service/PDF/PDFServer/BGT/01006504.pdf

Shu, K., Bhattacharjee, A., Alatawi, F., et al. *Combating disinformation in a social media age*. 2020, WIREs Data Mining Knowl Discov. 2020;10:e1385. https://doi.org/10.1002/widm.1385

Sicurezza Nazionale, *Spoofing*, Disponibile in: https://www.sicurezzanazionale.gov.it/comunicazione/glossario/372 [Data di accesso: 05/04/2024]

State of California, *AB-730 Elections: deceptive audio or visual media*, Legislative Counsel Bureau, 2019, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

Sullivan, D., *Google uses RankBrain for every search, im-pacts rankings of "lots" of them*, 2016. https://searchengineland.com/google-loves-rankbrain-uses-for-every-search-252526 [Data di accesso: 03/04/2024]

Treccani, *Donazione di Costantino*. Disponibile su: https://www.treccani.it/enciclopedia/donazione-di-costantino_%28Enciclopedia-Dantesca%29/ [Data di accesso: 29/03/2024]

Treccani, *Filter bubble.* Disponibile su: Treccani.it/vocabolario/filter-bubble_res-b92bdbdc-89c2-11e8-a7cb-00271042e8d9_%28Neologismi%29/ [Data di accesso: 08/03/2024]

Treccani, *Legge di Moore*. Disponibile su: https://www.treccani.it/enciclopedia/legge-di-moore_(Enciclopedia-della-Scienza-e-della-Tecnica)/ [Data di accesso: 08/03/2024]

Treccani, *QAnon*. Disponibile su: https://www.treccani.it/enciclopedia/qanon/ [Data di accesso: 02/03/2024]

Turrini, D., *Quello non sono io, ma un sosia creato dall'Intelligenza Artificiale": la denuncia di Tom Hanks contro lo spot pubblicitario con la sua faccia.* In Il Fatto Quotidiano,. Disponibile su: https://www.ilfattoquotidiano.it/2023/10/03/quello-non-sono-io-ma-un-sosia-creato-dallintelligenza artificiale-la-denuncia-di-tom-hanks-contro-lo-spot-pubblicitario-con-la-sua-faccia/7311791/ [Data di accesso: 30/03/2024]

US Congress, S.3805 - 115th Congress (2017-2018): *Malicious Deep Fake Prohibition Act of 2018*, 2018, https://www.congress.gov/bill/115th-congress/senate-bill/3805

Vasu, N., Ang, B., Teo, T., et al., *International Responses to Fake News*, Fake news: National security in the post-truth era, S. Rajaratnam School of International Studies, 2018, p. 18-25, http://www.jstor.com/stable/resrep17648.8

Vernice, A., *Il letto di Procuste dei sistemi informativi via Web: un pluralismo falsato?*, in Rivista italiana di informatica e diritto, 3, 1 (giu. 2021), 89-101. https://www.rivistaitalianadiinformaticaediritto.it/index.php/RIID/article/view/64/47

Waldrop, M.M., *What are the limits of deep learning?*, Proceedings of the National Academy of Sciences of the United States of America, January 22, 2019, Vol. 116, No. 4, pp. 1074-1077, National Academy of Sciences. https://www.jstor.org/stable/10.2307/26580207

Wang, Z., *Improved Link-Based Algorithms for Ranking Web Pages*, 2003, NYU. https://cs.nyu.edu/media/publications/TR2003-846.pdf

World economic Forum, *The Global Risks Report 2024, 19th Edition. Insight report*. https://www.weforum.org/publications/global-risks-report-2024/.