

UNIVERSITÀ DEGLI STUDI DI NAPOLI “FEDERICO II”

DIPARTIMENTO DI SCIENZE SOCIALI



DOTTORATO DI RICERCA

IN

SCIENZE SOCIALI E STATISTICHE

XXXVII CICLO

La regolamentazione pubblica delle piattaforme digitali.
Disegnando il codice dell'identità in Rete

*Public regulation of digital platforms.
Designing the code of networked identity*

Tutor
Ch.mo Prof.
Fortunato Musella

Candidata
Dott.ssa
Sara Concetta Santoriello

Co-Tutor
Ch.mo Prof.
Emiliano Grimaldi

Matricola
DR996210

Coordinatrice
Ch.ma Prof.ssa
Enrica Amaturro

2024/2025

Nelle radici, la forza

Indice

INTRODUZIONE	4
PER UN'ETICA DELL'INTELLIGENZA ARTIFICIALE	7
1.1 DAL CONTROLLO DEL <i>PROMPT</i> AL RIALLINEAMENTO DEI VALORI UMANI	7
1.2 LA NORMALIZZAZIONE DELLA RAZIONALITÀ TECNOLOGICA NELLA SFERA PUBBLICA	14
1.3 LA MATRICE DI CORPORATION DIGITALE	20
1.3.1 <i>Stato e mercato nell'ecosistema digitale</i>	23
1.3.2 <i>L'altalena dei diritti tra Giustizia e Libertà</i>	25
1.3.3 <i>Quattro fenomeni della digitalizzazione</i>	27
1.3.4 <i>Ulteriori correlazioni nei concetti periferici</i>	32
1.4 BIOPOLITICHE DEL DIGITALE NEL <i>GARBAGE IN, GARBAGE OUT (GIGO)</i>	34
I DIRITTI FONDAMENTALI ALLA PROVA DELLA MODERAZIONE AUTOMATIZZATA	38
2.1 LE FONTI DEL DIRITTO E IL CONCETTO DI SOVRANITÀ DIGITALE	38
2.2 IL PROCESSO DI <i>PRIVATE POLICY-MAKING</i> : GOVERNARE CON L'IA	46
2.3 IL MARGINE CONTESO. ANALISI COMPARATA SULLA <i>AUTOMATED CONTENT MODERATION</i>	54
2.3.1 <i>La libertà di espressione come principio regolatore</i>	57
2.3.2 <i>I "contenuti illegali"</i>	60
2.3.3 <i>Misure di restrizione e penalizzazione</i>	62
2.3.4 <i>Sanzioni da parte dell'AGCOM</i>	64
IL GOVERNO DEI CORPI INTERMEDI	66
3.1 IL DIRITTO D'APPELLO ONLINE	66
3.2 L' <i>OVERSIGHT BOARD</i> DI META (MOB)	70
3.3 PRESENTARE O RICEVERE UNA SEGNALAZIONE SU X	75
3.4 FARE APPELLO SU TIKTOK	77
3.5 LA DURA LEGGE DELL'INTEROPERABILITÀ	80
ALIENAZIONE E IDENTITÀ NELLO SPETTRO ALGORITMICO	82
4.1 <i>CODE AND CONTROL</i> . L'IPOTESI DI UNA STANDARDIZZAZIONE DELLA CITTADINANZA	82
4.2 IL GENERE TRA CORPI, VIOLENZA E IDENTITÀ	86
4.3 LA RAZZA TRA ETNIA, CULTO E COLORE DELLA PELLE	94
4.4 QUESTIONI DI CLASSE: MANSIONI, SALARIO E COSCIENZA	97
4.5 COS'È POLITICO? TRA ETICA E ORDINE SOCIALE	104
CORPI, AGENCY E GIUSTIZIA NELL'<i>ETHICALLY ALIGNED DESIGN</i>	111
5.1 LE TEORIE DELLA GIUSTIZIA OLTRE LA RAGIONE ARTIFICIALE	111
5.2 QUALE APPRENDIMENTO?	115
5.3 LA DISOBBEDIENZA DEL <i>GLITCH</i>	117
CONCLUSIONI	122
APPENDICE	134
A. MODERAZIONE DEI CONTENUTI ALLA LUCE DEL <i>DIGITAL SERVICES ACT</i>	134
B. FONTI	137
BIBLIOGRAFIA	144
FILMOGRAFIA	152

Introduzione

Anomalo è quel comportamento che devia da norme, aspettative e regole condivise in un contesto culturale e sociale; in altre parole, si discosta significativamente da ciò che è considerato “normale” o appropriato. Proprio come una narrazione, la valutazione è determinata asimmetricamente dallo sguardo di chi presenta i fatti, o i dati.

Nel racconto omerico, per esempio, a Ulisse spetta il nuovo, l’inaspettato, l’azione dell’eroe destinato a salpare per nuovi mondi, mentre a Penelope viene assegnata l’attesa, la conservazione, il confinamento in un luogo. Quel che in realtà si manifesta è la sottrazione di Penelope al ruolo, un atteggiamento che viene ritenuto anomalo dalle figure che la circondano, emissari di un intangibile ordine simbolico. Il lavoro sull’incertezza è per il suo personaggio una dimostrazione di consapevolezza¹: si tiene al presente (dubita), prolunga il tempo dell’identica ripetizione (sospetta), trova il modo di inserirsi negli eventi (chiede prove).

Come Penelope, questo lavoro resta con il problema, in controtendenza con i miti dell’innovazione, della velocità e dell’accumulazione; fa e disfa la tela delle relazioni di potere strutturalmente ramificate nell’industria tecnologica; pone enfasi sul politicamente non-detto delle sue contraddizioni.

L’implementazione di sistemi di intelligenza artificiale (IA) apre interrogativi non solo sull’efficacia e sull’efficienza delle macchine, ma anche su come si stanno trasformando valori fondamentali come la giustizia, l’autonomia e la protezione dei Diritti Umani. La discussione etica è interessata a garantire che il progresso tecnologico non si accompagni a rischi inaccettabili per la società. Nelle prossime pagine, l’attenzione sarà rivolta ai tre principali attori del processo di digitalizzazione: le Corporation digitali, lo Stato e la Comunità.

Innanzitutto, la cooperazione internazionale ha messo in moto l’affermazione di principi etici, rendendo imprescindibili gli sforzi a ogni latitudine. Nel Capitolo 1 si ripercorrono le tappe essenziali, i ragionamenti e le elaborazioni teoriche che hanno guidato la normalizzazione della razionalità tecnologica. Le Corporation digitali sono il fulcro della matrice concettuale elaborata nei primi mesi di questo percorso e ispirata alla più analogica versione contenuta in *Hyperpolitics*. Il paradigma della frontiera dell’innovazione si scontra con le evidenze biopolitiche di inclusione/esclusione, il fatto e finito regno del passato che si impone sul presente, dove logiche patriarcali e discriminatorie vengono rese alla stregua della tecnica.

¹ Un intero capitolo di contronarrazione è dedicato a Penelope in Cavarero, A. (1990). *Nonostante Platone. Figure femminili nella filosofia antica*. Castelvechi, Lit Edizioni, Roma.

Le piattaforme della conoscenza amministrate da Meta (Facebook, Instagram), X (già Twitter) e ByteDance (TikTok) hanno dimostrato di poter contare su un incondizionato potere definitorio delle condotte mediante modifiche unilaterali degli accordi, non senza ripercussioni per l'esercizio della libertà di espressione. Questo tipo di contrattazione manca di una reale negoziazione e gli utenti possono semplicemente accettare o rifiutare le condizioni predefinite ("take it or leave it"). Per rendere operativo il procedimento, sono stati via via introdotti sistemi di IA nella moderazione dei contenuti che, affiancati da operatori umani, consentono di gestire enormi quantità di dati di interazione uno-a-uno e uno-a-molti.

Ciò non avviene in un deserto normativo, allorché l'Unione europea ha reagito con iniziative che mirano soprattutto a proteggere la privacy, promuovere la trasparenza e limitare le pratiche abusive o sleali. La stagione della regolamentare europea, già avviata con l'entrata in vigore del GDPR nel 2018, ha visto l'apprestarsi sulla scena del *Digital Services Act* (DSA), del *Digital Markets Act* (DMA) e dell'*Artificial Intelligence Act* (AIA).

L'approccio della ricerca qualitativa è sembrato più idoneo a indagare in profondità l'argomento, dalla costruzione della base empirica al modo di organizzare i dati, per far luce sul fenomeno a partire dalla proprietà degli oggetti della ricerca. L'ipotesi proposta è che le Corporation digitali mettano in pratica un processo di standardizzazione dell'utenza, e dunque della cittadinanza, per adattarla alle proprie regole, che sono il risultato del sostrato politico e culturale da cui provengono. Indagandone le modalità, la ricerca è stata suddivisa in tre fasi: esplorativa, analitico-comparativa e applicativa.

La fase esplorativa ha riguardato lo studio della dottrina politica riguardante la sovranità, per una migliore definizione degli spazi liminari. Questo passaggio ha fornito gli strumenti per una migliore definizione del campo di indagine, chiarendo gli ambiti di disciplina, addestramento e applicazione delle decisioni automatizzate sulle piattaforme di social media, lo storico delle decisioni e le modalità di risoluzione delle controversie nell'ambito pubblicistico e privatistico. La fase analitico-comparativa consta di due livelli a partire dalla scomposizione in elementi più semplici, per calcolare le ricorrenze e classificarne i contenuti in maniera sistematica, con l'applicazione dei medesimi criteri a ciascuna unità.

I documenti di moderazione dei contenuti e le norme che disciplinano l'attività delle piattaforme sono di tipo pragmatico, ossia sono creati per regolare azioni o per trasmettere informazioni operative che orientano il comportamento delle persone e delle istituzioni. Sono progettati per essere chiari, diretti e vincolanti, in modo da evitare ambiguità interpretative, e classificabili come istituzionali perché prodotti e diffusi da enti od organizzazioni di cui si riconosce l'autorità. Leggi, sentenze e documenti amministrativi hanno valore normativo: non

sono solo descrittivi, ma anche prescrittivi. Hanno, quindi, l'obiettivo di stabilire cosa deve essere fatto e in che modo, rappresentando principi e direttive da seguire con l'intento di essere oggettivi e impersonali, in modo che il messaggio sia il più possibile neutro e universale.

Anche se il campo è circoscritto in maniera referenziale, sconfinando nella responsabilità sociale d'impresa, permette di comprendere come viene distribuito il potere tra Dipartimenti, deleghe e disposizioni formali e quali valori sono sostenuti, con il fine ultimo di regolare comportamenti e relazioni.

Per quanto concerne il campionamento di primo livello, le unità comunicative da esaminare sono state selezionate in base alla fonte e al contenuto di pertinenza nell'ambito della moderazione dei contenuti: le sorgenti sono state estratte dai portali delle piattaforme di social media (es. *Transparency Center*) e dalle risorse pubbliche messe a disposizione online (es. EurLex), come si vedrà nel Capitolo 2. L'Appendice A sulla Moderazione dei contenuti alla luce del DSA offre uno schema di sintesi delle strategie e delle tecniche impiegate da Meta, X e ByteDance. I regolamenti UE sono stati considerati come variabile indipendente mentre la regolamentazione interna come variabile dipendente (Appendice B), a cui sono stati associati attributi sulla base dei settori di applicazione, cioè il prodotto (*design*), il suo utilizzo (contenuto) e il processo di realizzazione (lavoro). Il Capitolo 3 prosegue l'analisi comparativa dei processi di moderazione, soffermandosi su alcuni aspetti quali la risoluzione extragiudiziale delle controversie, il diritto di appello e la gestione dei contenuti "illeciti".

Il campionamento di secondo livello ha, invece, riguardato il rapporto tra la piattaforma e l'utenza nella definizione del confine di ammissibilità dei contenuti: i documenti - le *policy* su termini di servizio, privacy, *cookies* e ogni ulteriore chiarimento interpretativo sulla permanenza o di pertinenza in materia di pubblicazione e diffusione dei contenuti sulle piattaforme in analisi - sono stati importati, esplorati e codificati. Questa fase della ricerca qualitativa si è avvalsa del contributo del *software* di analisi NVivo. Il Capitolo 4 contiene i risultati derivanti dai "nodi", suddivisi per i macrotemi "Genere", "Razza" e "Classe", che corrispondono alle tre principali forme di discriminazione algoritmica al centro della presente indagine. L'analisi confluisce in una più ampia osservazione, contenuta nel Capitolo 5, degli impatti sociali della predizione algoritmica in contesti cruciali quali la giustizia e l'educazione, nonché dei fenomeni emergenti di tutela delle lavoratrici e dei lavoratori dello spettacolo dal pericolo di riproduzione non consensuale degli attributi personali e della produzione intellettuale. Sebbene i processi di automatizzazione vengano descritti come irreversibili, sono crescenti i casi di disobbedienza che problematizzano, ribaltando o eludono, le norme implicite e la soggettivazione che deriva dall'esposizione a sistemi di IA.

Capitolo 1

Per un'etica dell'intelligenza artificiale

«Io - la mia porzione di cecità.
Io - la mia porzione di luce»
Carla Lonzi

1.1 Dal controllo del *prompt* al riallineamento dei valori umani

L'introduzione di macchine e innovazioni tecnologiche è stato un fattore di trasformazione per le comunità agricolo-artigianali che hanno ceduto il passo al sistema industriale. Coefficienti quali la minore quantità di energia necessaria a produrre informazione, l'aumento della potenza di calcolo e l'estensione della capacità di archiviazione/programmazione di dati e sistemi operativi hanno reso possibile il trasferimento e la fruizione della conoscenza su vasta scala. Anche il mondo analogico contemporaneo si reinventa, sperimentando una declinazione delle sue attività attraverso il digitale.

Nella “*Platform society*” (Van Dijck *et al.*, 2018) le persone beneficiano del cambio di paradigma perché in grado di “fare da sé” attraverso dispositivi elettronici, applicazioni e programmi disintermediati da istituzioni e organizzazioni tradizionali. La piattaforma è una base *hardware* o *software* su cui sono sviluppati e/o eseguiti programmi e applicazioni. In Rete, questa infrastruttura consente lo scambio di informazioni, servizi o prodotti mediante una interfaccia che facilita l'azione dell'utente. La proprietà ne orienta il funzionamento, che si caratterizza per la presenza di algoritmi programmati per raccogliere, elaborare e valorizzare i dati all'occorrenza codificati (indirizzi IP, geolocalizzazione, interessi, gusti), regolati mediante termini di servizio (*user agreements*), con il fine di filtrare i contenuti e collegare gli utenti a servizi e pubblicità (*ivi*, p. 10). Il modello di *business* è veicolato dall'attenzione e monetizza per mezzo di denaro, visualizzazioni o recensioni, a partire da funzioni “offerte” gratuitamente. Nel contesto internazionale si osservano mutamenti che riguardano gli stili di vita e di consumo di miliardi di persone (Musella, 2022).

L'algoritmo su cui poggia la piattaforma è una sequenza di istruzioni e procedure che possono risolvere o prevedere problemi collettivi sulla base di una logica condizionale “se...allora...” determinata da attori umani (Bucher, 2018). Essi sono portatori dei valori e della visione di chi li possiede e alla base dei sistemi di intelligenza artificiale (IA), integrati e utilizzati per categorizzare i comportamenti, personalizzare l'esperienza, individuare *pattern* ed elaborare

sintesi automatizzate sui dati raccolti. Le IA riproducono il processo di ragionamento (percezione e apprendimento) o di comportamento (azione e comunicazione) la cui misura di efficacia è l'essere umano o un agente ideale. L'esecuzione di un compito – tra cui si annoverano l'apprendimento automatico (*machine learning*), l'elaborazione del linguaggio naturale (*Natural Language Processing*, NLP e *Large Language Model*, LLM), la rappresentazione della conoscenza (*Generative AI*, GenAI) e la robotica - si osserva in termini di risultato, affinché esso sia buono o migliore se confrontato con il medesimo risultato ottenuto dall'intelligenza umana. Resta detto, quandanche non abbastanza chiarito, che il cervello e il computer non sono assimilabili e non funzionano allo stesso modo².

Le interconnessioni tra piattaforme, algoritmi e sistemi intelligenti, per lungo tempo di pertinenza esclusiva del mondo informatico, godono di un rinnovato interesse di ricerca nelle scienze umanistiche e sociali in virtù di una accresciuta fruizione a livello cittadino che deve far fronte a potenzialità e limitazioni tecniche, ponendosi l'obiettivo di chiarire le ambiguità che ne derivano come, a titolo non esaustivo, la gestione responsabile dell'incertezza su decisioni e previsioni, la comprensibilità dei processi di apprendimento, l'interpretazione delle risposte e l'attribuzione di responsabilità. L'idea che l'etica e la tecnologia siano incomunicabili, infatti, è stata ereditata da una organizzazione funzionalista del lavoro e del pensiero che separa e distingue le questioni tecniche dai problemi sociali e, di conseguenza, schematizza le competenze necessarie per affrontarli, adattando soluzioni considerate maggiormente idonee per ciascuna categoria: lo strumento tecnico, scisso dalle potenzialità dell'utilizzo, veniva concepito come un oggetto amorale.

Nell'informatica teorica si definisce “automa” una macchina in grado di eseguire algoritmi, mentre il termine “intelligenza artificiale” contiene una carica immaginifica e una componente di indeterminatezza che lo differenzia dal lessico tecnologico e gli attribuisce le peculiarità di un sistema-agente artificiale (Fossa, Tamburrini, & Schiaffonati, 2021). Il governo della complessità negli ambienti digitali è stato in larga parte delegato ai sistemi automatizzati proprio per una maggiore capacità di gestione della mole crescente di contenuti. È stato presto evidente in alcuni settori il verificarsi di un esercizio del “giudizio” privo di supervisione e non per forza allineato a esigenze valoriali umane (*ibidem*). La “politica del codice” acquisisce un'impostazione volta alla gestione e, dunque, al controllo di porzioni della società. Lawrence Lessing, in “*Code and Other Law of Cyberspace*” (2006), scriveva che:

«Il cyberspazio ha una architettura che ingloba certi principi; detta i termini di come uno utilizza lo spazio; definisce ciò che è possibile nello spazio».

² Noam Chomsky si è espresso sull'argomento in “*Decoding the Human Mind & Neural Nets*” (23 giugno 2023), disponibile sul canale di Eye on AI: <https://www.youtube.com/watch?v=wPonuHqbNds>

Le decisioni sono ancorate all'inferenza che deriva dall'analisi dei dati digitali, da cui è possibile distinguere due tipi di regolazione algoritmica: *Code-driven*, quando l'impostazione degli standard è integrata con la variazione della condotta; *Data-driven*, con suggerimenti a supporto della decisione, in base a predizioni e inferenze sugli standard per monitorare, predire e influenzare il comportamento (Hildebrandt, 2018).

L'economia delle piattaforme nell'ecosistema occidentale è dominata dalle *Giant Corporations* che operano in segmenti specifici di mercato e sconvolgono le relazioni di lavoro, «aggirando le istituzioni, trasformando le pratiche sociali e civiche e influenzando i processi democratici» (Van Dijck, 2019) in diversi ambiti della vita collettiva, tra cui figurano i trasporti, la medicina, la prevenzione delle emergenze e l'educazione. Se le pratiche di raccolta (*data mining*) consentono di creare correlazioni, gli algoritmi delle aziende digitali procedono a renderle preziose: questo è il potere e il potenziale significativo dei *Big Data*, ovvero la quantità di informazioni prodotte dalle persone, dalle cose e dalle loro interazioni (Bucher, 2018, p. 32). Mentre nel settore privato si verifica la centralizzazione del comando, il settore pubblico arranca per accumulo di strutture amministrative (Musella, 2022, p. 13) e il processo di digitalizzazione procede a prescindere, nonostante l'assenza di allineamento.

La connettività non si traduce in automatico in valore pubblico (Van Dijck *et al.*, 2018) e una delle questioni più dibattute è ancora quella della privacy: la trasparenza del flusso di dati può convivere o scontrarsi con problemi come la sicurezza e la non discriminazione degli utenti e far emergere profonde differenze di visione tra Stati Uniti, Cina e Unione europea. L'ambiguità esistente nel rapporto tra stati, aziende e consumatori può rimettere in discussione i principi che hanno guidato la stesura delle norme.

Dal 2018, le “Linee Guida Universali sull'IA” (*Universal Guidelines for AI*, UGAI) sono il primo esercizio di respiro internazionale che chiede divieti, obblighi e diritti, stabilendo 12 principi per massimizzare i benefici e minimizzare i rischi (Tab. 1), tra cui l'obbligo di cessazione (*Termination Obligation*) che presuppone il controllo umano ultimo sui sistemi. Promosso dalla *Public Voice Coalition*, fondata nel 1996 dall'*Electronic Privacy Information Centre* (EPIC), il testo offre un quadro di *policy* rivolto ai governi e alle compagnie private basato sugli impatti sociali delle tecniche di calcolo automatizzato. È stato osservato che, oltre a un inconsistente meccanismo di supervisione, un ulteriore elemento di debolezza consiste nel rischio di obsolescenza di fronte a innovazioni dirompenti come la rapida diffusione della GenAI a livello di utenza di base.

Altre due iniziative sulla formulazione di principi etici hanno riguardato l'OCSE (OECD) e il G20 nel 2019 (Tab. 1). Nel primo caso, il documento è stato sottoscritto da 47 Stati (tra cui

USA, Cina e UE) come primo standard intergovernamentale sull'affidabilità dei sistemi (*Trustworthy AI*) prodotto a partire da consultazioni con esperti internazionali, industrie tecnologiche, accademie e rappresentanti del settore pubblico e, successivamente, adottato dal G20. I principi sanciscono la responsabilità degli *stakeholders* nei processi e il coinvolgimento nella promozione di benefici per l'umanità e il pianeta, ragion per cui gli sviluppatori sono direttamente interpellati per assicurare il corretto funzionamento (*AI Deployment*) e l'appropriata implementazione. Inoltre, l'OCSE ha promosso l'adozione dell'*AI Incident Monitor* (AIM) per la trasparenza in caso di malfunzionamenti o errori (*AI Failures*). La convergenza del G20 ha garantito l'armonizzazione normativa degli standard globali, cruciale per scalare le tecnologie e diffondere un approccio unificato di *governance* (Giappone e Canada, per esempio, si sono allineati), anche se ancora insufficiente dal punto di vista della responsabilità in caso di danno (*liability*) e in ritardo sull'adozione di uno schema regolatorio di valutazione degli impatti sociali. Sia le *UGAI* che gli *AI Principles* mancano di una delimitazione chiara, univoca e condivisa di cosa sia un sistema di intelligenza artificiale.

Gli standard etici - dallo sviluppo all'utilizzo dei sistemi - sono stati definiti nel 2021 con la *Recommendation on Ethics of AI* dell'UNESCO, adottata da 194 Stati membri e non vincolante (Tab. 1). Il testo ha anticipato il divieto di sorveglianza di massa e *Social Scoring*, promuove l'idea che le tecnologie siano funzionali agli Obiettivi di Sviluppo Sostenibile (*SDGs*) e introduce due strumenti: il *Readiness Assessment Methodology*³ (RAM) per valutare la capacità (legale e tecnologica) degli Stati, e l'*Ethical Impact Assessment* (EIA) per l'analisi del rischio (basato su privacy, equità, sicurezza e Diritti Umani), ancora fermo alle fasi iniziali perché vincolato a un reale coinvolgimento dei governi. L'*Artificial Intelligence and Democratic Values* (AIDV) Report del *Center for AI and Digital Policy* (2023) segnala che soltanto 15 Stati hanno effettivamente implementato la Raccomandazione UNESCO. Tuttavia, il testo è diventato un riferimento (*benchmark*) per le successive norme europee, dove si riscontra traccia nei preamboli. Questo spirito è stato recuperato anche dall'Assemblea Generale delle Nazioni Unite, principale forum multilaterale per codificare gli standard, che nella sua Risoluzione 2024 propone un *Global Framework for AI Governance* e legittima gli sforzi della società civile, mentre risulta manchevole dal punto di vista dell'allocazione delle risorse.

La *Resolution 473* sulla "Necessità di intraprendere uno studio sui Diritti Umani e dei popoli e sull'IA, la robotica e altre tecnologie nuove ed emergenti in Africa" è stata approvata nel 2021 e mira a ridurre il divario digitale e a promuovere l'utilizzo etico delle tecnologie (Tab. 1). Nel

³ Panoramica sociotecnica globale dell'IA, basata sul completamento della RAM: <https://www.unesco.org/ethics-ai/en/global-hub>

2024, l'*African Union* ha promosso una Strategia per l'IA con l'obiettivo di unificare gli approcci nazionali tra gli Stati membri e favorire il posizionamento del continente come leader nello sviluppo inclusivo e responsabile delle tecnologie in settori chiave quali la salute, l'educazione, il cambiamento climatico e l'agricoltura.

In Cina, la Strategia Nazionale sull'IA (*Next Generation AI Development Plan*) è stata presentata nel 2017 e successivamente potenziata dalle disposizioni sulla Gestione delle Raccomandazioni algoritmiche (Creemers, Webster, & Toner, 2022), comprendenti requisiti di licenza per i fornitori dei servizi, protezioni specifiche per gli anziani e il divieto di monopolio di mercato, sotto il coordinamento generale della *Cyberspace Administration* (CAC). Dal 2023, la Repubblica popolare ha rivolto un'attenzione privilegiata ai modelli di GenAI, fissando standard nazionali, metodi di etichettamento e criteri di cybersicurezza. Le *Interim Measures for the Management of GenAI* (Tab. 1) riconoscono la necessità di un'evoluzione legislativa continua e hanno giovato di una consultazione pubblica. Tra le vulnerabilità riscontrate emergono: l'assenza di supervisione indipendente e centralizzata; una classificazione poco chiara, con ambiguità latenti sul sostegno ai valori socialisti; limiti alla libertà di espressione nella critica al governo.

La *Saudi Data & AI Authority* (SDAIA), istituita nel 2019, ha sviluppato la *National Strategy* nel 2021 e promosso i *Saudi AI Ethics Principles* nel 2023 (Tab. 1). Sebbene siano in linea con le norme precedentemente menzionate, i principi non menzionano la sorveglianza di massa e le tecnologie di riconoscimento facciale, per cui risultano meno convincenti dal punto di vista della privacy. Il Regno dell'Arabia Saudita non figura tra i firmatari della Convenzione sulle armi letali autonome (*Lethal Autonomous Weapons*, LAW) del 2022.

Nel 2024 è stato promulgato l'*AI Treaty* (CETS n° 255) del Consiglio d'Europa, il primo trattato internazionale e vincolante nel settore. Il suo obiettivo è garantire che il ciclo di vita dei prodotti sia rispettoso dei Diritti Umani, della democrazia e dello stato di diritto. La Conferenza monitora e valuta la sua implementazione e ciò vale come meccanismo di *follow up*. Il documento esclude ambiti come la sicurezza e la difesa nazionale, non prende debitamente in considerazione gli impatti ambientali delle tecnologie, non enumera gli utilizzi vietati e/o insostenibili (*redlines*) e chiama in causa soltanto il settore privato come attore chiave nello sviluppo dei sistemi.

Tabella 1. *Principi etici per l'intelligenza artificiale e riferimenti normativi internazionali dal 2018.*

Documento	Anno	Proponente	Principi
<i>Universal Guidelines for AI</i>	2018	EPIC	• Trasparenza • Determinazione Umana • Identificazione • Equità • Valutazione e Responsabilità • Accuratezza, Affidabilità e Validità • Qualità dei dati • Sicurezza Pubblica • Sicurezza Informatica • Divieto di Profilazione Segreta • Divieto di Punteggio Univoco • Cessazione
<i>AI Principles</i>	2019	OCSE	• Benefici per l'Umanità e il Pianeta • Valori e Diritti Umani • Trasparenza ed <i>Explainability</i> • Robustezza e Sicurezza • Responsabilità
<i>AI Principles</i>	2019	G20	• Crescita inclusiva e sviluppo sostenibile • Valori umani e Giustizia • Trasparenza ed <i>Explainability</i> • Robustezza, Sicurezza e Affidabilità • Responsabilità
<i>Recommendation on Ethics of AI</i>	2021	UNESCO	• Diritti Umani e Dignità • Autonomia • Equità • Non Discriminazione • Trasparenza • Responsabilità • Sostenibilità • Impatto Ambientale • Governo • Supervisione
<i>Resolution 473</i>	2021	ACHPR	• Protezione dei Diritti Umani • Equità e Inclusione • Trasparenza e Responsabilità • Sviluppo economico • Non Discriminazione
<i>Interim Measures for the Management of GenAI</i>	2023	Repubblica Popolare Cinese	• Protezione dei dati • Regolazione dei contenuti • Trasparenza • Responsabilità • Equità
<i>Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI</i>	2023	USA	• Innovazione e competizione • Sicurezza • Supporto ai lavoratori • Non Discriminazione ed Equità • Protezione dei consumatori • Privacy • Uso federale e leadership internazionale

Documento	Anno	Proponente	Principi
<i>Saudi AI Ethics Principles</i>	2023	SDAIA	• <i>Human-centric AI</i> • Inclusività ed Equità • Privacy e Sicurezza • Trasparenza e Responsabilità • Gestione del Rischio
<i>Declaration of Santiago</i>	2023	OAS	• Diritti Umani • Governance • Inclusività • Bene Sociale • Sostenibilità Ambientale
<i>Declaración de Montevideo</i>	2024	America Latina e Caraibi	• Regolamentazione • Alfabetizzazione • Protezione dei Gruppi Vulnerabili • Ambiente e Sostenibilità • Capacità Tecnologica
<i>UN General Assembly Resolution</i>	2024	Nazioni Unite	• Giustizia • Sicurezza • Inclusività • Trasparenza • Diritti Umani • Responsabilità
<i>Artificial Intelligence Act</i>	2024	Unione europea	• Classificazione dei Rischi • Obblighi di Trasparenza • Protezione e Innovazione • Valutazione di Conformità e Sorveglianza
<i>Council of Europe AI Treaty</i>	2024	Consiglio d'Europa	• Non Discriminazione ed Equità • Trasparenza e Supervisione • Affidabilità • Dignità umana e Autonomia • Privacy • Responsabilità • Innovazione sicura
<i>Continental AI Strategy</i>	2024	Unione Africana	• Collaborazione e Cooperazione • Diritti Umani • Privacy & Sicurezza • Trasparenza, Responsabilità e Inclusione

Fonte: Elaborazione a cura dell'autrice a partire dall'*AI Policy framework* del CAIDP

Un caso problematico è rappresentato dai sistemi di IA integrati nei social media, in grado di individuare contenuti potenzialmente illegali, dannosi od offensivi per portare a compimento operazioni di moderazione su grandi quantità di contenuti. Superando visioni parziali della platformizzazione, occorre considerare in maniera olistica l'affermazione della *governance* algoritmica, la contemporanea influenza su istituzioni, scambi economici e pratiche sociali e il modo in cui gli Stati sono tenuti ad adeguare le proprie strutture legali e democratiche (Van Dijck *et al.*, 2018). Come verrà approfondito, il trasferimento completo e/o parziale di tali

mansioni potrebbe comportare rischi considerevoli per l'esercizio di diritti fondamentali come la libertà di espressione (Celeste *et al.*, 2022).

1.2 La normalizzazione della razionalità tecnologica nella sfera pubblica

Nel dibattito contemporaneo, l'esistenza di una connessione tra calcolo e coscienza è parte di un immaginario ancora privo di significazione. Ciononostante, si scorge una sensibilità sottesa per le sorti del lavoro manuale e intellettuale, probabilmente causata dalla rapida diffusione di modelli automatizzati tra le pratiche comuni dell'utenza. La comparsa di LLM, come ChatGPT, Claude e DeepSeek, ha in parte modificato modalità e tempistiche necessarie per l'espletamento delle mansioni nelle industrie culturali e della conoscenza. Le stime contenute in uno studio pubblicato da ricercatori di OpenAI, OpenResearch e dell'Università della Pennsylvania (Eloundou, Manning, Mishkin, & Rock, 2023) rivelano che otto lavoratori statunitensi su dieci potrebbero subire un calo del 10% delle prestazioni, a fronte del 50% previsto per gli altri, con effetti estesi a tutti i livelli salariali perché, a prescindere dalle dimensioni dall'azienda, il 15% di tutte le attività lavorative saranno completate molto più velocemente a parità di qualità, con una percentuale in crescita fino al 47-56% qualora *software* e strumenti venissero incorporati. Ecco perché, come sosteneva Ulrich Beck (2000, p. 67-69), la società occidentale nel XXI dovrebbe compiere uno sforzo di "contro-interpretazione" della razionalità tecnologica per non rivivere l'"immiserimento" dei lavoratori inglesi della Prima Rivoluzione Industriale, le cui condizioni di vita furono stravolte dallo sviluppo delle forze produttive, dell'integrazione del mercato, dei rapporti di proprietà e di potere. Il riferimento storico non è casuale: in "*The Human Use of Human Beings*", Norbert Wiener (1950) associa proprio al Settecento i primi segnali di una svalutazione del braccio umano in favore della macchina a vapore nel settore tessile-metallurgico, durante il quale antiche abitudini e stili di vita convissero con la gestione dell'economia di fabbrica. «Essa configurò una nuova relazione economica fra gli uomini, un nuovo sistema di produzione, un nuovo ritmo di vita, una nuova società, una nuova era storica», scrisse a sua volta Eric J. Hobsbawm (1972, p. 72). Il passaggio risulta interessante perché ai capitalisti che raggiungono il gradino più alto del podio si recriminano, ancora oggi a distanza di due secoli, bassi livelli di retribuzione per lavoratori non sindacalizzati: che siano essi mercanti, petrolieri o proprietari di piattaforme, il potere economico di cui dispongono contribuisce a rimodellare le basi materiali della civiltà (Staglianò, 2022).

Con l'introduzione delle innovazioni tecnologiche, le riflessioni di Adam Smith, Joseph Townsend, David Ricardo e Thomas R. Malthus su terra, lavoro e capitale si imposero come principi regolatori di una situazione di equilibrio fondata sulla contrattazione e implicarono

l'emersione di due figure sociali: il proprietario dei mezzi di produzione e il lavoratore (Polanyi, 1974, p. 253-255). Robert Owen, al contrario, osservò che nel mettere al centro del discorso l'individuo, i suoi contemporanei negavano le responsabilità della società nel processo di organizzazione che, inevitabilmente, avrebbe influenzato anche la macchina. Gli "interventi utili" dello Stato avrebbero allontanato il male dalla comunità (*ivi*, p. 277-279):

«La generale diffusione delle manifatture in un paese genera un nuovo carattere tra i suoi abitanti e poiché questo carattere è formato su un principio del tutto sfavorevole alla felicità individuale o generale, esso produrrà i mali più lamentevoli o permanenti a meno che la sua tendenza non venga contrastata da un'interferenza e da una direzione legislativa» (R. Owen, *Observations on the Effect of the Manufacturing System*, 1815).

Gli economisti classici classificavano il governo come lavoro "improduttivo", contribuendo a consolidare la percezione di un'economia emancipata, se non del tutto scissa, dalla politica. Quando Karl Marx elaborò la sua teoria, il dibattito sull'entità del valore andava evolvendosi con la meccanizzazione, che aumentava di pari passo con la produttività riducendo la forza-lavoro, cioè la principale fonte di profitto. Intuì che il capitalismo fosse dinamico e in costante mutamento: la società dimostrava di aver acquisito una capacità di adattamento alle "novità" che repentinamente modificavano le sue abitudini (Mazzucato, 2018, p. 54-55).

Fu in quegli anni che comparve sulla scena della Londra vittoriana la macchina analitica di Charles Babbage, l'antenata del computer mai costruita, i cui scopi furono innovati dal fortunato incontro con Ada A. Lovelace. A margine della sua traduzione dell'articolo di Luigi Menabrea (1842), la contessa formulò un algoritmo - contenuto nella Nota G (consultabile grazie al *Project Lovelace*) - per calcolare i numeri di Bernoulli, il primo a essere concepito come *software*, diventando la madre dell'informatica e la prima programmatrice della storia⁴. Descrisse come i codici potevano essere creati includendo lettere e simboli e teorizzò un metodo di ripetizione in serie delle istruzioni (*looping*). L'elaborazione da parte della macchina prevedeva, in futuro avrebbe acquisito la capacità di superare il mero calcolo, pur senza sviluppare una coscienza propria.

A partire dal 1870, anno in cui convenzionalmente si colloca l'inizio della Seconda Rivoluzione Industriale, elettricità e petrolio furono protagonisti di una irreversibile alterazione che modificò il sistema di sviluppo internazionale. Karl Polanyi ne "La grande trasformazione" (1974, p. 46) scrisse che il XIX secolo aveva causato il crollo delle quattro istituzioni garanti

⁴ In suo onore, nel 1979 fu ideato un nuovo linguaggio che unificava gli altri linguaggi di programmazione. Il film "Conceiving Ada" (1997) è ispirato alla sua figura, che compare anche nelle serie tv Victoria (2x02) e Doctor Who (12x02). Nel 2018 è stato fondato l'Ada Lovelace Institute, con lo scopo di garantire che «i dati e l'Intelligenza Artificiale siano al servizio delle persone e della società».

degli assetti politici: l'equilibrio di potere, l'organizzazione economica (la base aurea), il benessere prodotto dall'autoregolazione del mercato e lo Stato liberale. Léon Walras, William Jevons e Carl Menger fecero il resto, ampliando la teoria economica con i concetti di utilità ed efficienza: essi consideravano questi fattori costitutivi del valore e, dunque, in grado di determinare i costi di produzione. Come si rese più evidente con Alfred Marshall, lo studio del cambiamento marginale dell'utilità in base alle preferenze dell'individuo forniva uno standard soggettivo (Mazzucato, 2018, p. 66-69). I servizi iniziarono a essere inquadrati nella categoria "produttiva" perché il prezzo assunse un ruolo cruciale nella determinazione del valore di mercato. Il meccanismo di equilibrio dove si incontrano i livelli di domanda e offerta restituiva l'immagine del capitalismo come di un sistema pacifico (*ivi*, p. 70) che, grazie alla concorrenza, poteva autoregolarsi. Fu accolto positivamente per le contingenze storiche legate al periodo, in cui facevano il loro ingresso nell'agone politico le dottrine socialiste e le formazioni sindacali. All'aumentare delle richieste del commercio, matematici e scienziati, commenta Meredith Broussard (2018, p. 142-143), si trovarono dinanzi due opzioni: incentivare il cambiamento sociale per aumentare la forza lavoro (dando accesso agli schiavi e alle donne al sistema educativo e avviando processi di emancipazione) o costruire macchine che potessero sostituire il lavoro umano; la scelta di quest'ultima strada, forse, comincia col spiegare l'origine del pregiudizio maschile, razziale e di classe presente nella tecnologia. A tal proposito, Matteo Pasquinelli (2023) ricorda che sono la produzione e la committenza degli artefatti tecnici a rivelare la forma della società che li governa, derivando da modelli astratti la concretezza di relazioni sociali e comportamenti collettivi.

ENIAC, il primo computer, fu lanciato nel 1946, appena trent'anni prima di Apple Computer Company, poi divenuta Apple Computers Inc. Nel Dopoguerra, lo sviluppo dei calcolatori si affermò rapidamente, al punto da ridurre l'intervento umano nel lavoro computazionale. Wiener (1950, p. 151) osservava che alcune parti delle macchine fossero in grado di comunicare tra loro attraverso un linguaggio di programmazione «senza parlare con nessuno e senza ascoltare nessuno, se non nelle fasi terminali e iniziali del processo».

Le agenzie di intelligence statunitensi e la *Defense Advanced Research Projects Agency* (DARPA) sono state le principali promotrici della ricerca sull'intelligenza artificiale, indirizzando gli obiettivi primari su esigenze e priorità militari.

Nel 1955 il comparto navale finanziò il “*Summer Research Project on Artificial Intelligence*” di John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon che intendeva sviluppare una macchina in grado di simulare ogni aspetto dell’apprendimento e dell’intelligenza umana per mezzo di calcoli matematici:

«Per il presente scopo il problema dell’intelligenza artificiale è quello di far sì che una macchina agisca con modalità che sarebbero definite intelligenti se un essere umano si comportasse allo stesso modo» (McCarthy, Minsky, Rochester, & Shannon, 1955).

Il documento aiuta a comprendere come è mutato lo studio della disciplina, ancora priva di una definizione unitaria. In seguito, IBM aggiornò con nuovi *dataset* le macchine attraverso cui sarebbe stato possibile elaborare non soltanto numeri, ma anche lettere, parole e documenti, espandendo la classe di dati mediante un approccio simbolico ai sistemi (Kaplan, 2016, p. 29). Gli studi di Alan Turing (1950) sono la base teorica per i moderni sistemi di IA. L’*Imitation Game* si fonda sul rifiuto alla domanda “Le macchine possono pensare?”: il test si struttura sulla partecipazione binaria di un uomo (A) e una donna (B) interrogati da un terzo (C) su una serie di quesiti a cui rispondere in forma scritta (X e Y), in modo tale che non siano decifrabili gli indizi, come il tono di voce; lo scopo è associare A e B a X e Y, sapendo – scriveva - che A è più incline a imbrogliare, mentre B avrà la tendenza ad aiutare C. Alla domanda "Cosa accadrà quando una macchina prenderà il posto di A in questo gioco?", l’ipotesi che ha raccolto maggiori consensi tra gli informatici concerne l’eventualità che C non riesca più a distinguere tra un essere umano e un computer, deducendo che quest’ultimo sia in grado di pensare. L’associazione *hardware*/corpo e *software*/cervello fu affiancata dalle prime simulazioni della struttura del neurone nei computer, dove *input* e *output* creavano connessioni autonome attraverso reti a più livelli: questi programmi, al quale il programmatore sottoponeva esempi di corretta esecuzione, vennero chiamati “*network* neurali” (Kaplan, 2016, p. 33).

In “*What Computers Can’t Do*”, Hubert L. Dreyfus osservò criticamente che l’intelligenza e la competenza hanno a che fare con fattori inconsci e subconsci (conoscenze implicite e induttive), mentre i computer richiedono che ogni procedimento e dato sia esplicito e formalizzato (Crawford, 2021, p. 6). L’abilità conversazionale dei *bot* (diminutivo di *robot*) per risolvere problemi o chiarire dubbi, per esempio, non comporta che lo strumento sia dotato di coscienza. Negli anni Sessanta, l’industria dei semiconduttori venne stravolta dall’introduzione del silicio, che permise la diffusione di massa del *Personal Computer* (PC). I progressi del settore tecnologico provenivano da ricerche condotte in collaborazione tra pubblico e privato (in laboratori come i Bell Labs di AT&T), ma di lì a poco la Silicon Valley si affermò come centro

dell'innovazione in cui imprenditori e industrie private lucravano sui frutti del finanziamento pubblico, che valse il titolo di “*Internet California Gold Rush*” (Mazzucato, 2014, p. 143-144). La Terza Rivoluzione Industriale, che ebbe convenzionalmente inizio nel 1970, comportò il definitivo *switch* dall'energia all'informazione e la riaffermazione del primato del mercato nella risoluzione dei problemi economici e sociali da parte della corrente neoclassica, con il consolidarsi del principio ricardiano fondato sull'accumulazione di profitto a partire dalla redistribuzione di risorse scarse (Mazzucato, 2018, p. 61-62). L'esclusione dalla produzione, applicata nell'epoca preindustriale con le *enclosure*⁵, le recinzioni delle terre comuni, si è ripresentata in tempi più recenti con la proprietà intellettuale e con i brevetti che, concedendo a un soggetto il monopolio limitato sulla sua invenzione, sottendono una «filosofia implicita di proprietà privata e dei suoi diritti» (Wiener, 1950, p. 113).

La nascita di Internet ha aperto nuove pagine per la democrazia, insinuandosi prepotentemente nel sentito comune sotto l'egida degli USA. La “teoria del tutto” – che per la fisica teorica equivale a unificare le quattro interazioni fondamentali (gravitazionale, elettromagnetica, debole e forte) – si è imposta come un'idea perseguibile fino a tradursi nella “teoria delle stringhe”, ovvero l'ambizione a creare un multiverso contenente «spazio, tempo, materia, energia, informazione e le leggi che li descrivono» (Staglianò, 2022, p. 49), che accompagna le più grandi industrie verso processi di integrazione. Basta pensare alla galassia Apple tra comunicazione cellulare, *mobile computing* e app di intrattenimento in un unico dispositivo (Mazzucato, 2014, p. 151) e al successivo inserimento di Siri⁶.

In “*The Discipline of Machine Learning*”, la testimonianza di Tom M. Mitchell spiega il funzionamento dei sistemi informatici:

«Diciamo che una macchina apprende rispetto a un particolare compito T, a una metrica di prestazione P e a un tipo di esperienza E, se il sistema migliora in modo affidabile la sua prestazione P nel compito T, in seguito all'esperienza E. A seconda di come specifichiamo T, P ed E, il compito di apprendimento può anche essere chiamato con nomi come raccolta di dati (*data mining*), scoperta autonoma (*autonomous discovery*), aggiornamento del database (*database updating*), programmazione per esempio (*programming by example*), ecc...» (citato in Broussard, 2018, p. 169).

Ciò ci invita a considerare l'esperimento di Turing replicabile allo stesso modo su oggetti immutabili, ma la società modifica e negozia continuamente l'interpretazione di concetti come il genere, inquadrandolo come un *continuum* (Broussard, 2018, p. 69-71), la razza e la classe di

⁵ Ritenute essenziali per la fornitura di materie prime per l'industria, Vandana Shiva sostiene che ora sono biodiversità e conoscenza a essere “racchiuse” nei diritti di proprietà intellettuale (citata in anatomyof.ai).

⁶ Nel 2000 la DARPA si rivolse allo Stanford Research Institute (SRI) per coordinare un progetto finalizzato allo sviluppo di un “segretario virtuale”. Il progetto CALO (*Cognitive Assistant that Learns and Organizes*) fu messo in commercio mediante una *start up*, la Siri, che fu acquistata da Apple nel 2010 (Mazzucato, 2014, p. 155).

appartenenza, mentre l'IA è una rappresentazione delle “verità” appartenenti a un periodo finito (Arendt ripresa in Browne *et al* 2023, p. 333). Il test contiene una codificazione culturale in cui sono presenti discriminazioni fisiche e morali che partono da costrutti sociali, come il più banale esempio del taglio di capelli, che presuppone una connotazione sessuale implicita.

Crawford (2021) ritiene che l'IA non sia né intelligente né artificiale, bensì integrata e materiale, composta da risorse naturali (carburante, lavoro umano, infrastrutture, logistica, storie e classificazioni) e fondata su un volano di strutture politiche e sociali: a causa del capitale richiesto per costruirla e delle possibilità generate dai suoi sistemi di ottimizzazione, è progettata per servire gli interessi dominanti esistenti (in altri termini, è un “registro del potere”). In tempi più recenti si sono affermate nuove teorie tra i gruppi di interesse industriali, in particolare nel mondo *tech*, fino a insediarsi presso governi e istituzioni internazionali. Tra queste, emerge l'altruismo efficace (*Effective Altruism*) che sostiene il finanziamento di progetti meritevoli per affrontare le sfide del futuro. La teoria, in sostanza, annulla il conflitto a favore delle cause e individua il “giusto” perseguibile in un rapporto di equazione dove il legame con l'utilitarismo è evidente: il bene (ciò che produce il maggior grado di piacere) e l'utile (per il maggior numero di persone), si equivalgono. A tal fine, i suoi seguaci perseguono l'accumulo di ricchezze mediante attività lucrative per reinvestirle in cause collettive, come la povertà nel Sud Globale. Si affianca al lungoterminismo (*Longterminism*), una dottrina considerata come un credo o un'ideologia fondata sul seguente principio: «Le azioni che compiamo oggi – sia noi come individui singoli sia collettivamente, come società – devono essere mosse dall'intenzione di migliorare il futuro a lungo termine di tutta l'umanità» (Doda, 2024, p. 20), valutando il tempo come dimensione centrale e la portata come misura del gesto. Ne hanno scritto Hilary Greaves e Willian MacAskill, il *Future of Humanity Institute* dell'Università di Oxford e il Future for Life Institute, fondato da Elon Musk. Le azioni compiute nel presente scongiurano il “rischio esistenziale” per le forme di vita intelligenti sulla Terra mediante l'accelerazione del processo economico e tecnologico, da cui consegue che le vite del tempo corrente contano quanto quelle delle prossime generazioni; è il manifesto dell'umanità in potenza basato sull'applicazione di concetti matematici a scelte etiche (*ivi*, p. 31).

Tra le voci critiche emerge quella di Émile Torres, che nel saggio “*Against Longterminism*” denuncia lo scarso valore attribuito alle singole persone, a favore, invece, di una concezione edonistica del “valore” inteso come piacere o felicità. Il rapporto con la tecnologia appare del tutto ambiguo: grandi esponenti si presentano come sostenitori della regolamentazione dell'IA

(c'è anche Sam Altman di OpenAI⁷) perché sono responsabili dello sviluppo e del commercio ma ne riconoscono i rischi, anche se considerano le tecnologie imparziali.

L'impostazione filantropica del *Longterminism* adotta un approccio gerarchico mediante erogazione unilaterale di fondi e gestione istituzionale *top-down* (Doda, 2024, p. 57). Questo modello si ripete anche nell'interpretazione della realtà, diffondendo l'idea che esista una comunità universale che condivide un destino, mentre si sostituisce allo Stato nel soddisfare i bisogni e crea criteri propri, scissi da una logica di comunità, per individuare meritevoli e non (Santoriello, 2024). All'orizzonte si intravede anche una corrente più libertaria e antistatalista, l'accelerazionismo efficace (*E/Acc*), che appare opposta perché fondata sul tecno-ottimismo. Guillaume Verdom (dopo Google, ora a capo della start up Extropic) è uno dei suoi volti ed è convinto che si possa programmare un sistema di civilizzazione.

Persone fisiche e Corporation possono essere parimenti committenti, agenti di reato ed essere incriminate per i danni commessi dalle intelligenze sintetiche che creano o detengono. La cornice legale è lo stato di persona⁸ (Kaplan, 2016, p. 80-85) che attribuisce a soggetti legali diritti e responsabilità. Queste sono le basi attraverso cui sono costruite le IA, il modo in cui decifrano l'ambiente e come si orienta la percezione di aspetti moralmente rilevanti.

1.3 La matrice di Corporation digitale

Seguendo un orientamento consolidato di analisi concettuale nelle scienze sociali (Calise & Lowi, 2010), è possibile costruire una matrice per osservare le trasformazioni che hanno interessato lo scenario tecnologico internazionale. La dimensione analitica è definita da due assi, il cui incrocio forma quattro spazi di osservazione. Ciascun quadrante, circondato dal concetto chiave e da due poli, contiene un sottotipo e un concetto chiarificatore che chiude lo spazio logico. La combinazione periferica restituisce il contesto nel quale si inserisce il concetto centrale. La ricerca recupera questa tradizione politologica per indagare il ruolo delle Corporation, identificate dal Dizionario Britannica come «una specifica forma giuridica di organizzazione di persone e risorse materiali, istituita dallo Stato, allo scopo di condurre affari», che risale al XI secolo. Si tratta di un ente artificiale istituito da una carta reale, da una prescrizione o da un atto legislativo, che conserva alcuni diritti di successione e, in tempi più recenti, un'organizzazione guidata dall'economia di mercato.

⁷ Si veda Rullo, L. (2024). *(Open) AI saga*, Rivista di Digital Politics, 4(1), 29-48.

⁸ Nel 1819 la Corte suprema stabilì che il Dartmouth College fosse autorizzato a proteggere la clausola contrattuale della Costituzione degli USA (art. I-10, comma 1), avviando il percorso verso il riconoscimento della personalità giuridica delle Corporation.

Come affermano Calise e Lowi (2010):

«Attraverso la corporazione, l'autorità politica poteva essere affermata su basi collettive piuttosto che sui legami personali della società feudale. Per molti aspetti, la Corporation può essere considerata come il prototipo dell'esperienza del potere statale» (p. 90).

Le Corporation digitali (Fig. 1) raccolgono in eredità il risultato di una lunga evoluzione che, fin dai conglomerati industriali del XX secolo⁹, è giunta alle più contemporanee imprese multinazionali in grado di realizzare profitti attraverso una produzione immateriale di informazioni, interazioni e tracce visive. La digitalizzazione ha consentito alla produzione di massa di espandere i propri mercati con all'attivo milioni di destinatari pronti a prendervi parte grazie alla progressiva riduzione delle barriere all'accesso (come l'accesso alla Rete e il possesso di un dispositivo elettronico), nello stesso modo in cui l'industrializzazione aveva consentito a Harry Ford di vendere il modello standardizzato T a centinaia di nuovi clienti. L'innesto tra l'iniziativa individuale e l'innovazione tecnologica è, secondo Hobsbawm (1972, p. 46-47), nel conseguimento dei profitti. La Corporation è interessata alla creazione del valore, vale a dire al processo mediante il quale la ricchezza produce oggetti tangibili e intangibili:

«Il valore può essere definito in molti modi, ma sostanzialmente consiste nella produzione di nuovi beni e servizi. Come queste merci vengano prodotte (produzione), come vengano distribuite all'interno dell'economia (distribuzione) e cosa si faccia del profitto che viene creato dalla loro produzione (reinvestimento) sono questioni chiave per definire il valore economico» (Mazzucato, 2018).

Misurare in cosa consista o meno il valore non è una considerazione neutra; anzi, è legata a una visione che distingue tra attività produttive e improduttive sulla base di una determinata idea di società (*ivi*, p. 17). Gli antichi dibattiti del Novecento tra Stato e Mercato incontrano nessi logici e trasformativi nella dottrina della gestione, da intendersi come aperta o chiusa, pubblica o privata. Pensiamo a due competitor come Google e Wikipedia: due motori di ricerca, entrambi aperti (non vi è richiesta alcuna iscrizione), sono in realtà diametralmente opposti nella gestione, passando da un modello di *business* fondato sulle inserzioni a un modello a trazione non-profit (Taylor, 2014, p. 23). Il potere di definire le caratteristiche e le linee operative delle piattaforme determina il controllo dell'ordine sociale e delle pratiche, perché esse sono "reciprocamente costitutive" (Calise & Musella, 2019, p. 19).

Le Corporation digitali sono la base delle piattaforme di social media, che si suddividono in 4 tipologie (Kennedy, 2016, p. 21): i siti di *Social Network* (SNSs) come Facebook, X (già Twitter) e LinkedIn, che promuovono il contatto e la comunicazione interpersonale; i siti per la

⁹ L'elaborazione originaria del concetto di "Corporation" è stata inserita da Calise e Lowi in "Hyperpolitics" (2010, p.17), dove si legge: «Due matrici concorrenti dello stesso concetto possono essere facilmente e sistematicamente confrontate, perché tutte le matrici fanno parte dello stesso sistema logico».

condivisione di contenuti creativi prodotti dagli utenti (*User-Generated Content*, UGC) come YouTube, Flickr, Instagram, TikTok e REDnote; i siti per la compravendita (*Trading and Marketing*, TMSs) come Amazon, eBay e Craigslist; i siti per il gioco (*Play and Game*, PGSs) come FarmVille, The Sims e Roblox. Questa categorizzazione non è rigida perché subisce l'effetto di convergenza dovuto alle integrazioni orizzontali e alla tendenza a creare ambienti digitali multifunzione per trattenere più a lungo gli utenti, come nel caso dell'app WeChat di Tencent, tramite cui è possibile scambiare messaggi, fare videoconferenze, condividere foto/video ed effettuare pagamenti.

Van Dijck (2019) propone una mappatura del mondo globale online, dominato da aziende e Stati, sviluppata a partire da due modelli geopolitici: l'ecosistema cinese, dove il mercato interno è sottoposto al controllo centralizzato (attraverso il *firewall*) e gestito da compagnie nazionali (Ali-baba, Tencent, Baidu, Jingodon Mall e Didi); e l'ecosistema statunitense, divenuto prevalente in Europa, Asia, Africa e America Latina con le sue *Big Five* (Alphabet, Meta, Amazon, Apple e Microsoft), in cui è il mercato a controllare l'infrastruttura. Essi non sono del tutto antitetici: le GAFAM hanno dimostrato interesse ad adeguarsi alle norme (è il caso del *Dragonfly Project* di Google) e le aziende cinesi hanno ottenuto partecipazioni nelle aziende occidentali. In Europa, le cosiddette “unicorno” (aziende tecnologiche valutate più di 1mld di dollari) sono poche e di dimensioni ridotte: persino le principali, Spotify (Svezia), BlaBlaCar (Spagna), Adyen (Olanda) e Transferwise (UK/Estonia), non riescono a competere con le *gatekeeper* del traffico online. Va da sé che l'impatto sociale delle compagnie statunitensi non si configura soltanto in termini di quote di mercato e fatturato, bensì come potere normativo (*ibidem*) a cui l'Unione europea ha risposto con una rigorosa produzione legislativa nell'ultimo decennio, affinché i valori pubblici, ovvero le libertà fondamentali e la dignità umana, non periscano sotto il peso degli interessi commerciali (i casi più rilevanti hanno riguardato l'IA e i servizi digitali, la privacy, il mercato concorrenziale, l'evasione fiscale, la condanna di *fake news* ed *hate speech*). Ciononostante, le Corporation suppliscono a una lacuna infrastrutturale dello Stato entrando in possesso di dati sensibili in settori cruciali come la sanità, l'istruzione e l'informazione, e finiscono per assumere alcune funzioni vitali degli enti pubblici:

«Le piattaforme diventano in effetti sempre più spesso fornitori di infrastrutture. Come ha osservato Mark Zuckerberg nel 2017, Facebook vuole essere una “infrastruttura sociale”, un termine che evoca la nozione di servizio pubblico» (*ivi*, p. 165).

La tensione tra Stati e Corporation digitali è poco percepita dall'opinione pubblica occidentale a causa di una narrazione sensazionalistica incentrata sull'*hype* dei prodotti e sulla conseguente commercializzazione. Secondo Mazzucato, Karl Marx non si sarebbe sorpreso della capacità della tecnologia di modificare la società (2018, p. 58-59): la sua distinzione tra capitale

produttivo (industriale) e capitale commerciale è, infatti, ancora applicabile. Amazon, per esempio, è un capitalista commerciale perché ha creato le condizioni per permettere ai capitalisti produttivi di vendere le merci e creare plusvalore. Dal punto di vista dei governi, invece, l'IA è considerata di interesse strategico in termini sia industriali, a giovamento delle filiere produttive, sia militari e di controllo interno. Si stima che, nel 2030, la ricchezza generata dall'IA sarà distribuita in maniera diseguale su più livelli: USA e Cina intercetteranno circa due terzi del fatturato mondiale, mentre il restante sarà spartito tra l'Europa, le economie più avanzate e, infine, i Paesi in via di sviluppo (Dini, 2023).

La distribuzione di grandi piattaforme e motori di ricerca in UE cristallizza l'esistenza di una "Guerra fredda dell'IA". Il modello europeo (detto renano) dello Stato sociale avrebbe, quindi, un ruolo chiave nella regolazione pubblica delle piattaforme perché in grado di bilanciare i poteri dello Stato, del mercato e della società civile (Van Dijck, 2019).

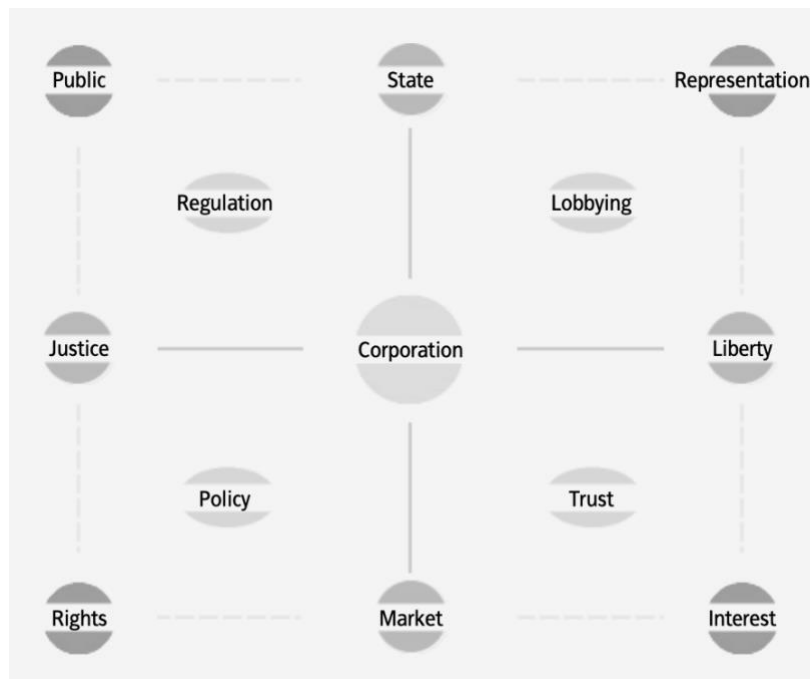


Figura 1. Matrice di Digital Corporation (2022).

1.3.1 Stato e mercato nell'ecosistema digitale

Le istituzioni sull'asse verticale sono lo Stato e il Mercato (Fig. 1), che rappresentano la dicotomia pubblico-privato basata sul livello di separazione tra la sfera politica e la sfera economica, da cui deriva, nel primo caso, il modello centralizzato e, nel secondo, il modello del *laissez-faire*. L'idea che il mercato sia autonomo rispetto alla legge era considerata un'utopia già da Polanyi (1974, p. 47), nonché una condizione che avrebbe rischiato di

disumanizzare la società. Tra il 1815 e il 1914 si consolidarono nuovi equilibri di potere in Occidente, guidati dall'interesse comune al mantenimento della pace: le guerre non convenivano più agli Stati per ragioni sia economiche che commerciali. In una cornice socioeconomica, l'affermazione del profitto come principio regolatore nelle relazioni internazionali aveva reso la Rivoluzione Industriale il motore della civiltà economica (Polanyi, 1974, p. 98-99). Come mai avvenuto prima, Stato e Mercato iniziarono a contendersi l'attività di governo e l'organizzazione della società¹⁰. È il periodo in cui le teorie di Keynes promuovono l'utilizzo della spesa pubblica per rafforzare la domanda e gli economisti ispirati a Schumpeter sostengono che gli investimenti dovrebbero essere indirizzati verso quei settori che accrescono la capacità di innovazione della nazione (Mazzucato, 2014, p. 62), sotto forma di ricerca e sviluppo (R&S), infrastrutture, formazione, incentivi diretti e indiretti a tecnologie e aziende. Dal punto di vista di Haggart e Tusikov (2023) la *governance* di dati e IP dovrebbe essere posta sullo stesso piano della politica monetaria vista l'influenza che esercita sul benessere dei cittadini, mentre il tecno-feudalesimo (USA) e il protezionismo digitale (Cina, UE, Canada), corrono entrambi il rischio di creare forme di colonialismo tra Nord e Sud Globale. Le Corporation esistono in relazione allo Stato e non detengono autorità di governo a causa di una limitata capacità sia di accesso alle risorse che di rappresentatività del pubblico¹¹.

Nella storia delle telecomunicazioni, il caso di AT&T negli USA ha segnato l'inizio di un processo di controlli ed equilibri con l'intervento pubblico nel commercio. Il direttore generale della Bell Telephone Theodore Vail considerava la sua impresa "un sistema, una politica, un servizio universale" e accettò di servire come *Common Carrier* (Wu, 2011), assicurando un accesso aperto ed equo ai concorrenti. Con l'acquisizione di Western Union, gli attori indipendenti del mercato furono invitati a unirsi alla società e a dividerne i profitti. Nel 1968 la *Federal Communications Commission* (FCC) obbligò la società AT&T a utilizzare una presa telefonica universale come connettore e tentò di creare tre *pool* di concorrenza distribuiti in servizi a lunga distanza, collegamenti e servizi di elaborazione dati. In qualche modo, la protezione concessa all'industria nascente dei servizi informatici online ne garantì anche lo sviluppo. Proprio come il *Trust Movement*, che riteneva impossibile il verificarsi della concorrenza spietata di Adam Smith in un'economia industrializzata, molti leader della Silicon

¹⁰ Secondo altri autori, il mercato nazionale - il più capitalista rispetto ai corrispondenti mercati locali e internazionali - è una creazione dello Stato, sfatando il mito della contrapposizione tra le due forze (Mazzucato, 2014, p. 61).

¹¹ La Cina, con l'ingresso nell'Organizzazione mondiale del commercio (WTO), ha rispettato la strategia di industrializzazione aspettando che le aziende nazionali fossero abbastanza mature per competere; il modello del Fondo monetario internazionale, al contrario, nega il ruolo attivo dello Stato nello sviluppo, come avviene in Germania, Gran Bretagna e Stati Uniti (*ivi*, p. 73).

Valley sostengono che sia il monopolio a guidare il progresso (Wu, 2018, p. 17). La reazione di AT&T fu quella di schiacciare le start up concentrando il mercato, finché il governo non decise, infine, di sciogliere l'azienda nel 1984.

Mutatis mutandis, pur non escludendo la proprietà fisica delle infrastrutture e dei rispettivi depositi (Blum, 2012, p. 301), una peculiarità delle piattaforme è il funzionamento determinato da algoritmi alimentati da dati invisibili, che richiederebbe una revisione degli strumenti di tutela. Con Amazon, per esempio, i prezzi possono calare sensibilmente per criteri stabiliti dall'azienda stessa. Un altro aspetto largamente sottovalutato, inoltre, è il lavoro gratuito svolto dagli utenti o quello invisibile e sottopagato di *Microworkers* o *Crowdworkers* (le persone che etichettano i dati o svolgono la revisione di contenuti sospetti) per alimentare l'IA (Gillespie, 2018; Crawford, 2021). L'estrazione stessa dei dati è avvenuta a partire da beni comuni commercializzati, in una sorta di privatizzazione della conoscenza.

Nel marzo 2018 è stata diffusa una notizia che coinvolgeva Facebook e una società di analisi. Raccogliendo in modo inappropriato i dati dai profili di circa 87 milioni di utenti, Cambridge Analytica aveva creato inserzioni personalizzate che miravano a influenzare le preferenze di voto delle persone in vista delle elezioni presidenziali USA del 2016. La tecnica del *Microtargeting* aveva sfruttato i dati personali in maniera non consensuale, condividendo disinformazione e sfumando i confini del contenuto pubblicitario (Heawood, 2018). Questa vicenda ha dimostrato il potere dei social media di condizionare un appuntamento democratico come le elezioni. Mentre si scrive, il mercato digitale risulta ancora in parte deregolamentato, favorendo l'acquisizione indebita di dati e l'accumulazione di capitale da parte di pochi proprietari. Lo Stato, invece, resta l'entità più potente nella difesa dell'interesse pubblico e l'unica incaricata di farsi carico di una cultura sostenibile, che ponga al centro del dibattito anche gli impatti ambientali delle tecnologie (Taylor, 2014, p. 2022).

1.3.2 L'altalena dei diritti tra Giustizia e Libertà

Nel perimetro disegnato dalle piattaforme, ogni utente è sia cittadino che consumatore¹². Sull'asse orizzontale (Fig. 1), Giustizia e Libertà sono concetti opposti, il cui orientamento oscilla su un'altalena sbilanciata tra la tutela dei diritti per la *Web Community* e la garanzia di iniziativa economica per i proprietari delle Corporation digitali. Secondo le più note teorie liberali, "Libertà" significa "protezione contro la tirannia dei governanti politici", inquadrati

¹² In anatomyof.ai, Kate Crawford scrive che l'utente di uno *speaker smart* Amazon Echo è «contemporaneamente un consumatore, una risorsa, un lavoratore e un prodotto».

come necessari antagonisti del popolo (Mill, 1859). Attraverso il riconoscimento delle immunità e l'istituzione di controlli costituzionali, la comunità esprime il suo consenso a essere governata. “Giustizia”, al contrario, è una parola collettiva intesa nella sua concezione democratica (Calise & Lowi, 2010). Nel corso del XX secolo, la diffusione della responsabilità sociale d'impresa aveva trasferito la discrezionalità dell'azione correttiva al *management*, imponendosi come dottrina dominante fino ai primi anni Duemila. Dal momento che la società negozia pratiche e rivendicazioni, la risignificazione del termine “Giustizia” nel digitale sottende ora l'urgenza di intervento sui pregiudizi algoritmici (sommersi da influenze culturali), sull'accumulazione incontrollata di capitale e la sua influenza sulla democrazia.

Per riflettere e trasferire un ordinamento democratico, la società civile chiede di implementare valori pubblici e interessi collettivi nella progettazione perché l'attuale ecosistema delle piattaforme è basato su un *design* «fondato e guidato da valori economici e interessi aziendali» (Van Dijck *et al.*, 2018, p. 139). Noto in letteratura come “capitalismo della sorveglianza” (Zuboff, 2019), la sua diffusione ha comportato la sedimentazione di tecniche di estrazione basate sulla profilazione dei *Data Subject* (agglomerati di dati), cioè delle abitudini di consumo degli esseri umani. Il passaggio successivo è stato sviluppare nella cultura di massa la falsa idea che le tecnologie intelligenti siano neutrali, oggettive e in grado di comprendere.

In realtà, tutta questa produzione umana è basata su standard scientifici che riflettono la visione sociale e culturale di coloro che li hanno sviluppati (Barassi, 2021, p. 122-124). La maggior parte degli errori algoritmici marca letture errate del corpo umano, per esempio basate su un'assunzione unilaterale del corpo incapace di distinguere il colore della pelle (*ivi*, 125-126). La pubblicazione del volume “*Data Feminism*” (D'Ignazio & Klein, 2020) ha acceso i riflettori sull'urgenza di agire sulla *Data Science* per favorire una *Data Justice*, rispettosa dell'intersezionalità e della complessità sociale, a partire da una risemantizzazione del campo. Privacy, consenso e governo dei dati non sono tecniche, bensì fondamenti della nostra società (Dencik *et al.*, 2019).

Nel mondo della tecnologia, la libertà si rivela sempre una variabile darwiniana e l'apertura a cui ambisce non difende l'equità, bensì la competizione (Taylor, 2014, p. 24). Karl Polanyi (1974, p. 84) sosteneva che il liberalismo economico avesse frainteso la storia della Rivoluzione Industriale valutando gli eventi sociali da una prospettiva economica laddove «un processo di cambiamento indiretto, il cui ritmo è ritenuto troppo veloce, dovrebbe essere rallentato per salvaguardare il benessere della comunità».

1.3.3 Quattro fenomeni della digitalizzazione

I quattro quadranti della matrice di Corporation digitale (Fig. 1) descrivono i fenomeni empirici e costituiscono, a loro volta, dei concetti.

Il quadrante in basso a destra restituisce uno dei tratti distintivi del neoliberismo: pochi *trust* (come i conglomerati tecnologici GAFAM – Google-Alphabet, Amazon, Facebook-Meta, Apple e Microsoft) controllano ampie quote di mercato, avvicinandosi a situazioni di monopolio. Per questo motivo, il termine *Trust* è usato come sinonimo per entrambi. I monopoli cercano di indirizzare le "innovazioni sostenibili" per i loro prodotti, ma sopprimono le "invenzioni dirompenti" che possono distruggerli (Wu, 2011); nel già citato caso di AT&T, fu il giudice federale a intervenire per vietare le interferenze da parte dell'azienda e consentire agli accessori telefonici prodotti da terze parti di collegarsi ai dispositivi telefonici.

Le *Big Tech* hanno superato il traguardo dei cinque anni di obsolescenza espandendosi con nuovi servizi gratuiti a disposizione di ogni utente e acquistando prodotti apparentemente concorrenti. Meta è riuscita a portare a termine 67 acquisizioni, Amazon ne ha intraprese 91 e Google addirittura 214 (alcune delle quali condizionate). Nel giro di pochi anni, i motori di ricerca alternativi a Google sono scomparsi dalla rete, i negozi indipendenti hanno ceduto ad Amazon e chiunque non avesse un account Meta è stato escluso dal dibattito pubblico, come descritto da Tim Wu in "*The Curse of Bigness*" (2018, p. 98). La popolarità crescente di Instagram ha reso la sua acquisizione da parte di Facebook una violazione della legge *antitrust* statunitense, in quanto tale strategia derivava dal desiderio di eliminare dal mercato un *competitor* tra i social media per la condivisione di foto. Lo stesso è accaduto per WhatsApp, la principale applicazione di messaggistica istantanea. Parallelamente, sono stati avviati alcuni progetti paralleli: sono nati *Marketplace*, l'app di incontri *Dating*, infine sono state annunciate – salvo poi ritirarle – la moneta *Libra* e il suo progetto più ambizioso, il *Metaverso*.

Ciascuna azienda non distribuisce mai un solo prodotto, ma attua continue espansioni orizzontali: Amazon è un "*Everything store*" (Staglianò, 2022, p. 51), un fornitore di servizi informatici (Amazon Web Services), in particolare di *cloud*, di *streaming* per video e musica (Amazon Prime), un rivenditore di lettori *ebook* (Amazon Kindle) e ha ottenuto alcuni *competitor* come Zappos, Diapers.com e Soap.com; parallelamente, il sistema operativo Android appartiene a Google, che ha comprato YouTube per la condivisione di video e Waze per il servizio di mappe.

Scrivo Tim Wu (2018, p. 96):

«Purtroppo, la legge antitrust non si è accorta che gli anni '90 erano finiti. Invece, per un decennio e più, ha concesso ai principali operatori tecnologici un lasciapassare, anche quando si sono trovati di fronte a pericoli piuttosto evidenti e a fusioni anticoncorrenziali».

Tra il 2006 (anno dell'acquisizione di YouTube) e il 2017, le *Big Five* sono cresciute a dismisura, raggiungendo una capitalizzazione pari al PIL della Germania (Calise & Musella, 2019, p. 18). Mentre ipotesi come la nazionalizzazione dei *provider* non sembrano convincere, l'imposizione di oneri minimi trova oggi sostegno (Taylor, 2014, p. 225). Secondo la presidente della Federal Trade Commission Lina Khan, esistono due strade per regolare i potenziali monopoli digitali: la prima è evitare le integrazioni verticali; la seconda è accettare gli oligopoli limitandone il potere, per poi inquadrarli come *Public Utilities*, pretendendo la non discriminazione sui prezzi e il *Common Carrier*, o come *Essential Facilities*, con la previsione dei soli oneri di condivisione (Staglianò, 2022, p. 103-104).

Istituzioni e imprese ricorrono alle *Policy*, un'ampia categoria che comprende sia i termini di servizio che gli standard tecnici, prendendo decisioni in merito a procedure e funzionalità. La *Platform Transience* (Barrett & Kreiss, 2019) delle Corporation digitali ha fini economici e riguarda l'articolazione degli asset organizzativi e pubblicitari, nonché di risoluzione delle controversie, comportando cambiamenti significativi per lavoratori e utenza.

Nel quadrante in basso a sinistra, si annovera il caso della creazione *ad hoc* di una giustizia parallela, elaborata da Meta con l'Oversight Board (MOB). Il Comitato per il controllo, investito dell'incarico di fornire un giudizio indipendente sul rispetto degli Standard della community, consente agli utenti, in caso di disaccordo, di avviare una procedura di ricorso per contestare le decisioni dell'azienda. Il MOB si esprime su temi già tutelati dalla Costituzione, come la cura della persona, l'identità performativa e il rispetto delle minoranze. Inoltre, ha un ruolo decisivo nel determinare "cosa rimuovere, cosa lasciare e perché", anche se non influisce direttamente sugli algoritmi e sulle strategie commerciali. L'analisi sul ruolo dell'Oversight Board rivela un conflitto tra legislazione pubblica e privata dovuto al rapporto ambivalente tra utente e Corporation, dove il primo agisce da cliente in base alla stipulazione del contratto per cui, registrandosi, accetta termini e condizioni del prodotto offerto dalla piattaforma (Santoriello, 2022b). Se considerato come una "Corte suprema" con funzioni pubbliche, si ottiene una "metafora costituzionale" (Cowls *et al.*, 2022, p. 2-3) in un contesto amministrativo agito da attori privati. Questa combinazione ha forti implicazioni per la democrazia e lo Stato di diritto, dove concetti come costituzionalismo, senso della legge, autorità e consenso pubblico sono

considerati soddisfatti senza un adeguato coinvolgimento delle istituzioni esistenti, come si approfondirà nel Capitolo 2.

Il quadrante in alto a destra delinea i casi in cui gruppi di interesse afferenti a imprese e organizzazioni pubbliche/private esercitano pressioni sui rappresentanti eletti attraverso attività di *Lobbying* per orientare l'agenda e la legislazione. Tra le più note vi sono le intese tra le aziende e gli Stati sulle sedi fisiche delle piattaforme. Un accordo stipulato tra Apple e l'Irlanda nel 1991 comprendeva la concessione alle filiali locali (Apple Sales International e Apple Operations Europe) di un trattamento fiscale privilegiato (lo 0,005% nel 2014). I diritti di sviluppo venivano trasferiti ad ASI a un prezzo simbolico e i contribuenti statunitensi venivano privati degli introiti derivanti dalle tecnologie incorporate nei prodotti che avevano a loro volta finanziato. La Commissione europea decise che l'aliquota massima sui profitti registrati in Irlanda dovesse attestarsi all'1%, a fronte del 12,5% pagato dalle società irlandesi. In seguito, ordinò di saldare gli emolumenti arretrati per "aiuto di Stato illegale", ma sia l'azienda che il governo irlandese rigettarono il provvedimento (Mazzucato, 2018, p. 4-5).

A The Dalles, in Oregon (USA), è situato il primo centro dati di proprietà di Google. I lobbisti hanno negoziato per mesi con i funzionari locali per ottenere un accordo che comprendesse esenzioni fiscali, garanzie di energia a basso costo e l'utilizzo della fibra ottica. Ciò vale a titolo di esempio per dimostrare come l'espansione dell'industria dell'IA sia stata sovvenzionata pubblicamente (Crawford, 2021, p. 215-217).

In tema di accesso ai dati acquisiti dalle aziende tecnologiche statunitensi, i tribunali europei hanno più volte affermato che i dati inviati agli USA possono essere spiati dalle agenzie di intelligence, come la *National Security Agency* (Meaker, 2022). Di fronte ai tentativi regolatori dell'Unione Europea, Meta ha inscenato un braccio di ferro, minacciando di rendere inaccessibili i suoi servizi. A luglio 2022, i negozianti statunitensi ed europei lavoravano per finalizzare un nuovo accordo di condivisione dei dati che avrebbe limitato le informazioni disponibili (Manancourt, 2022). Uno studio condotto sul *Digital Market Act* (Tarrant & Cowen, 2022) sostiene che, nonostante le ingenti spese per l'attività di *Lobbying* a Bruxelles, le *Big Tech* non abbiano ottenuto quanto sperato perché figurano come principali datori di lavoro e investitori soltanto in pochi Stati membri dell'UE che, al contrario, continuerà a introdurre vincoli normativi piuttosto che perseguire forme più tradizionali di diritto alla concorrenza.

Il rapporto del *Corporate Europe Observatory* (Schyns, 2023) testimonia l'influenza delle Corporation digitali sui negoziati che hanno sostenuto e affiancato la redazione dell'AI Act, testo che fornisce all'intelligenza artificiale un quadro comunitario di riferimento. Le aziende

statunitensi si sono presentate come “*AI first company*” (i cui prodotti sono trainati principalmente dai sistemi di IA) e, insieme alle associazioni di categoria, hanno contribuito alla definizione dei requisiti dei sistemi ad alto rischio tentando di escludere i “*General Purpose AP*” (GPAI¹³) per indebolire gli effetti della norma. Nello schema regolatorio, intento prioritario era garantire la valutazione dell’impatto per quei prodotti contenenti IA ad alto rischio e lasciare che rappresentanti dei comparti industriali partecipassero agli incontri con gli esperti. L’azione maggiormente finanziata riguarda l’equilibrio delle responsabilità tra utente, distributore e fornitore, per esimere questi ultimi dall’adeguamento e scaricarne il peso sulle piccole e medie imprese. Una delle argomentazioni presentate sostiene che la regolamentazione ridurrà il vantaggio competitivo a favore dei concorrenti cinesi (*ivi*, p. 20). Il testo finale presentato dal Consiglio contiene una definizione ristretta dei sistemi di IA, per distinguerli dai *software* più semplici, e la prof.ssa Joanna Bryson ha osservato la somiglianza con la proposta avanzata da Google. Altri studi hanno dimostrato che gli interessi commerciali prevalgono su quelli pubblici nell’influenzare le politiche digitali in UE, pur considerando le differenze esistenti tra le istituzioni: il Parlamento è influenzato dall’opinione pubblica, la Commissione ha una preferenza per le ONG e il Consiglio per le associazioni imprenditoriali (Keane, 2020, p. 65). I contorni risultano “sfumati” al punto che le attività di *Lobbying* possono essere esercitate senza che i gruppi di interesse entrino in conflitto tra loro.

Ciascuna Rivoluzione Industriale, se lasciata libera di fluttuare in un modello di organizzazione “innaturale” (Polanyi, 1974, p. 597), può avere conseguenze insostenibili per il sistema politico-economico internazionale. Per favorire equilibri a vantaggio dell’interesse pubblico, gli Stati possono intervenire regolando il mercato, così come viene presentato nel quadrante in alto a sinistra (Fig. 1). Per esempio, in Europa si sono susseguiti provvedimenti in materia di *Web Tax* (Staglianò, 2022, p. 106-107): nel 2015, il governo britannico ha varato la *Diverted Profits Tax*, che impone alle aziende con profitti superiori ai 10 milioni di sterline di autodenunciarsi all’Agenzia delle entrate; nella legge di Bilancio italiana del 2019 è stata inserita un’imposta minima del 3% che ha portato 233 milioni di euro all’erario. Dal punto di vista dei proponenti, le tasse devono essere pagate in virtù della creazione di valore, qualsiasi sia la sua origine (se tradizionale o digitale), e a beneficio della società.

¹³ Spesso si tratta di LLM, come nel caso di *DeepMind*, acquisita da Google-Alphabet nel 2014, oppure *Dall-E* e *ChatGPT* di OpenAI, in cui Microsoft figura come principale investitore. In un paper trasmesso da Google alla Commissione europea si evidenzia che il rispetto delle regole su *data governance*, applicazione e trasparenza sarebbe stato «*difficult or impossible to meet in practice*» (Schyns, 2023, p. 13).

L'impatto più significativo a livello internazionale è associato al *General Data Protection Regulation* (GDPR) 2016/679, che ha sancito alcune misure cruciali in materia di consenso, diritto all'informazione e diritto all'oblio negli ambienti digitali (conosciuti come standard di "terza generazione") e la cui interpretazione spetta alla Corte europea di Giustizia. Il testo ha influenzato altri provvedimenti legislativi al di fuori dell'UE, come il *Protection of Personal Information Act* (PoPIA) che regola la protezione dei dati e della privacy in Sud Africa, e significativi sviluppi si registrano anche in Asia, nei Caraibi e in America Latina (Greenleaf & Cottier, 2018), mentre il resto del continente africano fa riferimento all'*Economic Community of West African States (ECOWAS) Act on Data Protection* come strumento regionale vincolante e a principi internazionali sulla *governance* dei dati¹⁴ (Babalola, 2023).

I successivi Regolamenti UE *Digital Services Act Package* (DSA/DMA) e l'AI Act (AIA) puntano a creare «uno spazio digitale più sicuro in cui i diritti fondamentali degli utenti siano protetti e di stabilire condizioni di parità per le imprese» (Commissione europea, 2022), affiancati dall'Indice dell'Economia e della Società Digitale (DESI), una somma di indicatori sulle prestazioni per tracciare i progressi degli Stati membri. Non sono mancate iniziative locali sul contrasto alla disinformazione come la legge sulle *fake news* in Francia, il *Network Enforcement Act* (NetzDG) – conosciuto anche come Facebook-Gesetz¹⁵ - in Germania e il *Communication Platform Act* in Austria, che impongono una procedura *ad hoc* per i reclami su contenuti illeciti sui *Social Network*, da cui si sono mossi i primi passi verso l'entrata in vigore dell'*European Media Freedom Act* (EMFA) a livello comunitario.

Nell'ottobre 2022, l'Ufficio per le Politiche Scientifiche e Tecnologiche della Casa Bianca ha diffuso il "*Blueprint for an AI Bill of Rights*" (OSTP, 2022) con il fine di guidare la progettazione, l'uso e l'implementazione di sistemi automatizzati per proteggere il popolo statunitense nell'era dell'intelligenza artificiale. Questo documento non vincolante enuncia cinque principi etici rivolti a chiunque progetti o faccia uso di sistemi di IA, in cui si sancisce il diritto a:

- a) controllare l'utilizzo dei propri dati;
- b) non interagire con le decisioni automatizzate;
- c) vivere liberi da algoritmi inefficaci o insicuri attraverso alternative umane (avere accesso a una persona che possa comprendere e risolvere rapidamente i problemi);

¹⁴ Su 50 Paesi, soltanto 30 hanno un quadro normativo interamente dedicato alla protezione dei dati e 19 di questi hanno istituito delle autorità nazionali. Attualmente, soltanto in 10 (su un numero minimo di 16 per l'entrata in vigore) hanno ratificato la Convenzione di Malabo sulla sicurezza informatica (Rivista Africa, 2021).

¹⁵ Celebre la risposta di Facebook nel 2019: «*The constitutional state must not pass on its own shortcomings and responsibility to private companies. Preventing and combating hate speech and false reports is a public task from which the state must not escape*», citato da Lahoti (2019).

- d) conoscere i risultati dell'IA e il loro impatto;
- e) non essere discriminati dagli algoritmi mediante sistemi progettati in modo equo.

Stando al documento, l'amministrazione Biden riconosceva potenziale allo strumento tecnologico, al modo in cui potrebbe "ridefinire ogni aspetto della società" e "rendere la vita migliore per tutti", e a un progresso rispettoso dei diritti civili e dei valori democratici, che definiva "pietra miliare della sua amministrazione". La denominazione *Bill of Rights* era in aperto riferimento alla Carta dei diritti del 1791, in cui sono inseriti i primi dieci emendamenti della Costituzione, che introducono i diritti fondamentali¹⁶ dei cittadini. Dopo aver condotto audizioni con esperti, università, istituti di ricerca, aziende e start up di revisione dell'IA, l'OSTP ha ricevuto riscontri e avviato interlocuzioni con diverse parti interessate del settore privato (tra cui Google, Microsoft e Meta) e della società civile.

La *Regulation* è il tema centrale della matrice perché risponde alle questioni più urgenti, come quella di preservare la democrazia e l'integrità dei cittadini. I governi dispongono dell'autorità necessaria a identificare ordini di priorità per legiferare in maniera più efficace e determinare come organizzare la conoscenza. Per Wiener (1950, p. 105), è l'intervento legislativo che dà corpo alla dimensione etica:

«La legge può essere definita come il controllo etico applicato alla comunicazione e al linguaggio come forma di comunicazione, soprattutto quando questo aspetto normativo è sotto il controllo di un'autorità sufficientemente forte da assegnare alle sue decisioni una effettiva sanzione sociale. Si tratta del processo di regolazione degli "accoppiamenti" (*couplings*) che collegano il comportamento di diversi individui in modo tale che ciò che chiamiamo "giustizia" possa realizzarsi e le controversie possano essere evitate, o almeno giudicate».

Nell'industria culturale e dell'informazione, le Corporation digitali controllano sempre più spesso sia i mezzi di produzione che la connessione (Simon, 2022). Amazon, Google e Meta possono condizionare le scelte e le interazioni delle persone.

1.3.4 Ulteriori correlazioni nei concetti periferici

Dall'incontro tra interesse privato e funzione pubblica in settori cruciali, quali i trasporti, la sanità, l'istruzione e l'informazione, le aziende digitali e le loro piattaforme si espandono attraverso l'uso intensivo dei dati.

¹⁶ Si rammenta che i Diritti Umani sono prerequisito universale per diritto di nascita (indipendentemente dall'effettività, come da dottrina giusnaturalista), mentre i diritti fondamentali sono concepiti nella prospettiva di uno sviluppo del potenziale umano, intellettuale e fisico (inseriti principalmente nelle costituzioni nazionali degli Stati membri dell'UE).

La matrice di Corporation digitale (Fig. 1) crea un'opposizione tra il primo e l'ultimo quadrante, che rappresentano due conseguenze distinte: l'intervento normativo padroneggia la sfera pubblica, il *laissez-faire* risiede nella sfera privata. Uno dei quesiti a cui si proverà a rispondere nel prossimo capitolo è se sia o meno compito dello Stato regolamentare le imprese digitali.

Tra i concetti periferici emergono anche ulteriori correlazioni:

- a) "diritti - giustizia - pubblico" esprimono i principi del *welfare state*, «che garantisce a ogni cittadino il diritto a determinate soglie di reddito, alimentazione, salute, casa e istruzione» (Bobbio *et al.*, 2004);
- b) "pubblico - stato - rappresentanza" sono le peculiarità della forma di governo liberal-democratica, contraddistinta dalle elezioni;
- c) "*rappresentanza - libertà - interesse*" richiama i principi guida del Boston Tea Party "*No Taxation, Without Representation*" (dal punto di vista di Tim Wu "*an anti-monopoly protest*", 2018, p.20), agli albori della Rivoluzione americana;
- d) "interesse - mercato - diritti" trasmettono la prospettiva individuale, incline a cercare un equilibrio tra la Costituzione e il Contratto (Calise & Lowi, 2010).

La matrice analizza il contesto che vige nel capitalismo occidentale, motivo per cui non sono stati inclusi e/o approfonditi i riferimenti alle *Big Tech* cinesi BATX (Baidu, Alibaba, Tencent e Xiaomi).

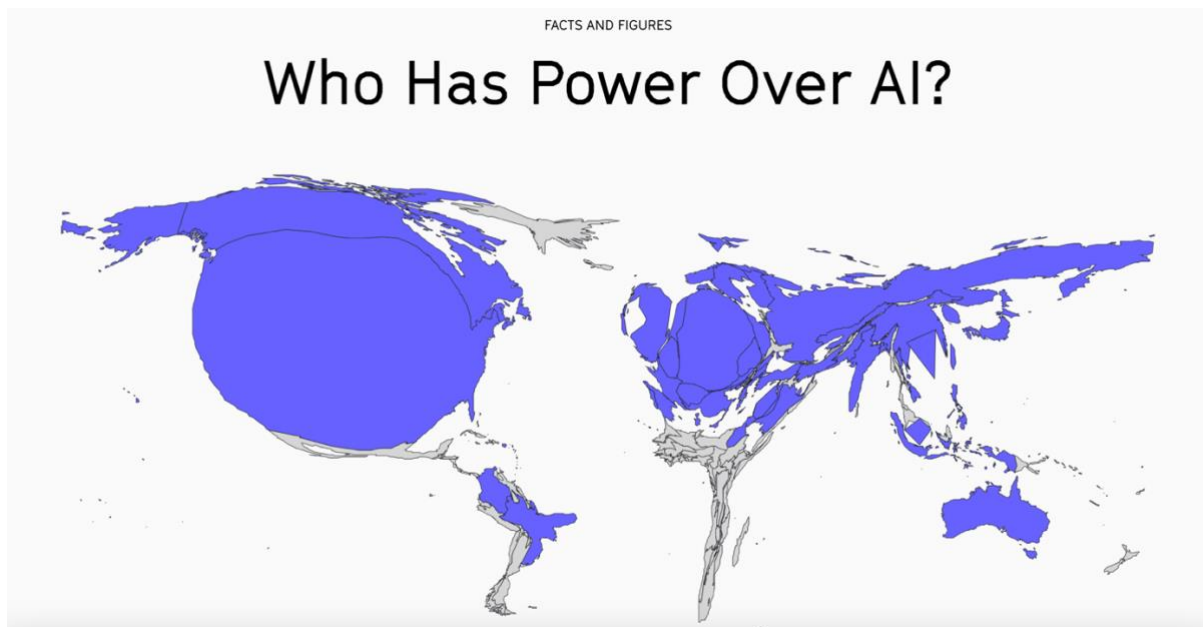


Figura 2. Chi ha potere sull'IA?
Fonte: Internet Health Report 2022

1.4 Biopolitiche del digitale nel *Garbage In, Garbage Out* (GIGO)

Sottorappresentato, quando non del tutto ignorato, è il ruolo dell'essere umano oltre la macchina. Le figure professionali che si occupano del funzionamento dei sistemi di IA sono *tech workers* o *IT workers*, ovvero quei lavoratori tecnici e creativi che intervengono sulle infrastrutture informatiche (Freeman & Aspray, 1999), anche detti “*code writers*” e “*techies*” (Lessig, 1999). Ciascuna fase di sviluppo prevede diversi livelli di responsabilità e l'intervento di informatici, programmatori e *data scientist*. Nelle discipline cosiddette STEM (scienze, tecnologia, ingegneria, matematica), di cui fa parte il settore *Information and Communication Technology* (ICT), la componente maschile è prevalente (O'Neil, 2017; Broussard, 2018; Crawford, 2021) a fronte di appena un terzo di donne occupate. Il fenomeno viene definito “*leaky pipeline*”¹⁷, un tubo che perde.

Nel volume “*Atlas of AI*” (2021), Kate Crawford spiega che nel *machine learning* la relazione tra chi trasmette (il lavoratore *tech*) e chi apprende ha una componente emotiva: le tassonomie che categorizzano le persone sono «binomi di genere forzati, raggruppamenti razziali semplicistici e offensivi, analisi stereotipate del carattere, del merito e dello stato emotivo». Come già anticipato nel precedente articolo “*Artificial intelligence's white guy problem*” (2016), le etichette cariche di valori trasmettono una visione del mondo e la legittimano su vasta scala. I dati raccolti – espressioni di un costrutto sociale - vengono processati da due categorie di algoritmi, *learners* e *classifiers*: i primi apprendono, per poi spiegare ai secondi come analizzare nel miglior modo la relazione tra *input* e *target* nel processo di predizione. Quando si parla di “errori tipici” (Palazzani, 2020, p. 58), si fa riferimento al verificarsi di una raccolta dei dati incompleta, una categorizzazione dei *file* fallace, una codificazione non adeguata, la registrazione di un utente non idonea, finanche l'intera analisi erronea. Gran parte del lavoro sui sistemi viene svolto senza processi etici di revisione perché considerati come formule prive di implicazioni per gli esseri umani.

Nei sistemi statistici, il *bias* si riferisce ai processi di generalizzazione (*Learning Bias*), classificazione e varianza. Nel *machine learning*, i *bias* sono cognitivi e legati al modo in cui i giudizi deviano sistematicamente dai risultati attesi su base probabilistica dove il problema della quantificazione genera a sua volta un problema di rappresentazione. Il meccanismo “*Garbage in, Garbage out*” (GIGO) ricorda il lavoro acritico svolto dagli algoritmi: una omissione nell'*input* è causa di una omissione nell'*output* (Santoriello, 2022a). Nascondere, come atto di

¹⁷ Si vedano i dati presenti nel rapporto Kuptsch, C., & Charest, É. (2021). *The future of diversity*. ILO, 13.

contenimento, alcuni soggetti e punti di vista non rende le interpretazioni più neutrali, anzi oscura l'orientamento che le determina (Numerico, 2021).

Crawford (2021) imputa questo fallimento alle asimmetrie di potere che mirano a identificare le persone per un imperativo di controllo e sorveglianza. La comprensione dei rischi è, ancora una volta, appannaggio dei gruppi più ricchi e protetti dei Paesi industrializzati (Beck, 2000, p. 69). Luciano Floridi (2022) afferma che la tecnologia può essere compresa in molti modi, ma l'approccio migliore è concepirla nei termini di un buon *design* e di una *governance* etica. Dal suo punto di vista, dobbiamo renderci conto che l'agire politico (democrazia diretta) e l'agire artificiale (IA) sono oggetto di un'interpretazione fallace: il ricongiungimento tra potere politico e delega, tra sovranità e *governance*, a cui si propende per mezzo del digitale è un rischio per la democrazia rappresentativa e non un'opportunità (*ivi*, p. 32-34).

La natura sociotecnica delle macchine è insita nel rapporto che instaurano con gli esseri umani: dalla scrittura al rilascio, il processo di creazione degli algoritmi è influenzato dalla cultura e dal pregiudizio delimitati nel tempo e nello spazio del presente. Costruire un modello matematico per il *machine learning* significa formalizzare questa esperienza in un codice. Concetti come l'equità non sono misurabili e non spetta ai *tech workers* codificarli. Assistiamo a una ingente produzione di ingiustizia che ha portato Cathy O'Neil a rinominare gli algoritmi "armi di distruzione matematica" (2017, p. 140-141), il cui problema principale è l'impostazione dell'efficacia quale obiettivo, anche a discapito del gruppo. Questo "effetto dell'IA" è la dimostrazione di una scissione della capacità di agire tra esecuzione (scopo) e cognizione (fine). L'habitat sociale si sta trasformando in uno spazio (le *Smart City*) in cui i *robot* possono operare con successo, con maggiore avvolgimento del 5G e di *Internet of Things* (IoT), per ospitare l'IA generativa (Floridi, 2022, p. 56-57).

Louise Amoore (2020, p. 135) spiega che la "verità di fondo" (*Ground Truth*) degli algoritmi non è incentrata sulla consapevolezza di un'entità, ma sull'ammontare di informazioni che ne dimostrano l'esistenza. L'essere umano diventa, in altre parole, l'*homo technologicus* (Longo, 2017), le cui parti, biologica e tecnologica, convivono rendendolo una creatura in continuo divenire. La corporeità viene replicata in diverse forme che spaziano dalla raccolta di dati biometrici tramite foto in cui siano chiare le espressioni per il riconoscimento facciale (il progetto *ImageNet* le estraeva direttamente da Internet) alla riproduzione vocale presente nei dispositivi di assistenza (come Google Home, Alexa e Siri), senza che sia richiesto alcun consenso per l'inserimento nei *dataset* (Crawford, 2021, p. 106-109).

L'invasione della dimensione digitale nella sfera del privato accresce l'esposizione dell'individuo a una messa in discussione dell'identità: il codice informatico è un regolatore di

comunità (Lessig, 1999) perché l'autenticazione è un processo mediante il quale l'identità di ciascuno viene riconosciuta¹⁸. In tale scenario, le forme di tutela degli equilibri naturali, alterati per via artificiale, rivolgono la propria attenzione ai diritti di “quarta generazione” (Amato S. , 2021, p. 12-14). In questo senso si esprime la biopolitica del digitale, un processo che mira a creare forme di soggettività osservabili¹⁹ nell'auto-realizzazione. Il soggetto della governamentalità neoliberale è l'oggetto da monitorare e organizzare alla stregua delle pratiche e dei saperi. L'IA sostituisce la deduzione della logica matematica con l'inferenza e la correlazione statistica²⁰, registrando e monitorando il “chi” e il “cosa” di tutti e ciascuno. La privacy – la vita personale, privata, dell'individuo - è meritevole di protezione in quanto fondamento della dignità umana; nell'Unione europea questo principio poggia sul legame tra “persona” e “dato”. Scrive Laura Bazzicalupo (2015):

«I gruppi e le popolazioni (piuttosto che il popolo e l'individuo, che sono i classici soggetti della liberaldemocrazia tradizionale) sono individuati e tra loro differenziati in base all'esposizione al rischio che li caratterizza o in base alle potenzialità produttive che li accomunano. In entrambi i casi (rischio e potenzialità) sono messi a fuoco per essere gestiti».

In quali situazioni²¹ le persone diventano consapevoli delle macchine? Se la complessità sociale si riducesse a frequenze ritenute statisticamente rilevanti, la maggiore/minore probabilità che si verifichi un fenomeno (c.d. *group identity*), si correrebbe il rischio di emarginare categorie di individui che non aderiscono allo standard prestabilito compiendo discriminazioni e ingiustizie. La standardizzazione di principi etici in fase di progettazione (*Ethically Aligned Design*) contribuirebbe, come da auspici costituzionali, alla promozione dello sviluppo autentico dell'identità personale, soprattutto nel campo dell'educazione.

Alle istituzioni pubbliche si richiede di agire nel perseguimento di questo interesse, che passa attraverso la convalida, l'*audit* e la revisione, cessando di utilizzare applicazioni programmate mediante *black box* (Broussard, 2018, p. 352). La rivoluzione tecnologica interessa enormemente gli equilibri internazionali, incluso il funzionamento della democrazia, perché

¹⁸ Con il proliferare di casi di *fake identity*, sempre più spesso l'accesso a un servizio online è subordinato a un'operazione di associazione tra concetto e foto per convincere un sistema di *reCaptcha* dell'esistenza umana oltre lo schermo, allenando gratuitamente un algoritmo di riconoscimento delle immagini.

¹⁹ Proprio come nell'aneddoto della “scatola di sardine” di Lacan, il “modo di vedere invisuale” opera attraverso meccanismi invisibili. La visione non è più qualcosa di neutrale e naturale, ma una costruzione che coinvolge tecnologia, algoritmi e controllo sociale.

²⁰ Crawford (2021, p.97) dice: «Se, nel nostro esempio, tutte le immagini di addestramento di mele sono rosse e nessuna è verde, allora un sistema di ML potrebbe dedurre che “tutte le mele sono rosse”. Si tratta di un'inferenza induttiva, un'ipotesi aperta basata sui dati disponibili, piuttosto che di un'inferenza deduttiva, che segue logicamente da una premessa. Dato il modo in cui questo sistema è stato addestrato, una mela verde non verrebbe riconosciuta come una mela. I *dataset* di addestramento, quindi, sono al centro del modo in cui la maggior parte dei sistemi di ML produce inferenze».

²¹ Con lo scoppio della crisi del 2008, l'opinione pubblica “prese coscienza” delle persone che stavano dall'altra parte degli algoritmi (O'Neil, 2017, p. 65).

privati cittadini possono compiere azioni in passato riservate soltanto agli Stati (Staglianò, 2022) e sono responsabili dell'estrazione di risorse non rinnovabili²² con impatti negativi sull'ambiente e sulle comunità locali (Crawford, 2021, p. 31).

Il rifiuto di Rosa Parks di cedere il posto a un bianco in autobus nel 1955 ha fatto scuola, dando vita ad azioni di boicottaggio che hanno, infine, inciso sulla legislazione. L'opposizione alle norme, quando percepite come ingiuste, mette in moto un processo politico di ricodificazione. La stessa "etica del digitale" è a rischio (Floridi, 2019): i gruppi di interesse possono esercitare pressioni per orientare le scelte del legislatore (*Lobbying*), selezionare a piacimento o adattare i principi più comodi (*Shopping*) e promuovere prodotti poco trasparenti o informazioni fuorvianti sui valori (*Bluewashing*); situazioni classificabili come distruttive potrebbero verificarsi qualora si decida di esportare attività di ricerca, prodotti e servizi in altri luoghi, in modi che sarebbero considerati inaccettabili nei contesti di origine (*Dumping*) e di valutare come vantaggiosa la possibilità che ciò avvenga quanto più basso è il rendimento del lavoro (*Shirking*), che interessa il campo della moderazione dei contenuti online (Santoriello & Franchi, 2023). La tesi si occupa delle Corporation digitali dal punto di vista della regolamentazione pubblica e privata: si sosterrà che, mutato il contesto, emergono nuovi spazi per la contrattazione collettiva e per l'intervento sulle vite e le libertà individuali. Campo d'indagine per esaminare questi cambiamenti è la definizione dell'identità di cittadine e cittadini con e attraverso la Rete. Se il codice è la legge (Lessig, 1999), bisogna controllare che venga scritto in conformità con il diritto.

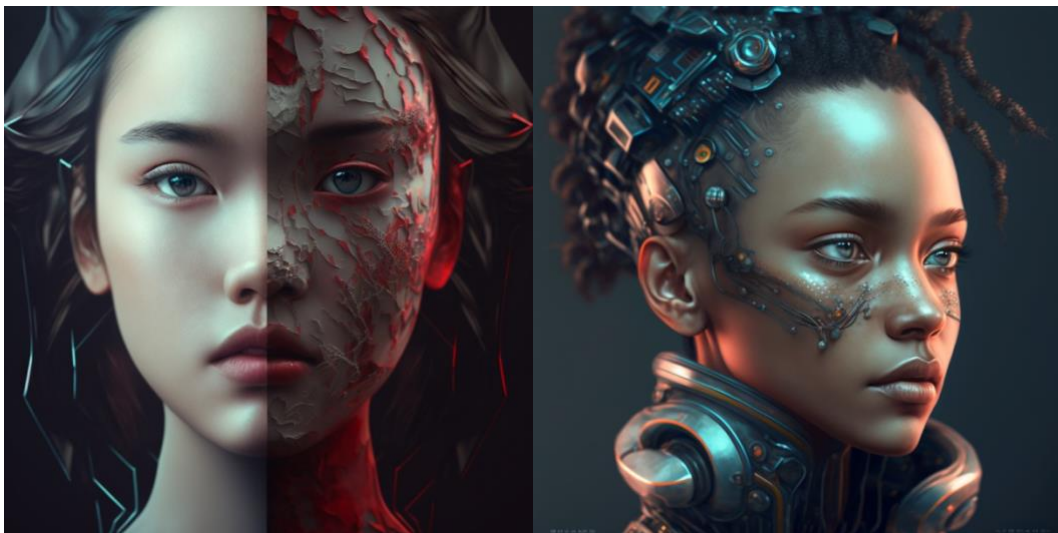


Figura 3 e 4. *AI Bias in Midjourney (2023)*.

Fonte: Midjourney è un laboratorio e un LLM che crea immagini da descrizioni testuali. Inserendo la query "AI bias", nel 2023 l'algoritmo ha riprodotto due immagini di donne con connotazioni etniche non occidentali e un'alterazione tecnologica su parte del volto.

²² In "*A Geology of Media*", Jussi Parikka suggerisce di pensare i media come estensioni della Terra.

Capitolo 2

I diritti fondamentali alla prova della Moderazione Automatizzata

*«It can follow analysis; but it has no power of
anticipating any analytical relations or truths»*

Ada A. Lovelace

2.1 Le fonti del diritto e il concetto di sovranità digitale

Oltre due miliardi di persone hanno scelto di iscriversi sulle più popolari piattaforme di social media. «Se Facebook fosse una nazione, sarebbe la più popolata al mondo» recita l'incipit del documentario *Facebookistan* (2015), precisamente il 70% della popolazione residente nell'Unione europea nel 2022 (*Internet World Stats*). Far parte di una rete giova quanto più elevato è il numero dei suoi utenti, creando le condizioni affinché utilità e valore siano foraggiate da esternalità positive come interazioni, informazioni, opportunità di esprimersi liberamente e condividere esperienze. Al crescere della comunità, i numeri e le differenti tipologie di contenuti pubblicati comportano l'introduzione di norme e strategie di gestione, decise dai vertici amministrativi, "a porte chiuse" (Gillespie, 2018). La comparsa di nuove "istituzioni" e l'implementazione di sistemi privi di negoziazione politica con le parti sociali è diretta conseguenza della scissione democratica contemporanea. Diventa essenziale comprendere come si manifesta il governo su porzioni dello spazio politico erose dal digitale. La pietra miliare della scienza politica è il concetto di "sovranità", ovvero il monopolio riconosciuto della forza legittima. Autorità e potere caratterizzano la dimensione ambivalente entro cui si sviluppa (Polin, R. in Calise *et al.*, 2021): interna, la cui univocità è definita da un territorio e verso la popolazione che vi risiede, fin dalla formula latina *rex superiorem non recognoscens in regno suo est imperator*; esterna, determinata dall'indipendenza dagli altri Stati e dal mutuo riconoscimento, originati nel 1648 con la Pace di Vestfalia. La sua affermazione non ha alternative, un'entità statale può essere o non essere sovrana, *tertium non datur* (*ibidem*): difendere la sicurezza e gli interessi nazionali spetta allo Stato.

Nella dimensione *tech*, il concetto viene ripreso e replicato: la sovranità digitale si riferisce alla capacità di una nazione di esercitare controllo sulla propria infrastruttura, sui flussi di dati e sulle attività online entro i propri confini, mantenendone l'autonomia, in un momento storico in cui il settore privato partecipa alle relazioni internazionali e diplomatiche (Schaaake, 2024).

Gli approcci possono essere molto diversi e rispecchiare le linee guida di ciascun sistema politico, da un processo decisionale decentralizzato e condiviso tra vari attori a un controllo più centralizzato da parte dello Stato, con annessi regolamenti e *policy* per prevenire accessi non autorizzati o usi impropri da parte di entità estere. Ciò può comportare l'attuazione di misure di localizzazione dei dati o l'emanazione di norme più severe sui flussi transfrontalieri, motivate dal desiderio di ridurre la dipendenza dalla tecnologia straniera e di attenuare le vulnerabilità a essa associate. Il controllo su importanti settori industriali è un fondamento dello stato di diritto (Schaake, 2024, p. 240), ma si tratta anche di una "reazione a uno *status quo*" (Bertola, V. in Abba *et al.*, 2022, p. 53-54) che ha preso forma nel corso di diversi decenni per smentire chi sosteneva l'effetto implicito di Internet: rendere superflui i governi. Il riferimento più opportuno, e meno deterministico, andrebbe rivolto alla più ampia "interdipendenza complessa" (Polin, R. in Calise *et al.*, 2021) derivata dalla globalizzazione, che ha reso necessario l'intervento preventivo della comunità internazionale per definire quadri etici, regolamenti e standard, con il doppio scopo di allineare le tecnologie a valori nazionali e norme culturali, come la trasparenza e la tracciabilità (*Explainability*) degli algoritmi, e controllare l'impatto dell'IA sul mercato del lavoro e sulle disuguaglianze sociali.

Schaake sostiene che “il golpe tecnologico” (2024, p. 243) sta riscrivendo il contratto sociale tra democrazie e cittadini, privatizzando le responsabilità, un tempo monopolio dello Stato, e senza nessi elettivi. Già Berg e Hofmann (2021) avevano osservato quattro ambiti in cui assistiamo alla trasformazione: il ruolo del governo, la sfera pubblica, la dinamica tra partecipazione e rappresentazione, il problema del controllo e dei diritti. Con l'avvento della sfera pubblica generativa e il manifestarsi di forme di *decision-making* fuori da istituzioni indebolite, sono i confini costituzionali e il bilanciamento dei poteri a essere messi alla prova, e non necessariamente in senso emancipatorio.

La presenza di corpi intermedi, ovvero l'equilibrio tra pluralismo sociale e frammentazione del potere, è l'elemento critico della sovranità. Nei forum dei paesi occidentali cresce il consenso per una valutazione critica delle piattaforme dominanti (Cammaerts & Mansell, 2020) sul rapporto tra valori economici e pubblici, dando priorità a temi che riguardano la privacy, la libertà di espressione e la sicurezza. Le *Digital Public Policies*, in cui rientrano iniziative legislative e codici di condotta, ne evidenziano la priorità - pur non prevedendo un intervento sul mercato che metta in discussione i fondamenti dell'attuale datificazione commerciale – e azionano il processo sanzionatorio in caso di mancato rispetto delle norme (*ibidem*).

Negli anni 2000, il modello economicista dominante frenava l'intervento pubblico dalla regolazione delle infrastrutture²³ virando verso l'istituzionalizzazione di alcuni principi chiave, individuati da Santaniello (2021): *Private Sector Leadership*, *Business Self-regulation* e *Pro-market Regulation*. Dopo vent'anni il vento è cambiato e si parla ora di *Platform Regulation* (ivi, p. 589), che incrocia le due dimensioni politiche della regolazione come la fonte del processo (modelli di *governance*) e l'oggetto (attività di interazione o funzionamento della piattaforma). La tendenza più diffusa riguarda il rapporto tra piattaforma e utente, definito dai proprietari in maniera unilaterale, attraverso *Design* e *User Agreements*, base normativa del *Platform Order*. Nel contesto europeo, dove per molto tempo sono state implementate direttive anziché regolamenti, è in atto un "turn to regulation" (ivi, p. 592) per circoscrivere i sistemi di funzionamento delle piattaforme, garantire la compatibilità e l'interoperabilità attraverso i confini. Il progetto Gaia-X (*Data Storage Cloud*) del 2019 fu promosso da Francia e Germania al fine di assicurare una gestione dei dati sicura e indipendente, ma anche di fronteggiare il *Cloud Act* degli USA. Ciononostante, l'assenza di servizi alternativi in UE rese necessario un approccio inclusivo di gruppi come Amazon, Google e Microsoft (Schaake, 2024, p. 171).

In materia di sistemi di IA, la creazione di un quadro legale globale sembra il percorso privilegiato per favorire gli interessi degli Stati membri e consentire la cooperazione che, secondo Roberts *et al.* (2024), si trova in affanno a causa delle sfide geopolitiche derivanti da: l'anarchia internazionale, ovvero di una guida comune nella politica mondiale; il vantaggio competitivo percepito dai singoli Stati, sia per ragioni economiche che di sicurezza, che pone un tema di fiducia; la frammentazione istituzionale, che complica gli esiti del multilateralismo (si veda, per esempio, la "coalizione ideologica" nel G7). La Cina, per esempio, ha promosso politiche industriali *Made in China 2025* e gli standard tecnologici della *Belt and Road Initiative*, ma si è astenuta dalla *Global Partnership on Artificial Intelligence* (GPAI) e le sue aziende non sono coinvolte nella PAI; gli USA controllano le esportazioni di semiconduttori, mentre l'UE sta rafforzando la capacità nazionale in materia di *hardware* e *software* per ridurre la dipendenza dall'estero, esercitando influenza internazionale ("Effetto Bruxelles") soprattutto per l'accesso al suo mercato (*ibidem*).

Le *Big Five* dell'UE comprendono il *Digital Services Act* (DSA), il *Digital Markets Act* (DMA), il *Data Governance Act* (DGA), il *Data Act* (DA) e l'*Artificial Intelligence Act* (AIA) e sono state concepite per dare vita a un ambiente normativo "meta-livello" di riflesso alle Corporation digitali di Stati Uniti e Cina (Nieminen *et al.*, 2023), facendo direttamente seguito al cambio di

²³ Lo ricorda anche Schaake (2024, p. 171), in quel periodo le democrazie in mercato libero temevano che gli appelli alla sovranità esprimessero un richiamo al nazionalismo e al protezionismo.

paradigma avvenuto con il *General Data Protection Regulation* (GDPR) e dell'entrata in vigore nel 2018 di un sistema sanzionatorio a difesa dei cittadini europei, incluso il diritto a essere informati, alla portabilità e all'oblio. Dal 2022, il *Digital Services Act Package* (DSA e DMA) rimodula le responsabilità della *eCommerce Directive* del 2001, esentando soltanto i fornitori di servizi di telecomunicazione e i *Server Hosting*. *De facto*, le piattaforme digitali vengono inquadrare come ecosistemi soggetti a sovranità digitale. L'art. 33 del DSA (Tab. 1) introduce le definizioni di Piattaforme online di dimensioni molto grandi (*Very Large Online Platform* – VLOP) e Motori di ricerca online di dimensioni molto grandi (*Very Large Online Search Engine* – VLOSE), che ospitano un numero medio di utenti europei (destinatari attivi del servizio) pari o superiore al 10% del totale dei cittadini UE (45 milioni di persone), e affina gli obblighi di trasparenza e *accountability*, tra cui rientrano l'accesso ai dati da parte di ricercatori e *policymaker*, le informazioni sulle inserzioni pubblicitarie e i meccanismi partecipativi sulla moderazione dei contenuti. L'approvazione del *Media Freedom Act* (EMFA) è un ulteriore passo in questa direzione, istituendo norme comuni per i servizi di media nel mercato interno.

Tabella 2. *Elenco di VLOP e VLOSE nel 2024*

Very Large Online Platforms	Very Large Online Search Engines
AliExpress (Alibaba) Amazon Apple Booking.com Facebook, Instagram (Meta) Google Play, Maps, Shopping (Alphabet) LinkedIn (Microsoft) Pinterest PornHub Snapchat Shein Stripchat Temu TikTok (ByteDance) Wikipedia X (già Twitter) XNXX XVideos YouTube (Alphabet) Zalando	Bing (Microsoft) Google (Alphabet)

Fonte: Elaborazione a cura dell'autrice aggiornata al 17/12/2024

Il DMA dettaglia gli obblighi per i *gatekeeper* e delega la supervisione alla Commissione europea per evitare i colli di bottiglia tra Corporation e consumatori. Particolare attenzione viene dedicata agli effetti di rete prodotti dalle piattaforme, esplicitando l'obbligo di interoperabilità e il divieto di manipolazione del mercato rivolto ai proprietari.

La sovranità digitale europea conduce verso un nuovo approccio la *governance* di Internet, una sorta di "terza via" (Santaniello *et al.*, 2022, p. 63), i cui pilastri sono i diritti fondamentali, la democrazia e la giustizia, alternativa al modello neoliberista e alle aspirazioni protezioniste dei regimi autoritari. Proprio il “grande balzo” della Cina nella corsa all’IA dimostra che il sistema liberale non è la fonte del progresso tecnologico, il quale non si arresta neanche di fronte alla regolamentazione (Bradford, 2023). È la scalabilità il segreto della crescita e della competitività, ma presuppone anche coesione, sia economica che culturale. Il rallentamento dell’UE non è dovuto alle regole, ma alla frammentazione del mercato interno e alla diversità culturale e linguistica di cui si compone (*ivi*, p. 390).



Figura 5. Approvazione dell'AI Act (2023). MEP Relatori Benifei e Tudorache.

L'accordo raggiunto sul Regolamento europeo che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act) ha determinato l'entrata in vigore di parametri vincolanti di qualità, trasparenza e controllo umano sul prodotto e lungo tutta la filiera (dall'immissione sul mercato alla messa in servizio, per finalità generali), a prescindere dalla collocazione. Il suo approccio è basato sul rischio: divieti, obblighi e requisiti sono calibrati in base al prodotto e a un utilizzo specifico. La legge stabilisce regole per il monitoraggio, la vigilanza del mercato e misure a sostegno dell'innovazione rivolte a PMI e start up. Si distinguono due figure, che possono essere entrambe persona fisica o giuridica, autorità pubblica, agenzia o altro organismo: il fornitore, chi sviluppa il sistema o lo commissiona per immetterlo sul mercato o in servizio con il proprio nome o marchio, a titolo oneroso o gratuito, indipendentemente da dove sia stabilito; il *deployer*, chi utilizza il sistema sotto la sua autorità, a meno che l'attività non sia classificabile come personale e non professionale, con sede o situato in UE. Fornitori e *deployer* di un paese terzo sono ugualmente coinvolti se l'*output* prodotto è utilizzato in UE,

così come importatori e distributori, fabbricanti, rappresentanti autorizzati e persone interessate. Non si applica, invece, a sistemi di IA in settori che esulano dal diritto UE, per scopi militari, di difesa o sicurezza nazionale, per ricerca e sviluppo scientifico, per *beta testing*, in caso di licenze *free* e *open source* (salvo che non rientrino nei sistemi ad alto rischio). La sovranità europea assume le sembianze di un puzzle legislativo non esente da critiche (Schaake, 2024).

Secondo Cammaerts e Mansell (2020, p. 145), una deliberazione democratica ha bisogno di: un dibattito politico e normativo in una istituzione indipendente, un pluralismo radicale che dia maggior peso agli interessi pubblici, l'articolazione di misure e pratiche (nonché organismi rappresentativi) per realizzarli, denaturalizzando il senso comune, e monitorare gli impatti delle piattaforme sui cittadini. Le infrastrutture digitali, però, hanno condotto Stato e Mercato verso interessi convergenti in tema di fornitura: questa tendenza a esternalizzare l'applicazione della legge al settore privato, che concepisce i proprietari delle piattaforme come "governatori primari" della comunicazione online, ha conseguenze poco chiare in termini di qualità del controllo pubblico e responsabilità democratica, non scongiurando il rischio che i privati restino o diventino soggetti dominanti; anzi, pare che promuovano appuntamenti come il *Brussels Forum* o l'*European Internet Forum* per migliorare la propria reputazione²⁴ e ridurre il rischio di essere colpiti dall'antitrust (Schaake, 2024, p. 174). Alcuni autori, come Berg e Hofman (2021), preferirebbero un maggior coinvolgimento di questi attori nel "momento costituzionale" perché temono che il rimando alla sovranità giustifichi il rafforzamento dello Stato piuttosto che promuovere la democrazia²⁵.

La gerarchia delle fonti delle *Digital Private Policy* si articola, dunque, in una commistione di norme nazionali e internazionali, giacché la struttura delle Corporation digitali è distribuita in sedi fisiche, localizzate in USA, Cina e Irlanda, e presenza immateriale online, non per questo meno tangibile tra domini, applicazioni e codici algoritmici. Nessuna GAFAM è nata in Europa e nessuno dei proprietari possiede la cittadinanza europea: il *Platform Gap*, unitamente al contesto geopolitico che vede Stati Uniti e Cina contendersi lo scacchiere globale, rende l'Unione europea un legislatore svincolato da interessi di parte, in cui risiedono consumatori benestanti (Bradford, 2023; Schaake, 2024). Strategie e investimenti vengono continuamente rivalutati per stare al passo con i progressi tecnologici e le minacce in evoluzione, ma l'UE è sempre più dipendente dai servizi delle società digitali (Nieminen *et al.*, 2023).

²⁴ Durante l'approvazione del GDPR, Facebook provò la strada del *pinkwashing*, coinvolgendo Sheryl Sandberg, Chief Operating Officer e prima donna nel Consiglio d'Amministrazione (tra le più influenti del mondo). Zuboff (2019) la definì la "Typhoid Mary del capitalismo di sorveglianza", ovvero colei che ha infettato il mondo.

²⁵ L'India, per esempio, ha il più alto numero di interruzioni dei servizi di Rete nel mondo, accompagnati da casi di censura da parte del governo di Modi (Schaake, 2024, p. 174).

Tra le ideologie dominanti affermatesi nel settore ci sono il datismo (*Dataism*) e il soluzionismo tecnologico (*Technological Solutionism*): il primo assume che i dati restituiscano una rappresentazione neutrale della realtà e la principale, nonché empirica, forma di conoscenza; il secondo, riposiziona la tecnologia come base della questione politica²⁶, articolandola in capacità e limiti di coloro che la utilizzano (Haggart & Tusikov, 2023, p. 250).

La *Datification* interseca la giustizia sociale nel complesso paradigma del capitalismo contemporaneo e delle asimmetrie di potere esistenti. L'affermarsi della *Data Justice* (Dencik *et al.*, 2019), sotto forma di critica al sistema culturale costituito da infrastrutture e gestioni autonome, riconosce la responsabilità della tecnologia nella produzione di rappresentazioni distorte o manchevoli, e la facoltà, ovvero il diritto, delle persone di scegliere se prendere parte o meno all'economia dei dati.

L'assenza di una leadership globale spinge a scavare nel passato, per giungere all'esperienza dei negoziati di Montego Bay. L'esercizio di sovranità contenuto nella Convenzione delle Nazioni Unite sul diritto del mare (UNCLOS) del 1982 ha stabilito un quadro globale per la gestione degli oceani e dei mari (Treves, 2008), in sostituzione del vecchio concetto del XVII secolo, secondo il quale i diritti nazionali erano limitati a specifici lembi di mare e lo spazio marino antistante rientrava nelle "acque internazionali", di proprietà di nessuno Stato e liberamente accessibili. Il trattato fu articolato in un momento storico di accesi cambiamenti, in cui la decolonizzazione stava consentendo a nuovi attori statali indipendenti di esprimere il proprio punto di vista modificando i trattati. Se considerare le Corporation equivalenti agli Stati sarebbe un errore, auspicabile è raccogliere lo spirito della Convenzione. Le aziende sembrano muoversi in concorrenza e in un rapporto di mutua dipendenza dallo Stato, ma a quest'ultimo spetta preminenza perché queste "isole" sono situate sotto la sua giurisdizione; in altre parole, è la Costituzione a rendere validi i contratti. Attenuare il potere delle multinazionali con lo spirito di cooperazione delle Nazioni Unite sarebbe una ulteriore affermazione di sovranità degli Stati a tutela dei Diritti Umani e della libertà di espressione.

Sono state avanzate diverse proposte per colmare il deficit di *governance* globale dell'IA (Roberts *et al.*, 2024): a) sviluppare nuove istituzioni o b) rafforzare il coordinamento delle istituzioni esistenti, entrambe nel rispetto di principi di legittimità politica quali il vantaggio reciproco e il processo decisionale democratico.

²⁶ Lo è senz'altro l'identificazione di un "interesse superiore": nel caso delle *smart city*, da un lato ci sono i *Sidewalk Labs* della città di Toronto e dall'altro l'approccio dal basso alla progettazione dei sistemi adottato a Barcellona "mettere i dati al servizio delle persone e non le persone al servizio della tecnologia"; la variabile principale è il contesto socio-economico in cui vengono implementati (Haggart & Tusikov, 2023, p. 261).

Con riferimento alla prima proposta, l'opzione di tipo centralizzato immagina la creazione di un organismo come l'AIEA in ambito nucleare per la definizione degli standard, ma non risulta fattibile per la natura intrinseca dell'intelligenza artificiale che, oltre a differenze materiali, presenta impatti trasversali sia a monte che a valle del processo. Altre istanze di tipo semi-centralizzato immaginano un "IPCC per l'IA" per sostenere il consenso scientifico e un "CERN per l'IA" per concentrarsi sulla sicurezza, anche se i modelli a cui si ispirano sono emersi per effetto di uno scenario storico e politico di gran lunga differente e corrono il rischio di vedere la *governance* plasmata delle grandi aziende, indebolendo il mandato democratico (Roberts *et al.*, 2024). La seconda proposta è di tipo complesso e chiama in causa uno sforzo di cooperazione internazionale in grado di facilitare l'inclusione di nuovi *stakeholders* (tra cui le *Big Tech*) nel processo che, nel consolidare gli obiettivi, allineerebbe gli attori coinvolti alle priorità internazionali, alla condivisione delle informazioni, allo sviluppo di partenariati istituzionali e contribuirebbe a creare meccanismi di risoluzione dei conflitti (*ibidem*). Il modello a cui si ispira è la Conferenza sul Clima, che ha incoraggiato i governi locali e le aziende private ad attuare politiche in base a specifiche responsabilità, migliorando i risultati nel tempo grazie a un approccio decentralizzato e incrementale. La sfida sarà quella di trovare un equilibrio tra l'affermazione del controllo e la conservazione di un ecosistema digitale sicuro, prospero ed etico. I cittadini europei chiedono al legislatore di intervenire su più fronti.

Uno studio condotto dal *Center for the Governance of Change* (De Tena & De Viedma, 2023) fotografa uno scenario in cui emerge che i cittadini confidano nell'UE (44%) per regolare l'IA e la leadership tecnologica (88%). Una maggioranza significativa di europei è favorevole alla regolamentazione globale di Internet, in cui venga mantenuto il carattere aperto della Rete pur con regole universali (posizione sostenuta dal 74% degli italiani), dando priorità alla protezione dai contenuti dannosi (79%) rispetto alla libertà di pubblicare qualsiasi cosa online (21%). La comparsa di *Generative AI* come BART e Midjourney preoccupa per l'incapacità di discernere tra contenuti ingannevoli e autentici (soltanto il 27% si dichiara in grado di distinguerli). Alla domanda "Come pensa che l'uso dei dati influenzerà la distribuzione della ricchezza nell'economia digitale?", i cittadini hanno risposto che l'economia digitale aggraverà le disuguaglianze economiche (74%) e otto persone su dieci in Italia, Spagna e Regno Unito pensano che l'uso dei dati personali sia principalmente a vantaggio delle grandi aziende tecnologiche e degli inserzionisti. Nonostante ciò, in caso di minacce globali (terrorismo, attacchi informatici, ecc...), più della metà degli intervistati si dimostra disponibile a condividere dati sensibili con aziende e governi a discapito della propria privacy.

2.2 Il processo di *Private Policy-making*: governare con l'IA

La “Retorica”, fin dai testi di Aristotele, è pura capacità di persuasione giacché il processo di selezione e rappresentazione prescinde dalla verità. Plasmare le percezioni sostituisce la presentazione dei fatti, dalla concettualizzazione alla promozione. L'esercizio retorico della scelta tra coppie dicotomiche di ammissibilità coinvolge ogni aspetto del *design* (Browne *et al.*, 2023), eticamente orientato dai proprietari delle Corporation digitali, che stabiliscono i termini, le condizioni e le modalità attraverso cui dovranno essere rispettati, aventi peso-forza in materia di moderazione dei contenuti (*Content Moderation*).

Rispetto ai dati, fin dal primo momento i contenuti hanno goduto di una maggiore libertà di movimento nell'architettura della Rete come parte imprescindibile della sua resistenza alla censura (Roberts S. T., 2017). Ciò non ha impedito l'affermarsi di forme di moderazione volontaria nelle prime comunità online, con pratiche e stili che rispecchiavano la vocazione del gruppo (come Usenet, BBS, MUD e listserv) per includere: la richiesta di modificare un contributo, la cancellazione dei post, il divieto di accesso (per account o indirizzo IP), i filtri testuali su specifiche categorie di parole/contenuti. In alcuni casi, gli utenti svolgevano mansioni per prestigio o per scopi altruistici (ancora oggi avviene su Wikipedia), in altri casi dietro remunerazione o accesso a servizi gratuiti o a tariffa ridotta (per esempio, su AOL). Nel tempo, le aziende hanno assunto personale dedicato al controllo del flusso di scambi, fatto ricorso ai sistemi di IA o chiuso definitivamente i canali di interazione con il pubblico (*ibidem*). Le piattaforme digitali che hanno sede negli Stati Uniti possono scegliere quali modelli di *Governance by Design* applicare per moderare (vale a dire come vagliare, classificare, filtrare e bloccare) i contenuti diffusi dagli utenti senza i vincoli del Primo emendamento²⁷, con cui la Costituzione degli Stati Uniti sancisce il principio di terzietà della legge rispetto al culto e al suo libero esercizio, la libertà di parola, di stampa e il diritto di appellarsi al governo. A completamento del quadro normativo, i proprietari delle piattaforme vengono svincolati dalla responsabilità editoriale secondo quanto stabilito dalla Sezione 230 (Titolo 47) del *Communications Decency Act* (CDA), Titolo V del *Telecommunications Act* (1996):

«Nessun fornitore o utente di un servizio informatico interattivo può essere considerato editore o diffusore di qualsiasi informazione fornita da un altro fornitore di contenuti informativi».

²⁷ «Il Congresso non promulgherà leggi per il riconoscimento ufficiale di una religione, o che ne proibiscano la libera professione; o che limitino la libertà di parola, o della stampa; o il diritto delle persone di riunirsi pacificamente in assemblea e di fare petizioni al governo per la riparazione dei torti».

Insieme alla Sezione 512 del *Digital Millennium Copyright Act* (DMCA, 1998), forma parte di uno scudo giuridico fondato sulla *Secondary Liability* (Longo E. , 2023, p. 247), che considera gli intermediari online come estranei dalla responsabilità per le condotte illecite di terzi attraverso i loro siti. L'immunità interviene per far fronte a grandi quantità di *upload* di UGC (Klonick, 2020), ma nell'ambito dei social media è soggetta a continue erosioni a seguito di alcuni eventi critici, quali, tra gli altri, lo scandalo Cambridge Analytica e l'arresto del fondatore di Telegram Pavel Durov²⁸. I tribunali sono stati accusati di aver interpretato la norma in senso ampio senza riconoscere l'effettiva responsabilità agli erogatori dei servizi.

Di conseguenza, nel 2023 la Corte Suprema ha accettato di accogliere due ipotesi di "favoreggiamento" in atti di terrorismo internazionale quando a promuoverlo sono sistemi automatizzati. Il dibattito continua a svilupparsi intorno all'idea di migliorare la legge piuttosto che abrogarla, arricchendola di differenziazioni tra contenuti (per esempio, monetizzabili e non), requisiti di cura, neutralità.

A discapito del tradizionale regime di *self-regulation*, la pressione per la moderazione delle piattaforme, appartenenti a società quotate in borsa, proviene spesso dagli azionisti, per cui tutelarsi dai rischi legali, finanziari e reputazionali comporta investire in approvvigionamento tecnologico (Roberts S. T., 2017) e, in egual misura, assicurarsi un vantaggio competitivo.

In "*Content or Context Moderation?*", Robyn Caplan (2018) distingue tre modelli di Corporation (*Platform Companies*) con conseguenti approcci di moderazione:

- Tradizionale (*Artisanal*), presente su Vimeo, Medium, Patreon e Discord, in cui prevale il contesto nell'analisi manuale, caso per caso, svolta da 5-200 dipendenti *in-house* per operazioni di piccola scala;
- di Comunità (*Community-reliant*), come largamente osservato su Wikimedia e Reddit, dove si integrano lavoro volontario "federale" e rispetto delle *policy*;
- Industriale (*Industrial*), adottato da Facebook, YouTube e Google, che privilegiano la coerenza delle *policy* redatte da gruppi distinti rispetto a chi si occupa della supervisione (impiegando terze parti), nell'ordine di migliaia di persone operanti nel processo.

All'interno del suo rapporto l'autrice riscontra una distinta sensibilità che deriva principalmente dalle risorse a disposizione e dalle dinamiche organizzative, come le caratteristiche della tecnologia, del modello di business e delle dimensioni dell'azienda. Ciò è particolarmente evidente nel rapporto piattaforma-motore di ricerca: sebbene entrambi utilizzino forme di moderazione, la prima se ne serve per mescolare i contenuti ("ospita, organizza e fa circolare",

²⁸ In libertà vigilata da agosto 2024, risponderà di 12 capi di imputazione per mancata collaborazione nelle inchieste su Telegram che riguardano immagini di abusi sessuali su minori, traffico di droga e transazioni fraudolente.

per dirla con Gillespie, 2018), mentre la seconda le implementa in maniera funzionale (per esempio, per la correzione automatica).

Registrandosi alle piattaforme digitali, gli utenti acconsentono alla sottoscrizione di condizioni generali, un accordo (*user agreements*) in cui vengono enunciati i doveri individuali e i diritti della piattaforma, accettati integralmente al momento dell'iscrizione con una spunta. Non è, infatti, possibile concordare modifiche e/o sostituzioni parziali, fatta eccezione per alcune sezioni aggiornabili in maniera unilaterale dal pannello di controllo.

Il preambolo di questi contratti fa esplicito rimando a due norme internazionali: la Dichiarazione universale dei Diritti Umani delle Nazioni Unite e la Dichiarazione sui Principi e i Diritti fondamentali nel Lavoro dell'Organizzazione internazionale del lavoro (ILO).

Le condizioni previste nei documenti conosciuti come *Community Standards*, *Privacy Policy* e *Terms of Service* costituiscono i regolamenti interni della maggior parte delle piattaforme digitali, comprese le già note Facebook e Instagram di Meta, X (ex Twitter) e TikTok di ByteDance (Tab. 2). A questi documenti si affiancano condizioni di servizio che articolano le caratteristiche tecniche di *design* e contenuti.

Gli Standard della community regolano la convivenza *inter pares* tra gli utenti (Santoriello, 2022, p. 287), definendo il confine di idoneità con l'obiettivo implicito di dirimere le controversie. Spesso alla redazione prendono parte esperti di tecnologia, sicurezza e pubblica e Diritti Umani e possono subire modifiche a partire dai riscontri degli utenti.

La moderazione dei contenuti è, dunque, l'insieme dei meccanismi di *governance* che struttura la partecipazione a una comunità online, in cui si riscontrano due diverse modalità di intervento: la rimozione dei contenuti contrari agli accordi o l'imposizione di sanzioni (*Hard Moderation*); l'organizzazione, la distribuzione e la diffusione dei contenuti secondo una gerarchia (*Content Curation* o *Soft Moderation*) e/o una personalizzazione dell'esperienza dell'utente (Dunn, P. in Abba *et al.*, 2022, p. 177). La moderazione può essere effettuata da esseri umani, da sistemi di IA (moderazione automatizzata o algoritmica) o da una combinazione di entrambi. Quest'ultima è largamente diffusa perché facilita il controllo ad ampio raggio, soprattutto in contesti in cui la gestione dell'abbondanza non sarebbe altrimenti possibile, per la quale si sono affermati sistemi di apprendimento basati su reti neurali, come il *Natural Language Processing* (NLP) per l'analisi dei contenuti testuali al fine di trarre conclusioni sul testo, seppur prive di comprensione semantica (*ivi*, p. 179).

Tabella 3. *Le Digital Private-policy di Meta, X e ByteDance*

Corporation	Normativa	Altre condizioni e normative
Meta (Facebook e Instagram)	• Standard della Community • Condizioni d'uso • Informativa sulla privacy • Linee guida della community	• Introduzione agli Standard pubblicitari • Normative per gli sviluppatori • Pagine, gruppi ed eventi • Linee guida sui suggerimenti • Normative sui contenuti brandizzati • Condizioni della Piattaforma • Condizioni commerciali • Condizioni per gli strumenti di Facebook Business • Condizioni d'uso e marchio di Instagram • Diffusione di contenuti terroristici online • Normativa sull'uso improprio
X (Twitter)	• Regole di Twitter • Termini di servizio • Informativa sulla privacy	
ByteDance (TikTok)	• Linee Guida della Community • Condizioni generali • Policy pubblicitarie	• Trasparenza • Privacy Policy • Informativa sui Cookie dei siti web

Fonte: Transparency Center. Elaborazione a cura dell'autrice

Rimane imprescindibile, infatti, l'intervento umano, con il proprio portato di conoscenze e competenze per prendere decisioni sugli UGC, con maturata esperienza "in questioni di gusto" del pubblico, di mediazione culturale (a prescindere dalla presenza geografica) e linguistica (inclusi i dialetti), delle leggi locali e delle norme interne di gestione (Roberts S. T., 2017).

Non è escluso che l'intervento pubblico sulle policy di moderazione orienti i modelli automatizzati per allinearli agli interessi dell'esecutivo, individuando e bloccando le informazioni ritenute sensibili (tra cui il dissenso politico, le proteste e alcune parole chiave) a partire da una comprensione contestuale, così come il monitoraggio in tempo reale della violazione delle norme o delle politiche governative. Le *Provisions on Ecological Governance of Network Information Content* emanate dal *Cyberspace Administration of China* (CAC) impongono ai fornitori di servizi online l'obbligo di dotarsi di meccanismi di *governance*, sviluppare regole dettagliate e migliorare i sistemi di registrazione, gestione degli account, revisione dei contenuti e delle informazioni, assumere personale qualificato, il tutto proporzionato alla portata dell'offerta (Zhang & Barata, 2020). Essendo prevista responsabilità legale, in caso di violazione delle disposizioni, le piattaforme e i produttori di contenuti possono essere soggetti a misure come avvertimenti, richieste di correzione, restrizioni funzionali,

sospensione degli aggiornamenti e chiusura degli account, nonché sanzioni imposte dalle autorità competenti. Aziende tecnologiche come Tencent, Alibaba e Baidu si sono prodigate a sviluppare sistemi di conformità alla legge, allenando le macchine a identificare e filtrare in modo proattivo contenuti inappropriati o illegali. Il programma di *Social Scoring* che valuta le condotte degli utenti sulla base dei dati prodotti assegnando un punteggio, potrebbe altrettanto rientrare nel quadro normativo soprattutto per l'influenza esercitata su vari aspetti della vita dei cittadini, tra cui l'accesso ai servizi.

I Dipartimenti di Sicurezza delle piattaforme si occupano dell'esclusione (*Select out*), mentre i team di Sviluppo lavorano sui sistemi di raccomandazione (*Select for*), orientandosi in base all'*engagement*, cioè a un indice derivante dal tempo speso sulla piattaforma, dal numero e dal tipo di azioni intraprese e dal grado di soddisfazione espresso (Gillespie, 2022). A queste tecniche si aggiunge l'inserimento discrezionale di etichette con funzioni di verifica, demonetizzazioni e punizioni. L'azienda può decidere se avviare processi di moderazione (*Commercial Content Moderation*, CCM) prima che il materiale venga inviato/pubblicato o dopo che è stato caricato, ovvero a seguito di reclami da parte di amministratori, moderatori, altri utenti (persone fisiche o giuridiche) o forze dell'ordine e attori governativi, purché ciò avvenga entro tempi congrui (Roberts S. T., 2017).

Nel dibattito pubblico temi come la rimozione e la riduzione della visibilità dei contenuti, così come la sospensione degli utenti - il caso della sospensione di Trump su Twitter (*Deplatforming*) è uno degli esempi più noti - hanno trovato varie forme di interpretazione. Innanzitutto, è utile chiarire che le strategie comunicative aziendali hanno impatti personali profondi sugli individui, che non sempre trovano riscontro nella controparte (Divon *et al.*, 2025), e vivono la piattaforma come "massima autorità epistemica" (Cotter, 2023). L'esempio più significativo è lo *shadowban*²⁹, una forma di riduzione della visibilità che impedisce al *creator* di avere contezza delle pubblicazioni: il contenuto o il profilo sono formalmente online, ma gli algoritmi li raccomandano meno fino a farli scomparire dai *feed*, limitando l'esposizione degli altri utenti, che possono accedervi soltanto mediante specifica ricerca. L'effetto della negazione pubblica di pratiche che gli utenti percepiscono come effettive è stato definito *Platform Gaslighting*³⁰ (Blunt *et al.*, 2020).

²⁹ Carolina Are (2022) ne ha individuato le origini nei forum online negli anni '70, con picchi di popolarità negli anni 2000, spiegando che è stato introdotto su Instagram per gestire contenuti "vagamente inappropriati", come foto di corpi femminili e contenuti sessuali non espliciti. Nel 2016 si è diffusa su Breitbart la teoria che la tecnica fosse stata creata per diminuire l'influenza dei repubblicani su Twitter.

³⁰ Il termine deriva dall'opera teatrale e cinematografica "*Gaslight*" (1938; 1940; 1944). Ambientata in epoca vittoriana, la vicenda riguarda il logoramento psicologico causato da azioni impercettibili e dalla presenza di un rapporto fiduciario tra le parti. Nel film, il marito affievolisce l'intensità della lampada a gas per poi negare di

La moderazione automatizzata dei contenuti (MAC) ha visto l'implementazione di sistemi appartenenti alle seguenti tipologie non esclusive (Bloch-Wehba, 2020):

- *Hash Technology*, differenziazione per contrasto su database di azioni ritenute illegali, quali pedopornografia, violazione della libertà intellettuale ed estremismo organizzato;
- *Age Barriers* e *Geo-blocking*, dividendo il pubblico per target e posizione dell'indirizzo IP per discriminare i contenuti in base alle norme vigenti e ai confini geografici;
- *Antispam*, suddivisione dei contenuti in relazione alla quantità di volte che un'informazione viene pubblicata o raggiunge il suo pubblico di riferimento.

La rilevazione avviene a partire dal *dataset* utilizzato per l'addestramento, intervenendo sui nuovi contenuti in caso di corrispondenze o forti somiglianze con un analogo post già contrassegnato in precedenza come non conforme (Santoriello, 2022), in virtù di strategie censorie che cristallizzano il binarismo di liceità. Quando il procedimento non viene notificato agli utenti, diventa complicata sia l'identificazione che la giustificazione. Infatti, mentre la rimozione è visibile e, quindi, esplicita e problematica, la classificazione flessibile del contenuto ritenuto ambiguo (o legato a fenomeni emergenti, come QAnon) come “*borderline*” comporta un depotenziamento sia in visibilità che in priorità, relegandolo in coda (*Up next*), quando non del tutto oscurato. Oltre alla soppressione algoritmica non notificata, rientrano in questa categoria anche i dinieghi automatizzati, i pregiudizi nella moderazione a favore di account più grandi e gli ostacoli nell'accesso all'appello (Divon *et al.*, 2025).

Secondo Gillespie (2022), il campo semantico creato *ad hoc* dalle piattaforme è un tentativo per evitare critiche “politiche” che trova margini di manovra nelle maglie delle proprie *policy* e agisce in nome della sicurezza delle altre persone coinvolte, preservando la propria autonomia. Anche se si preferisce la restrizione alla rimozione, non si evitano gli impatti iniqui sulla società e su diversi gruppi/comunità, come la presenza di *bias* e le implicazioni per la libertà di espressione (*ivi*, p. 8). I processi di MAC e di raccomandazione algoritmica fanno parte della *governance* digitale, per cui vengono redatti e sottoscritti specifici accordi di riservatezza (*Nondisclosure Agreements*, NDA) con dipendenti, collaboratori e fornitori. L'autore, dipendente Microsoft, ammette di possedere un privilegio di status nella conduzione della ricerca, conservando problematicità dal punto di vista metodologico (*ibidem*).

La sua indagine ha riguardato la moderazione di YouTube, definito da Tufekci “*the Great Radicalizer*” (2018) a causa del sistema di raccomandazione, tendente a dare risalto a video estremi a prescindere dall'argomento. Le *policy* introdotte nel 2019 si apprestavano a ridurre,

averlo fatto, facendo credere alla moglie di essere colpevole dei danni emotivi che prova e mostrandosi, al contempo, compassionevole e disposto al dialogo, su cui è conscio di avere un grande ascendente (si veda la scena “*You think I'm insane*”).

pur mantenendo disponibile in archivio, la diffusione di quei contenuti che si avvicinavano, senza oltrepassare il limite, a violazioni (come la cospirazione) delle Linee Guida della Community oppure volti a promuovere forme acute di disinformazione. I segnali raccolti dai sistemi più sofisticati possono, infatti, modificarsi e adattarsi in base al profilo dell'utente, alla sua collocazione geografica o al suo orientamento di genere. I lavoratori *tech* che lavorano per YouTube, intervistati da Gillespie (2022, p. 6), segnalavano la presenza di due passaggi chiave nella definizione dello schema di raccomandazione: l'inventario (*Candidate Generation*, cioè la definizione del profilo e la possibilità che un video possa essere ritenuto rilevante) e il *ranking* (cosa mostrare e in quale ordine). La riduzione della visibilità avrebbe comportato una mancata raccomandazione “del tutto”, al di fuori dell'inventario, “in parte”, diminuendo la valutazione nel *ranking*, o “per alcuni/e”, in base ai dettagli disponibili (età, regione, tracce raccolte, se collegato con modalità incognita o in linea con le preferenze espresse).

Il 10 febbraio 2024 Mosseri, CEO di Instagram, ha annunciato su Threads (Fig. 5) l'adozione di una nuova *policy* di raccomandazione dei contenuti politici sulle sue piattaforme, per restituire all'utente il potere di fare “*Opt in*”, ovvero di esplicitare il consenso a seguire un profilo e continuare a essere esposto ai suoi contenuti nelle sezioni *Feed* e *Stories*. Gli account pubblici hanno subito un'alterazione della propria visibilità nelle sezioni *Esplora*, *Reels* e *Utenti*, suggeriti secondo la determinazione della natura “politica”, ovvero riguardante temi come leggi, elezioni e argomenti a sfondo sociale: «Il nostro obiettivo è quello di preservare la possibilità per le persone di scegliere di interagire con i contenuti politici, rispettando al contempo l'appetito di ciascuno», ha dichiarato Mosseri, giustificando la decisione come una richiesta proveniente dal pubblico.



Figura 5. Mosseri descrive su Threads il Political Content Approach
Fonte: Threads.net/@mosseri/post/C3IjVOovcS5 (12 febbraio 2024)

L'impostazione *Account Status* garantisce agli account professionali l'idoneità alla raccomandazione soltanto a patto che rinuncino ai contenuti politici (pubblicati o in pubblicazione), scrive Todd Spangler (2024). Se segnalati, potranno “smettere di pubblicare questo tipo di contenuti per un periodo di tempo” per riottenere la visibilità.

È un esempio di de-responsabilizzazione della Corporation perché la moderazione è ritenuta controversa tanto quanto il contenuto³¹.

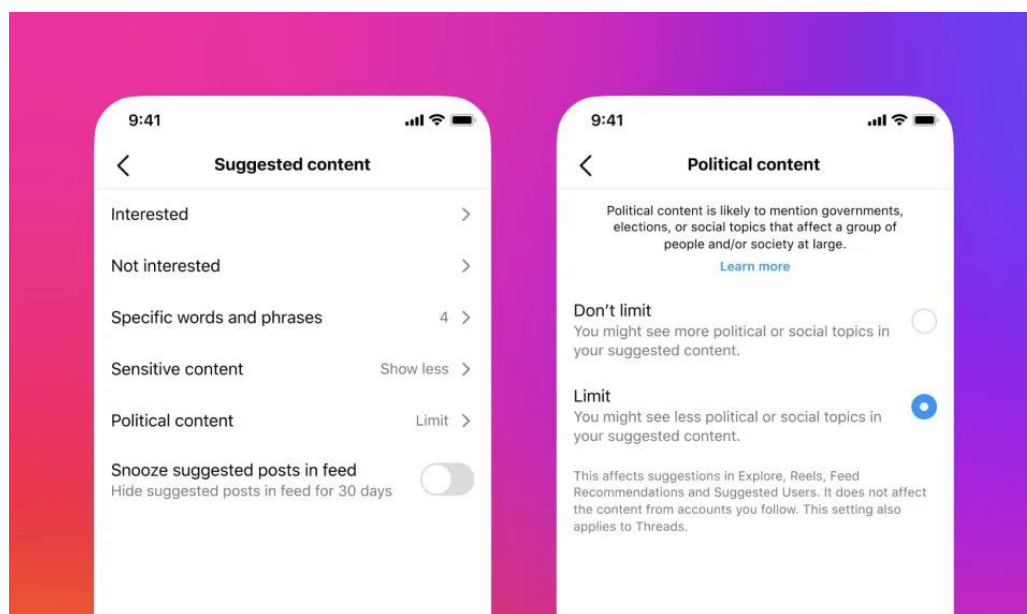


Figura 6. La funzione di consenso alla fruizione di “Contenuti politici”

La moderazione della libertà di espressione online, nella sua connotazione politica, è stata affrontata anche in contesti non occidentali, dove sono state proposte o approvate nuove norme per regolare il flusso delle pubblicazioni «costringendo le società di social media a fungere da censori», secondo Human Rights Watch (2018). A Singapore, il Governo si era posto in prima linea per affrontare le *fake news*; nelle Filippine è entrato in vigore nel 2021 l'Atto 1942 per sanzionare la distribuzione di contenuti malevoli, quali notizie false e violazioni correlate. In Russia, su due proposte presentate alla Duma dal partito Russia Unita, una si rivolgeva direttamente agli “organizzatori della diffusione di informazioni”, inserendo tra i contenuti illegali le informazioni sulla propaganda di guerra, l'incitazione all'odio nazionale, razziale o religioso, la diffamazione su onore, dignità e reputazione e le violazioni delle leggi amministrative e penali nazionali. Il Venezuela, nel 2017, ha adottato la *Ley Constitucional contra el Odio, por la Convivencia Pacífica y la Tolerancia (Ley contra el Odio)* contenente

³¹ L'AGCOM ha avanzato una richiesta di informazioni in vista della consultazione elettorale europea 2024. «Le piattaforme online sono tenute ad assumere ogni utile iniziativa volta ad assicurare il rispetto dei principi di tutela del pluralismo della libertà di espressione, dell'imparzialità, indipendenza e obiettività dell'informazione» secondo l'art. 30 della delibera 90/24/CONS (Comunicato stampa, 22 maggio 2024).

sanzioni cospicue per le piattaforme di social media che non provvedono alla rimozione dei contenuti su «propaganda di guerra o di odio nazionale, razziale, religioso, politico o di qualsiasi altro tipo» entro sei ore dalla pubblicazione. Pochi mesi prima, l'Autorità nazionale per le comunicazioni del Kenya ha pubblicato linee guida contenenti l'obbligo di chiusura degli account responsabili della diffusione di contenuti politici indesiderati (cioè messaggi politici offensivi, violenti, fuorvianti, osceni o blasfemi) entro 24 ore dal reclamo.

2.3 Il margine conteso. Analisi comparata sulla *Automated Content Moderation*

Nell'ambito della Scienza Politica, l'analisi comparata è un tassello cardine del metodo scientifico (Lijphart, 1971, p. 682). Differenze e specificità dei fenomeni politici in diversi contesti del mondo non sarebbero altrimenti comprensibili se approfonditi soltanto in uno studio di caso o in assenza di un'associazione in parallelo tra sistemi, istituzioni, politiche e risultati che possono, invece, essere identificati in modelli, nessi causali e fattori *ceteris paribus* in grado di plasmare la dinamica politica. Benché sia manchevole di una misurazione ordinale (*ibidem*), questa metodologia costruisce teorie e sviluppa ipotesi in cui si osservano variabili indipendenti e dipendenti che causano una dinamica politica in relazione con le istituzioni che determinano le regole del gioco (Caramani, 2023, p. 25). Nell'analisi delle *policy*, la procedura è essenziale per valutare l'applicazione e l'impatto dei provvedimenti: è ragionevole affermare che, durante la fase istruttoria, il paragone con altre giurisdizioni costituisce evidenza empirica sull'argomento nonché fonte di ispirazione per processi di riforma e progettazione.

L'analisi qualitativa comparata (QCA) sulla regolamentazione interna ed esterna delle piattaforme è idonea a confrontare i confini della sovranità tra attori pubblici e privati nello spazio conteso del digitale. Inoltre, muovendoci nel quadro normativo, è possibile analizzare quanto i provvedimenti possano garantire o meno il rispetto dei principi democratici, degli standard in materia di Diritti Umani e formulare considerazioni etiche per promuovere la *governance* democratica e la giustizia sociale. Secondo la prospettiva costruttivista, dati e valori sono socialmente costruiti (*ivi*, p. 37) e non sarebbero altrimenti comprensibili senza un'analisi contestuale da parte di chi conduce la ricerca, che si adopera per fornire un'interpretazione sociale e culturale dei fenomeni osservati. L'affermazione dei valori collettivi nel digitale e le modalità attraverso cui vengono declinati rappresentano un campo di osservazione distinto dall'azione sociale del singolo individuo, che quasi prescinde dal suo operato.

Il metodo comparativo viene impiegato quando il numero di casi disponibili per l'analisi è talmente esiguo da rendere impossibile un'ulteriore tabulazione incrociata (Lijphart, 1971, p. 684), come nel caso della regolamentazione nell'ambito della MAC. Attraverso il metodo delle

differenze (Lijphart, 1971, p. 687), si indagano le variazioni concomitanti in contesti simili, osservabili per quelle piattaforme che hanno inserito la moderazione come elemento di *design* e, dunque, di *governance*, intesa come quella discrezionalità che mira a stabilire obiettivi, trova i mezzi per raggiungerli e si modella in base agli impatti prodotti dalle proprie decisioni.

Gli approcci politologici utilizzati per osservare i sistemi amministrativi sono per lo più funzionalisti e si concentrano sugli attori che li determinano a prescindere dalla ripartizione formale delle mansioni stabilita dalla legge (Caramani, 2023, p. 47-48). Nel complesso ecosistema digitale, la *governance* si esplicita con le scelte di moderazione dei contenuti orientate dal singolo o dal gruppo che detiene la proprietà della piattaforma.

Fulcro della ricerca è comprendere se la libertà di espressione sia diventata uno spazio conteso di sovranità tra le Corporation e lo Stato, per approfondire la tensione del digitale tra privato e pubblico nel “sistema politico” dell’Unione europea (*ivi*, p. 445), in cui sono presenti istituzioni, organizzazioni intermedie e decisioni collettive che hanno impatto sulla distribuzione delle risorse, così come sull’allocazione dei valori, con continuità. Il tema della moderazione dei contenuti, infatti, trova spazio nel DSA e in altri atti legislativi (dedicati, per esempio, ai contenuti terroristici), in continuità con la competenza esclusiva dell’UE sulla regolazione del mercato unico, inclusa la gestione delle politiche di concorrenza, doganali e commerciali, da cui si evince l’esercizio della sovranità. La disciplina, se da un lato ha riguardato la rimozione delle barriere fiscali, fisiche e tecniche per gli Stati membri, dall’altro concerne questioni riguardanti le politiche sociali come i diritti di lavoratori e lavoratrici, la parità di genere e il contrasto alle discriminazioni di razza, religione, disabilità, età e orientamento sessuale, intervenendo per correggere i “fallimenti di mercato” (*ivi*, p. 448).

L’art. 1 (DSA, 2022) chiarisce che il legislatore, stabilendo norme armonizzate per il «corretto funzionamento del mercato interno dei servizi intermediari», intende facilitare l’innovazione in un ambiente online «sicuro, prevedibile e affidabile» e in cui trovino applicazione e tutela effettiva i diritti fondamentali «compreso il principio della protezione dei consumatori», nonché attuazione ed esecuzione degli stessi mediante cooperazione e coordinamento tra le autorità competenti. L’indagine riguarda, pertanto, le piattaforme di social media (indipendentemente dal luogo di stabilimento, art. 2 DSA), che rispecchiano i seguenti requisiti:

- a) l’utente “destinatario del servizio” è libero di postare contenuti (ovvero “ricercare informazioni o renderle accessibili”, art. 3, lett. b-p, DSA);
- b) la *governance* della piattaforma operante in UE prevede un meccanismo di moderazione delle informazioni (incluse le operazioni di *mere conduit*, *caching* e *hosting* che la qualificano come servizio intermediario, art. 3, lett. g-i, DSA);

c) la Corporation raggiunge un numero di utenti congruo con la definizione di VLOP (artt. 3, lett. e, e 33 DSA).

Si considera l'ipotesi che la regolamentazione esterna, in vigore nell'Unione europea e/o considerata nella sua legittimità statale, sia una variabile indipendente (X) e che la regolamentazione interna sia una variabile dipendente (Y), come raffigurato nella Figura 7.

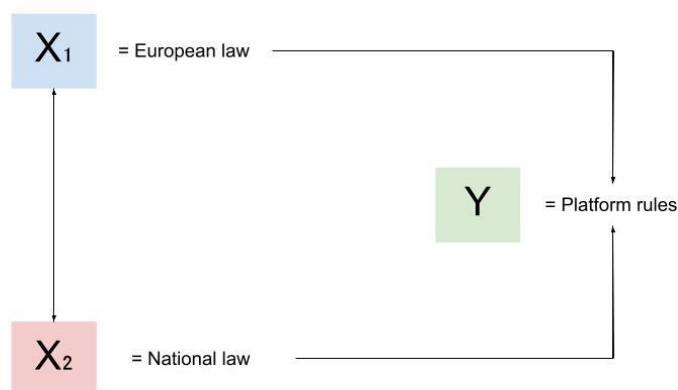


Figura 7. *Relazione tra autorità della legge e potere di moderazione dei contenuti digitali*
Fonte: Elaborazione dell'autrice adattata a partire da Caramani (2023)

A partire dal 25 aprile 2023, le piattaforme e i motori di ricerca inseriti all'interno degli elenchi VLOP e VLOSE hanno dovuto conformarsi agli obblighi stabiliti dal DSA, comprendenti una moderazione dei contenuti più diligente (Commissione europea, 2023) composta da:

- misure per affrontare la diffusione di contenuti illegali online e gli effetti negativi sulla libertà di espressione e di informazione;
- condizioni generali chiare e applicabili in modo diligente e non arbitrario;
- meccanismi di segnalazione dei contenuti illegali con provvedimenti rapidi;
- analisi dei propri rischi specifici e misure di attenuazione – per i casi di disinformazione e uso non autentico dei loro servizi.

Grazie alla pubblicazione dei *Transparency Report* molte delle attività di gestione sono state palesate e tracciate, in ottemperanza al DSA (artt. 15, 24 e 42).

L'operativizzazione dell'analisi prende in considerazione alcune delle principali caratteristiche delle piattaforme (casi), quali l'organigramma interno e la struttura societaria, unendole all'implementazione esplicita o implicita della MAC (unità di osservazione), a come viene applicata e giustificata. Infine, si indagano gli impatti sulla libertà di espressione che possono aver subito alterazioni in base a variazioni delle norme o delle regole di funzionamento tra il

2018 e il 2024. Le piattaforme delle Corporation digitali Meta (Facebook e Instagram), Twitter (X) e ByteDance (TikTok) sono presenti nell'elenco delle VLOP (Tab. 2).

2.3.1 La libertà di espressione come principio regolatore

Il problema di Ganton (Caramani, 2023, p. 63), dove differenze e similarità causate da fattori esogeni incontrano le particolarità derivanti da fattori endogeni, è la proprietà delle Corporation. Meta e X sono accomunate da una gestione proprietaria con leadership esplicita: Mark Zuckerberg (fondatore, presidente e CEO) ed Elon Musk (presidente e direttore tecnico, con Linda Yaccarino CEO) fungono da protagonisti dell'evoluzione societaria. TikTok (o Douyin) è sviluppata dalla società ByteDance appartenente a Zhang Yiming (fondatore e CEO fino al 2021), l'attuale CEO è Shou Zi Chew e, come avviene in altri settori produttivi, il governo cinese ha una rappresentanza nel CdA di una sussidiaria (con una *golden share* dell'1%), anche se non è presente nella società madre. Tutti e tre sono tra gli uomini più ricchi di USA e Cina e, dunque, del mondo. Zuckerberg e Yiming condividono la formazione in informatica e ingegneria informatica, mentre Musk ha condensato competenze tra studi in economia e fisica e attitudini personali di programmazione.

L'inquadramento della libertà di espressione risente di questi fattori nella sua declinazione. Nella sezione *Transparency Center*, Meta afferma che la tutela della libertà di espressione è considerata al pari della sicurezza degli utenti, in un connubio che si realizza attraverso il processo di informazione tra piattaforma e utente. Il *Transparency Report* di X si apre, invece, con uno *storytelling* sul ruolo che ha avuto la piattaforma nel diffondere contenuti rilevanti dal punto di vista sia storico-politico che simbolico («dai movimenti sociali ai meme più potenti del mondo») in un luogo che ha accolto l'espressione delle idee anche quando considerate scomode o non convenzionali. X è uno spazio di estrema libertà di parola purché le prospettive espresse non risultino moleste, minacciose, disumanizzanti o usino la paura per reprimere le voci altrui, garantendone la sicurezza. Questa tensione comporta un lavoro minuzioso di svisceramento che tutela prima di ogni altro chi ha espresso la propria opinione:

«I nostri sistemi di moderazione dei contenuti sono progettati e adattati per mitigare i rischi sistemici senza limitare inutilmente l'uso del nostro servizio e i diritti fondamentali, in particolare la libertà di espressione. Le attività di moderazione dei contenuti sono implementate e ancorate a *policy* di principio e sfruttano un insieme di interventi per garantire che le azioni intraprese siano ragionevoli, proporzionate ed efficaci» (*X Transparency Report*, 2023).

La restrizione dei contenuti viene considerata un detrimento da ridurre laddove possibile, in assenza di una giusta motivazione («Se si fa l'una o l'altra cosa, bisogna avere ragione. E viviamo in un mondo con molte sfumature di grigio», *X Transparency Report*, 2023).

Per TikTok l'espressione del *creator*, chi produce il contenuto (utente o inserzionista), è vincolata prioritariamente alla legge locale e l'azienda sostiene di avere un record nella pubblicazione proattiva dei report sulla trasparenza, che risalgono al 2019.

I prestatori di servizi intermediari sono tenuti ad agire in modo diligente, obiettivo e proporzionato, secondo quanto stabilito dall'art. 14 (DSA, 2022) e tenuto conto degli interessi legittimi delle parti coinvolte e dei diritti fondamentali dei destinatari, tra cui si segnalano «la libertà di espressione, la libertà e il pluralismo dei media».

I tre approcci trovano significazione anche analizzando la *core mission* aziendale: mentre Meta propone un ambiente in cui possano transitare scambi di natura sociale e commerciale e X si pone al servizio della conversazione pubblica, TikTok si posiziona come un canale di intrattenimento e divulgazione. La dimensione della comunità, seppur imprescindibile come elemento di funzionamento per i social media e costante nelle ambizioni delle tre piattaforme, viene declinata secondo l'interesse particolare di ciascuna, talvolta come nicchia di mercato, opinione pubblica o *audience*.

La strategia di gestione adottata dalle tre piattaforme riflette il posizionamento designato. Meta presenta un approccio globale, ponendosi come un mezzo di comunicazione universale intercettando un bisogno delle persone di mettersi in contatto che prescinde dalla provenienza geografica. In X è tangibile il riferimento ai valori e alle norme nazionali, in particolare al Primo emendamento della Costituzione degli Stati Uniti - che sancisce la terzietà della legge sul culto, la libertà di parola e di stampa, il diritto a riunirsi pacificamente - nell'applicazione delle scelte strategiche. TikTok adotta un organigramma decentralizzato con particolarità che si modellano e adattano al contesto specifico. Le sedi centrali si trovano a Los Angeles e Singapore, ma dispone di ulteriori uffici a New York, Londra, Dublino, Parigi, Berlino, Dubai, Giacarta, Seul e Tokyo e società sussidiarie, specificando nei suoi Termini di Servizio che «la società con cui l'utente stipula il contratto dipende dal suo luogo di residenza anagrafica o fiscale». L'azienda si è dotata anche di otto Consigli consultivi sulla sicurezza in Asia-Pacifico, Brasile, Europa, America Latina, MENAT (Medio Oriente, Nord Africa, Turchia) e di un Consiglio consultivo sui contenuti negli Stati Uniti (Tik Tok, 2023).

In base agli artt. 9, 10 e 15 del DSA, le piattaforme sono tenute a rendere conto dei “contenuti illegali” diffusi, secondo quanto stabilito dal diritto nazionale e dell'Unione, e a fornire informazioni sui destinatari dei servizi quando richiesto dalle autorità nazionali degli Stati membri a livello giudiziario o amministrativo. Va osservato che già i precedenti articoli contemplano la possibilità di intervento da parte di un'autorità nazionale che «esiga che il prestatore del servizio impedisca o ponga fine a una violazione». Ricevuto l'ordine,

comprensivo delle motivazioni e delle informazioni sui meccanismi di ricorso, in capo ai prestatori spetta l'obbligo di darvi seguito e di informare senza indebito ritardo l'autorità che ha emesso l'ordine delle azioni intraprese. A partire da questo momento l'autorità nazionale provvede a informare il Coordinatore dei servizi digitali dello Stato membro³² (DSC), che trasmette una copia dell'ordine ai suoi analoghi secondo il sistema descritto nell'art. 85, mentre il prestatore informa il destinatario del servizio del provvedimento e degli annessi (motivazione, possibilità di ricorso e ambito di applicazione). Qualora sussista il sospetto di un reato commesso o prossimo, il prestatore deve notificare alle autorità giudiziarie o all'Europol le informazioni pertinenti disponibili (art. 18 DSA). È tenuto, inoltre, a indicare un unico canale per comunicare con le autorità, la Commissione e il comitato e di renderne fruibile almeno uno per i destinatari del servizio unitamente a «mezzi di comunicazione che non si basino unicamente su strumenti automatizzati» (artt. 11-12). Francia e Germania sono i primi Stati per intervento nel 2024: 1562 ordini da parte delle autorità competenti tedesche su Facebook (Meta), pari al 74% del totale; 2624 ordini dal governo francese su TikTok su questioni riguardanti la sicurezza nazionale contro il terrorismo; su X, 2465 dalla Germania sul discorso d'odio e 1985 dalla Francia sui rischi per la pubblica sicurezza.

All'art. 22, il DSA introduce l'incarico di “Segnalatori attendibili” (*Trusted flaggers*³³) per enti designati dai DSC con requisiti di capacità e competenza particolari ai fini dell'individuazione, dell'identificazione e della notifica dei “contenuti illegali”, indipendenti e che assicurino diligenza, accuratezza e obiettività. Alla fine del 2024, soltanto Austria, Danimarca, Finlandia, Svezia, Francia, Ungheria, Grecia, Germania e Romania hanno già avanzato le nomine.

A cadenza annuale gli enti sono tenuti alla pubblicazione di un report di sintesi sulle notifiche presentate, sui tipi di contenuti illegali segnalati e sulle azioni intraprese dalle piattaforme online. Alle VLOP, inoltre, spettano periodiche valutazioni dei rischi che possono scaturire dall'introduzione di funzionalità con impatti critici individuati dall'art. 34 del DSA: diffusione di “contenuti illegali”, effetti negativi attuali o prevedibili in materia di diritti fondamentali (dignità umana, libertà di espressione informazione, non discriminazione, tutela dei consumatori), dibattito civico e processi elettorali, sicurezza pubblica, violenza di genere, protezione della salute pubblica e dei minori, benessere fisico e mentale della persona. La valutazione deve essere completa anche dell'eventualità che i rischi possano essere influenzati dalla manipolazione intenzionale del servizio («mediante l'uso non autentico o lo sfruttamento

³² AGCOM ha assunto l'incarico per l'Italia: www.agcom.it/dsa

³³ Una forma istituzionale di moderazione decentralizzata dei contenuti che fa eco alle “Community Notes” di X e Meta. Save The Children ha ottenuto il titolo sia in Finlandia che in Romania. Elenco disponibile su digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa

automatizzato») e adottare misure di attenuazione ragionevoli, compreso l'adeguamento delle condizioni generali, delle procedure di revisioni e di tutti i processi decisionali (art. 35 DSA). Infine, è riscontrabile un disallineamento nei periodi di osservazione a cui si riferiscono i *Transparency Report*: nel 2023, il monitoraggio di Meta ha riguardato il semestre aprile-settembre, X il periodo che va dal 28 agosto al 31 ottobre e TikTok si è limitato al solo mese di settembre; nel 2024, Meta e X hanno pubblicato il secondo report per il semestre ottobre-marzo, TikTok ha coperto gennaio-giugno.

2.3.2 I “contenuti illegali”

L'art. 3 (DSA, 2022) fornisce chiarimenti semantici per rendere effettiva l'interoperabilità del regolamento e la corretta applicazione nel contesto frammentato in cui operano le piattaforme. Tra questi, rientra la definizione di “contenuto illegale”:

«Qualsiasi informazione che, di per sé o in relazione a un'attività, tra cui la vendita di prodotti o la prestazione di servizi, non è conforme al diritto dell'Unione o di qualunque Stato membro conforme con il diritto dell'Unione, indipendentemente dalla natura o dall'oggetto specifico di tale diritto» (art. 3, lett. h, DSA).

I contenuti illegali sono definiti dalla violazione della legislazione europea o nazionale. Se il contenuto è illegale solo in un determinato Stato membro, sarà rimosso solo nel territorio in cui è illegale. Meta lo applica anche a livello globale e rende noti i dati in una apposita sezione sulle Restrizioni³⁴. Al contrario, il “contenuto sano”, come lo definisce Gillespie (2022, p. 8), è parte di uno specifico disegno dei proprietari delle piattaforme. L'operazione di scelta dei criteri che determinano l'assegnazione di un'informazione alla categoria “illegale” è direttamente connessa all'esercizio della libertà di espressione e alle modalità stabilite e implementate per rendere esecutiva la sua rimozione e/o restrizione, ovvero l'incompatibilità con quanto stabilito nelle *policy*. Per questo motivo l'art. 3 definisce anche cosa si intende per “moderazione dei contenuti”:

«Le attività, automatizzate o meno, svolte dai prestatori di servizi intermediari con il fine, in particolare, di individuare, identificare e contrastare contenuti illegali e informazioni incompatibili con le condizioni generali, forniti dai destinatari del servizio, comprese le misure adottate che incidono sulla disponibilità, sulla visibilità e sull'accessibilità di tali contenuti illegali o informazioni, quali la loro retrocessione, demonetizzazione o rimozione o la disabilitazione dell'accesso agli stessi, o che incidono sulla capacità dei destinatari del servizio di fornire tali informazioni, quali la cessazione o la sospensione dell'account di un destinatario del servizio» (art. 3, lett. t, DSA).

³⁴ Meta, *Content Restrictions Based on Local Law*, Transparency Center, <https://transparency.meta.com/reports/content-restrictions> (30 dicembre 2024).

Le Corporation digitali si sono dotate della MAC per la classificazione dei contenuti illegali, allenando i propri modelli alla riproduzione della valutazione umana sulla tipologia di contenuto e sul significante espresso. Il giudizio umano viene simulato dai *classifier* (Gillespie, 2022, p. 7) al punto da poterlo replicare su milioni di pubblicazioni quotidiane. La moderazione degli account e dei contenuti segue principalmente due orientamenti: orizzontale, tra utenti mediante apposite funzioni (*affordances*) di segnalazione³⁵, e verticale, con metodi di revisione automatizzata e/o umana (Santoriello, 2022). In particolare, Meta e TikTok sono accomunate dalla coesistenza di una moderazione proattiva e reattiva, mentre X preferisce la moderazione contestuale (Appendice A). Ciò non avviene senza criticità (Crawford & Gillespie, 2014): gli utenti possono disapprovare il discorso altrui, relegando l'opportunità di dire alcune cose alla mercé e al giudizio della *governance* che, come si vedrà nel Capitolo 3, lascia ancora a desiderare sul piano del reclamo e dell'opposizione. Inoltre, al momento del *flag*, è la piattaforma a fornire il ventaglio di categorie tra cui scegliere, predeterminando sia la tipologia che il lessico per articolare i "sentimenti complessi" (*ibidem*) degli utenti.

Il meccanismo di segnalazione e azione (art. 16 DSA) deve essere predisposto per qualsiasi persona o ente al fine di notificare la presenza nel servizio di informazioni ritenute illegali, esclusivamente per via elettronica, e, qualora il trattamento dell'istanza o la decisione vengano presi mediante strumenti automatizzati, tale dettaglio va comunicato al momento della segnalazione. Nel *Transparency Report* di TikTok (2023) si esplicita che la piattaforma informerà gli utenti sulla idoneità del contenuto e della decisione di ripristino.

Su Meta i sistemi di IA contribuiscono alla classificazione "reputazionale". I modelli di apprendimento automatico adattano le previsioni sull'utilità e la pertinenza dei contenuti sui segnali di input (informazione o dato), tra cui rientrano le caratteristiche del post, il livello di coinvolgimento e le interazioni passate, un sistema in scala per cui un maggiore punteggio è associato alla maggiore probabilità che venga rimosso e un avvicinamento alla soglia di legittimità corrisponde a una riduzione algoritmica della visibilità (Gillespie, 2022, p. 7). Nei *feed* di Facebook e Instagram, per esempio, la classificazione viene ponderata in base a una "combinazione bilanciata" di più sistemi di IA che suggeriscono i contenuti secondo criteri quali prossimità (amici/follower), affinità, affiliazione (appartenenza a gruppi o cerchie) e disponibilità economica (inserzioni), un apprendimento dei gusti costante, in grado di prevedere quali eventuali nuove cose potrebbero piacere.

³⁵ Il *flag* (Crawford & Gillespie, 2014) fu intensificato a seguito delle numerose minacce di stupro che nel 2013 colpirono la giornalista Caroline Criado-Perez su Twitter; ritenuto inadeguato per gestire abusi online, 120.000 firmatari chiesero alla Corporation di cambiare le *policy* sulla segnalazione.

X basa la sua tecnica di moderazione sulla filosofia “*Freedom of Speech Not Reach*” (FoSNR) ed evidenzia nel suo *Transparency Report* 2024 che il contenuto non riceverà restrizioni a meno che non sia in violazione delle *policy*:

«Non abbiamo applicato alcun filtro di visibilità basato su contenuti illegali. Nei casi in cui riceviamo una richiesta di contenuti illegali e il post risulta essere anche in violazione della politica in cui viene applicato il FoSNR, verrà sempre applicata la nostra applicazione per i contenuti illegali» (*X Transparency Report* 2024).

L’approccio alla moderazione di TikTok si concentra su quattro principi: rimozione dalla piattaforma di contenuti inappropriati, restrizioni in base all’età, controllo degli standard di idoneità, informazione tra piattaforma e community (inclusi strumenti e risorse).

2.3.3 Misure di restrizione e penalizzazione

L’art. 17 DSA indica che i prestatori di servizi di memorizzazione sono tenuti a fornire agli interessati una “Motivazione” chiara e specifica per le restrizioni imposte, tra cui:

- «a) eventuali restrizioni alla visibilità di informazioni specifiche fornite dal destinatario del servizio, comprese la rimozione di contenuti, la disabilitazione dell’accesso ai contenuti o la retrocessione dei contenuti;
- b) la sospensione, la cessazione o altra limitazione dei pagamenti in denaro;
- c) la sospensione o la cessazione totale o parziale della prestazione del servizio;
- d) la sospensione o la chiusura dell’account del destinatario del servizio» (art. 17 DSA).

Fatta eccezione per i contenuti commerciali ingannevoli, l’informazione sulla decisione deve contenere i fatti e le circostanze, l’origine della segnalazione (se da parte di un utente, da indagini volontarie o da strumenti automatizzati), la portata territoriale e la sua durata, se riguarda contenuti illegali o incompatibilità con le condizioni generali e, infine, informazioni chiare e di facile comprensione sui mezzi di ricorso.

All’interno dei *Transparency report* di Meta, X e TikTok non viene direttamente menzionato lo *shadowban*, ma già nell’art. 3, lett. t, si contempla la pratica della “retrocessione” tra le misure che incidono su disponibilità, visibilità e accessibilità dei contenuti. Meta, tuttavia, dichiara che i suoi sistemi di IA sono in grado di prevedere l’idoneità alla permanenza sulla piattaforma e, di conseguenza, attribuire, quando non rimosso, un valore inferiore rispetto all’indice di priorità stabilito («[...] se si prevede che un contenuto risulterà sgradevole in base ai nostri standard, alle nostre linee guida o alle nostre normative in materia di integrità, il contenuto verrà rimosso o visualizzato più in basso nel *feed*»³⁶). Ai modelli di apprendimento

³⁶ Il nostro approccio per spiegare la classificazione (Transparency Center, 31 dicembre 2023), <https://transparency.meta.com/features/explaining-ranking> (24 aprile 2024).

automatico viene, dunque, delegato il giudizio non soltanto qualitativo ma di notiziabilità del contenuto. Ciò diventa particolarmente problematico, riprendendo l’informativa che compare nel *Transparency Center* di Meta sulla strategia “rimuovere, ridurre, informare”, perché l’azienda afferma che la riduzione avviene anche in assenza di una accertata violazione:

«[...] riduciamo la distribuzione dei contenuti problematici anche quando non sono in violazione rispetto alle nostre normative e informiamo gli utenti, fornendo loro dettagli aggiuntivi sul contesto in modo che possano decidere con maggiore facilità su cosa cliccare, che contenuti leggere e quali post condividere».

Si evince che la pubblicazione di un elemento che differisce dalla norma, pur non trasgredendo ad essa, sarà svantaggiato nella probabilità di raggiungere il pubblico. I destinatari del messaggio possono visionare un *disclaimer* che informa sulla caducità del contenuto, ma non sussiste garanzia per chi pubblica e/o condivide.

La già citata strategia sui “contenuti politici” aggiorna una *policy* di riduzione della visibilità dei contenuti lanciata per Facebook nel 2018 attraverso una esaustiva guida (*Content Distribution Guidelines*) comprendente la tipologia di contenuti problematici (*borderline*, disinformazione, *spam* e *clickbait*) che sarebbero stati declassati in Bacheca. A seguire, nel 2021 Instagram comunicò l’introduzione di una nuova possibilità per gli utenti di abilitare/limitare (opzione base) o limitare ulteriormente i contenuti sensibili (riguardanti sesso, droga e armi) nella sezione *Esplora*. Gillespie (2022, p. 4) osserva che, implicitamente, l’azienda stava comunicando un’operazione di moderazione già in essere di *default* e non sembra dissimile rispetto alle versioni più recenti. Zuckerberg affermò che gli utenti erano naturalmente inclini a interessarsi a contenuti al limite tra il consentito e il proibito dai *Community Standards*, per cui oltre ad agire sulla domanda si tentò di silenziarli del tutto (*ibidem*).

TikTok, per alcuni contenuti (*spam*, video in revisione o video che potrebbero turbare il pubblico), riduce la visibilità fin dalla funzione di ricerca, inclusa la probabilità che l’utente venga reindirizzato su altri risultati, e ne limita la distribuzione nella sezione *Per te*. Dichiarò, altresì, di attuare azioni simili di riduzione della visibilità nel paese in cui viola la legge, in linea con la strategia regionale, o di procedere con la rimozione integrale del contenuto quando in palese violazione delle condizioni generali. X preferisce distinguere i contenuti per l’impatto che generano sulla platea: quando l’informazione diffusa infastidisce o trasmette sensazioni controverse pur essendo lecito, la piattaforma preferisce restringere il pubblico e apporre un’etichetta piuttosto che rimuoverlo (Fig. 8). Per quanto concerne la sospensione dell’account, a seguito di un avviso preventivo, è prevista sia per coloro che forniscono contenuti manifestamente illegali che per coloro

che abusano dei meccanismi di segnalazione e dei sistemi interni di reclamo presentando con frequenza istanze infondate (art. 23 DSA). Fungono da prove circostanziali il numero di azioni presentate entro un determinato arco temporale, la proporzione sul totale delle informazioni, la gravità degli abusi e l'intenzione.

X tende a evitare quanto più possibile l'incidenza della sanzione sull'account per cui prevede che la violazione comporti il passaggio dalle piene funzioni alla “sola lettura”, con limitazioni della possibilità di postare e interagire sulla piattaforma.

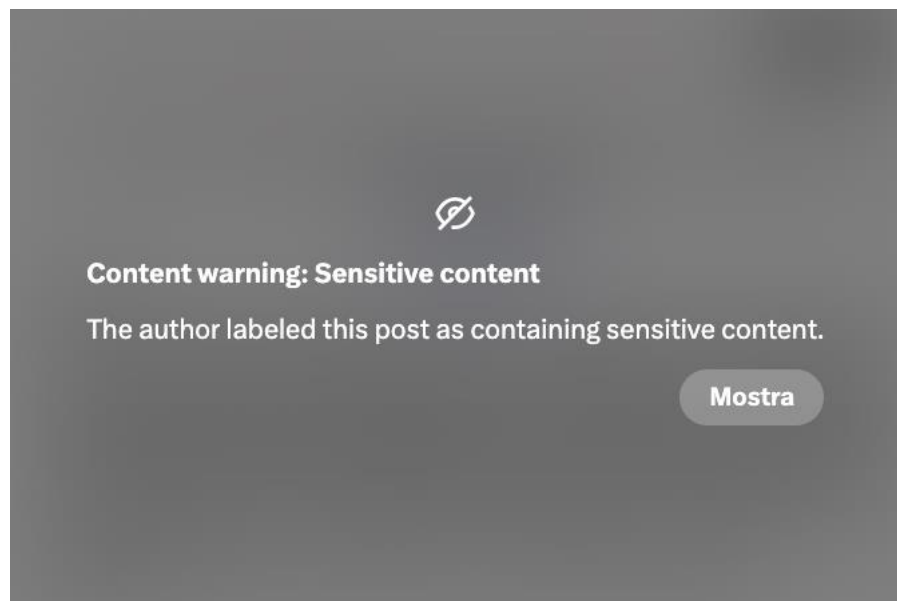


Figura 8. Schermata di X che avverte l'utente della presenza di un contenuto sensibile

Meta misura la limitazione sull'entità del contenuto e prevede che una violazione ripetuta nel tempo comporti sospensione (per un periodo limitato), disabilitazione o chiusura dell'account nell'art. 4.2 delle sue Condizioni d'uso. Inoltre, l'azienda ha prodotto la “Normativa sull'uso improprio” (*Misuse Policy*)³⁷ per chiarire la sua posizione.

TikTok prevede la sospensione o disattivazione definitiva dell'account in caso di violazione delle *policy* o della Proprietà Intellettuale, ovvero in mancato soddisfacimento dei requisiti di permanenza, quali l'età minima, l'appropriazione di identità altrui, la diffusione di “contenuti illegali” e le violazioni multiple e frequenti dei regolamenti.

2.3.4 Sanzioni da parte dell'AGCOM

Nei primi sei mesi dalla designazione dell'Autorità per le Garanzie nelle Comunicazioni (AGCOM) come DSC in Italia, sono state avviate indagini su tutte e tre le Corporation digitali.

³⁷ Normativa sull'uso improprio (Transparency Center, 16 aprile 2024), <https://transparency.meta.com/it-it/enforcement/taking-action/misuse-policy/> (27 aprile 2024).

Il 20 dicembre 2023 Meta è stata sanzionata per 5.850.000 euro per aver violato il divieto di pubblicità del gioco d'azzardo ai sensi dell'articolo 9 del decreto-legge del 12 luglio 2018, n. 87 (c.d. "Decreto Dignità")³⁸. L'Autorità ha accertato la presenza di contenuti di promozione o pubblicità su 18 account (5 su Instagram e 13 su Facebook) e 32 contenuti sponsorizzati. Meta, già sanzionata nel gennaio 2023, è stata ritenuta responsabile dei contenuti a pagamento diffusi sulle sue piattaforme di social media, che hanno ospitato i contenuti caricati dagli utenti e hanno offerto loro un servizio pubblicitario, altrimenti evitabile previo controllo. Dei 18 account incriminati, Meta è stata ritenuta responsabile soltanto per 5 di essi perché, a seguito delle dovute segnalazioni, ha provveduto a rimuovere i profili.

Sulla tutela dei minori, TikTok ha rimosso i video relativi alla "cicatrice francese", una sfida diffusa sulla sua piattaforma che ha coinvolto migliaia di giovani in pratiche di autolesionismo per procurarsi ematomi sugli zigomi. Si tratta del primo provvedimento dell'AGCOM in applicazione del Testo Unico dei Servizi Media Audiovisivi - TUSMA atto a limitare la circolazione di programmi, video e pubblicità audiovisive su una piattaforma di condivisione video (VSP) diretti al pubblico italiano, se questi vengono ritenuti nocivi per lo sviluppo psicofisico dei minori, incitano all'odio o rappresentano una minaccia³⁹.

Infine, l'Autorità ha comminato una sanzione di 1.350.000 euro a X per nove violazioni del divieto di pubblicità del gioco d'azzardo su account con "Spunta blu" che hanno condiviso contenuti aventi natura di comunicazione pubblicitaria. Soltanto 7 account sono stati oscurati direttamente dalla piattaforma e ai restanti due profili ha impartito un ordine di inibizione all'accesso⁴⁰.

³⁸ Comunicato stampa di AGCOM (22 dicembre 2023), <https://www.agcom.it/sites/default/files/migration/article/Comunicato%20stampa%202022-12-2023.pdf> (27 aprile 2024).

³⁹ Comunicato stampa di AGCOM (16 febbraio 2024), <https://www.agcom.it/comunicazione/comunicati-stampa/comunicato-stampa-16-febbraio-2024> (30 dicembre 2024).

⁴⁰ Comunicato stampa di AGCOM (12 marzo 2024), <https://www.agcom.it/comunicazione/comunicati-stampa/comunicato-stampa-12-marzo-2024> (30 dicembre 2024).

Capitolo 3

Il governo dei corpi intermedi

*«Ricorda che queste regole sono le stesse
di un sistema di elaborazione dati:
alcune possono essere eluse, altre infrante»*
The Matrix

3.1 Il diritto d'appello online

L'esercizio della sovranità, nella sua dimensione interna, si inverte nel monopolio della forza in un dato territorio e nel governo dei corpi intermedi. Mentre nella tradizione di *common law* il potere giudiziario era considerato parte del potere esecutivo, Montesquieu differenziava gli organi a cui affidare le funzioni legislativa, esecutiva e giudiziaria, indipendenti tra loro, per evitare che la libertà potesse essere minacciata. La separazione dei poteri⁴¹ e la tensione tra *iurisdictio* (l'insieme dei diritti, pregi, principi) e *gubernaculum* (il potere politico) è il cuore di ogni regime democratico (Rullo, L. in Calise, M. & al., 2021, p. 134).

Le corti oscillano tra il rendere giustizia, per rispondere alle necessità di una logica sociale fondata sulla pacifica convivenza tra gli individui, e il dichiarare legge, ovvero di applicarla e interpretarla. La condizione di indipendenza è stata soddisfatta con la creazione di «una struttura della società politica organizzata tramite e mediante la legge, allo scopo di limitare l'arbitrarietà del potere di sottometterlo al diritto» (Sartori, 1995 [1987], p. 23). Tale potere è stato ulteriormente frammentato dall'emergere di nuovi organismi come le autorità amministrative indipendenti, quei soggetti o enti pubblici tenuti a esercitare con imparzialità funzioni in ambiti sensibili o ad alto contenuto tecnico. Fungono da riferimento le *Independent Agencies* statunitensi con autorità di regolamentazione, affermatesi con il *New Deal*, e le francesi *Autorités Administratives Indépendantes* (AAI), che agiscono per conto dello Stato dal 1970. Tra queste, la prima fu istituita con la legge n°17/1978 sul trattamento dei dati, l'archiviazione e le libertà individuali (*Loi Informatique et Libertés*), che predispose la creazione della commissione CNIL (*Commission Nationale de l'Informatique et des Libertés*) per permetterne l'applicazione. Entrambe le tipologie sono state concepite come isolate dal controllo e

⁴¹ Art. 16 della Dichiarazione dei diritti dell'uomo e del cittadino (1789): «Ogni società nella quale la garanzia dei diritti non è assicurata e la separazione dei poteri non è determinata non ha una costituzione».

dall'autorità del governo. Il processo di integrazione europea ha rappresentato un nuovo impulso verso l'introduzione delle autorità indipendenti negli ordinamenti con specifici incarichi di regolazione.

L'art. 21 del DSA si concentra sugli organismi di risoluzione extragiudiziale delle controversie, cioè quelle strutture o procedure progettate per risolvere dispute legali al di fuori del sistema giudiziario tradizionale. La disposizione si rivolge ai destinatari dei servizi, annoverando persone o enti che hanno presentato segnalazioni con diritto di scegliere qualsiasi organismo per contestare le decisioni, in combinato disposto con l'art. 20, inclusi i sistemi interni di gestione dei reclami. La norma fa riferimento a procedure del diritto privato conosciute come Metodi Alternativi di Risoluzione delle Controversie (MARC) o *Alternative Dispute Resolution* (ADR), di cui le più comuni sono la negoziazione, la mediazione civile, la conciliazione, l'arbitrato e il giudizio privato (Direttiva 2013/11/UE)⁴². Uno dei primi casi nell'*e-commerce* fu il servizio di *Online Mediation* di eBay e in seguito l'*Online Schlichter* in Francia e Germania; differente e meritevole di menzione è il *British Columbia Civil Resolution Tribunal* (CRT) del Canada che combina l'espedito tecnologico di mediazione elettronica al sistema giudiziario nazionale vigente (Longo E. , 2023, p. 218-219). La risoluzione dei conflitti di consumo senza ricorso al giudice ("il processo fuori dal processo", *ibidem*) è pensata per essere più rapida ed efficace e per evitare che le persone rinuncino alla tutela dei diritti per ragioni di tempo e costo: il reclamo viene presentato presso un organismo – pubblico o privato, istituito su base permanente - sul quale vigila un'autorità competente, che nell'ambito delle comunicazioni elettroniche e postali italiane è l'AGCOM (2015).

Gli elementi caratterizzanti degli organismi certificati sono l'indipendenza e l'imparzialità, anche sul piano finanziario, dalle piattaforme e dagli utenti, liberamente accessibili in Rete e in grado di fornire un servizio rapido (entro 90 giorni dal ricevimento del reclamo, prorogabili una sola volta per controversie complesse), efficiente (nei costi) secondo regole procedurali chiare ed eque, il cui parere, tuttavia, non è vincolante per le parti (art. 21, lett. 2 DSA). Il certificato viene rilasciato dal Coordinatore dei Servizi Digitali (DSC) per un periodo massimo di cinque anni, rinnovabile e revocabile, specificando le questioni concrete sotto la sua competenza e le lingue in cui l'organismo può esprimersi, con l'obbligo di riferire annualmente il proprio operato al fine di produrre una relazione biennale complessiva sul funzionamento. Il fornitore dei servizi è tenuto al rimborso dei diritti e delle spese sostenute in relazione alla risoluzione

⁴² Dal 2013, la quota di *e-commerce* è passata dal 2% al 4% del PIL dell'UE, aumentando la sua rilevanza per i consumatori. Circa 300.000 controversie sono esaminate da ADR ogni anno, con tassi di risoluzione tra il 17% e il 100% a seconda degli Stati membri. EP Think Tank (29 novembre 2024). Briefing su "*Alternative Dispute Resolution*", [www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2024\)757639](http://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2024)757639)

della controversia in caso di torto, viceversa il destinatario non è tenuto al rimborso a meno che non sia stata accertata la mala fede.

Gli artt. 20 e 53 del DSA disciplinano il diritto a presentare un reclamo. All'utente deve essere fornita la possibilità di accedere a un sistema efficiente di gestione dei reclami contro le decisioni prese all'atto della segnalazione o della notifica sui contenuti considerati illegali o incompatibili con le condizioni generali, tra cui si evidenziano decisioni sulla rimozione di informazioni, impedimento all'accesso o limitazione della visibilità, sospensione o cessazione della prestazione o degli account, finanche della possibilità di monetizzazione. Anche se gestito in autonomia dalle piattaforme, il sistema interno deve contemplare le seguenti funzionalità:

- libertà di accesso;
- per un periodo di almeno sei mesi;
- presentazione per via elettronica e gratuita.

La gestione dei reclami deve essere operata in maniera tempestiva, non discriminatoria, diligente e non arbitraria, in grado di ritirare, senza indebito ritardo, la decisione presa qualora si verificassero le condizioni di infondatezza o errore materiale e di darne immediata comunicazione al reclamante. L'art. 20 si completa con l'obbligo a non utilizzare esclusivamente sistemi di IA nell'ambito di applicazione:

«I fornitori di piattaforme online provvedono affinché le decisioni [...] siano prese con la supervisione di personale adeguatamente qualificato e non avvalendosi esclusivamente di strumenti automatizzati» (art. 20, lett. 6, DSA).

In prima o seconda istanza, i destinatari del servizio (siano essi enti, organizzazioni o associazioni) possono presentare un reclamo in caso di violazione del regolamento direttamente al DSC dello Stato membro in cui è residente o domiciliato, il quale, a seguito di una valutazione del caso, esprime un parere o trasmette a un'altra autorità competente il reclamo (art. 53 DSA).

Il *Digital Services Act* era stato preceduto alcune leggi nazionali che avevano reso meno uniforme l'applicazione delle regole aziendali. Il *Communication Platform Act* in Austria e il *Network Enforcement Act* (NetzDg o anche "*Facebook Act*") in Germania già prevedevano una procedura *ad hoc* per i reclami basata su tre casistiche (Santoriello, 2022, p. 298):

- la corrispondenza tra reclamo, Linee guida della community e disposizioni di legge comporta la rimozione a livello globale e la conclusione del processo di analisi;
- la mancata violazione delle Linee guida, a cui fa seguito l'analisi di legittimità in base alla segnalazione;
- la violazione del Codice penale nazionale, che provoca interdizione dall'accesso soltanto in quello Stato.

Il NetzDg, in vigore dal 2018, prevede l'obbligo di rimozione dei contenuti illegali definiti nel Codice penale (come insulti, minacce, violenza), pena una sanzione massima di 50 milioni di euro, per le piattaforme con oltre 2 milioni di utenti. La norma è stata criticata da giornalisti, attivisti per i Diritti Umani e accademici perché percepita come un incentivo alla censura proattiva della libertà di espressione, legittimando il potere delle piattaforme di articolare il dibattito nazionale, mentre fu considerata un modello da Russia, Singapore e Filippine. Human Rights Watch (2018) sosteneva che anche per i tribunali sarebbe stato difficile determinare la liceità dei contenuti, valutando nel dettaglio le sfumature culturali e legislative, ma proprio in presenza di revisioni rapide (24 ore per contenuti palesemente illegali e fino a una settimana per gli altri, con possibilità di esternalizzare la decisione previa autorizzazione del governo) e del rischio di multe, le aziende private non avrebbero salvaguardato la libertà di espressione. Pur prevedendo un meccanismo lineare dal ricevimento del reclamo alla trasmissione della giustificazione, la disciplina non imponeva un processo di appello giudiziario, altrimenti reso da Facebook, YouTube e Twitter (*ibidem*).

Attraverso la piattaforma europea per la risoluzione extragiudiziale delle controversie online (*Online Dispute Resolution – ODR*)⁴³ – disponibile in UE, Norvegia, Islanda e Liechtenstein – il consumatore ha facoltà di dialogare con un professionista e inviare allegati o programmare una riunione⁴⁴. L'accordo può essere raggiunto entro un massimo di 90 giorni o ricorrere ad altri organismi di risoluzione delle controversie autorizzati entro 30 giorni. Le VLOP sono tenute a inserire nel proprio *Transparency Report* il numero di controversie sottoposte agli organismi di risoluzione extragiudiziale, il tempo mediano per portare a termine le procedure e il numero di sospensioni imposte secondo quanto stabilito dall'art. 24 del DSA.

I *Transparency Report 2024* contengono i dati aggiornati sull'*Automation Overturn Rate* (AOR), l'indicatore di accuratezza della MAC (artt. 15 e 42 DSA), ovvero «la percentuale di contenuti sottoposti a intervento automatico che vengono successivamente ripristinati» (*Meta Transparency Report*). ByteDance, che riporta la distribuzione per ciascuno Stato UE, afferma che l'indicatore di appropriatezza è «la percentuale di video e annunci per i quali la decisione di *enforcement* originale è stata confermata o mantenuta» e di errore «la percentuale di contenuti video e pubblicitari per i quali la decisione di applicazione originale è stata annullata» (*Annex B, TikTok Transparency Report*).

⁴³ Online Dispute Resolution: <https://ec.europa.eu/consumers/odr/main/>

⁴⁴ La Commissione ha proposto di eliminare la piattaforma ODR, attiva dal 2016, in quanto si è occupata, in media, di soli 200 casi all'anno, non giustificando i costi di mantenimento. La decisione è stata sostenuta dal Parlamento e dal Consiglio. EP Think Tank (29 novembre 2024). Briefing su “*Alternative Dispute Resolution*”, [www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2024\)757639](http://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2024)757639)

Tabella 4. *Automation Overturn Rate* su Meta, X e ByteDance

	Meta	X	ByteDance
AOR	8,4 %	14,92 %	0,9 % (media)

Fonte: Elaborazione a cura dell'autrice sui dati dei Transparency Report al 06/2024

I sistemi interni di gestione dei reclami sono tra loro molto diversi e rispecchiano la strategia di ciascuna Corporation (Appendice A) per affrontare la complessità digitale. Condividono, tuttavia, la pretesa di poter delegare a terzi la decisione finale sulle questioni più scottanti riguardanti la moderazione dei contenuti, senza mai uscire dalla piattaforma digitale, al culmine di un processo decisionale altamente centralizzato e sproporzionato a favore dell'azienda.

3.2 L'*Oversight Board* di Meta (MOB)

Le piattaforme amministrate da Meta prevedono una procedura di ricorso che concede all'utente la possibilità di fare appello qualora i sistemi automatizzati o i revisori umani dei contenuti commettano errori di valutazione. L'idea di introdurre una sorta di "Corte suprema" risale al 2018⁴⁵, quando si presentò l'urgenza di un'istituzione dotata del potere di difendere la libertà di espressione e di indagare e valutare gli episodi di censura (Klonick, 2020). Dopo due anni di consultazioni, Mark Zuckerberg annunciò la nascita dell'*Oversight Board* (MOB), un organismo indipendente il cui compito sarebbe stato quello di rivedere la moderazione dei contenuti, fornire interpretazioni di *policy* ed esaminarne l'applicazione, concentrandosi su alcuni casi di rilevanza globale (Klonick, 2021). Fin dai primi mesi di attività, il Comitato per il controllo ha esaminato casi, condotto indagini e aperto consultazioni pubbliche per raccogliere le prospettive di varie parti interessate prima di pubblicare le sue decisioni vincolanti per l'azienda. Secondo Caplan (2023), il MOB è un esempio di come la Corporation stia tentando di distribuire responsabilità verso organizzazioni della società civile, esperti, agenzie governative e comunità locali, creando una rete di interessati (*Network Governance*). È composto da un Consiglio che comprende accademici ed esperti di tecnologia, legge e Diritti Umani, e può variare da 11 a 40 membri (attualmente ne conta 21), molti dei quali provengono dagli Stati Uniti e nessuno dai Paesi in cui le relazioni con Meta sono maggiormente a rischio o interrotte, come Russia e Cina (Santoriello, 2022). Al Consiglio è stata conferita l'autorità di trasmettere raccomandazioni per migliorare le sue politiche e pratiche. Ogni membro dura in

⁴⁵ I *Santa Clara Principles* furono di impulso per migliorare il rapporto con la trasparenza e la responsabilità nella moderazione, esortando le piattaforme alla responsabilità (numeri-notifica-appello) con processi di revisione esterna indipendenti: santaclaraprinciples.org

carica tre anni e può essere rinnovato per un massimo di tre volte, pena la decadenza in caso di conflitto di interessi. Il MOB è un organismo indipendente (Tab. 5) supervisionato da cinque Garanti (*Trustee*) più il Presidente e il Direttore amministrativo, con competenze in ambito giuridico legate alla tutela della libertà di espressione, esercitate in accademia, organizzazioni pubbliche e private. La sua disciplina è regolata da un Atto costitutivo, dagli Statuti e dal Codice di condotta.

Tabella 5. Governance dell'Oversight Board di Meta e mansioni

Governance del MOB	Mansioni
Comitato	Selezione ed esame dei casi Emanare decisioni Proporre nuovi membri
Trustee	Gestione e adesione del Comitato ai valori Doveri fiduciari
Facebook	Supervisione indipendente e attuazione Ripartizione delle risorse

Fonte: Atto costitutivo, Statuti e Codice di condotta

Sulla base delle segnalazioni ricevute, i membri del MOB eleggono un numero esiguo di casi «idonei, difficili, significativi e pertinenti a livello globale, in grado di arricchire le normative future» che necessitano di una revisione. L'organismo stabilisce i criteri di ammissibilità delle richieste, tra cui il livello di gravità e priorità necessari per la riapertura del caso, sulla base delle indicazioni contenuto nello Statuto e i 4 principali valori aziendali: autenticità, sicurezza, privacy e dignità (Santoriello, 2022).

Il ricorso è vincolato al soddisfacimento dei seguenti requisiti:

- 1) La contestazione proviene da un account attivo (il proprietario è in grado di accedervi);
- 2) L'utente, previa richiesta di revisione della decisione di moderazione, avvia la procedura entro i 15 giorni successivi all'aggiornamento della decisione;
- 3) Le decisioni devono essere idonee al ricorso.

A ciascun caso vengono assegnati un identificativo (ID) e una giuria, formata da almeno un membro proveniente dalla regione coinvolta, a cui verranno trasmesse le informazioni rilevanti ai fini della comprensione del contesto. La giuria può avvalersi del parere ulteriore di consulenti esperti del settore per deliberare una decisione vincolante nei successivi 90 giorni. Il MOB produce, infine, una motivazione articolata e disponibile online, notificando al ricorrente la pubblicazione e concordando con gli interessati quali informazioni rendere note per

autorizzare/impedire il riconoscimento dei soggetti coinvolti. La dichiarazione potrebbe contenere anche un parere consultivo sulle normative rivolta a Meta, che implementerà la decisione «purché tale applicazione non violi la legge» (Santoriello, 2022).

L'Oversight Board si pone come l'ultimo grado di appello per ottenere giustizia in caso di moderazione errata o mendace, prima di uscire dalla piattaforma per adire a tribunali "tradizionali". Secondo quanto stabilito dall'art. 21 del DSA, un organismo di risoluzione extragiudiziale delle controversie deve rispettare i requisiti di imparzialità e indipendenza, anche finanziaria, dai fornitori di piattaforme, di cui il MOB non dispone.

L'interpretazione degli Standard della Community, delle Linee Guida e delle Altre Normative di Meta è un'operazione complessa perché l'azienda modifica arbitrariamente le singole voci, le tecniche di moderazione che influiscono sulla quantità e sulla qualità dei contenuti in circolazione e il ritmo con cui il procedimento viene messo in atto (Grafico 1). Stando ai dati a disposizione, complessivamente gli Standard della Community di Facebook sono stati modificati in media 41 volte all'anno (7 volte per singola voce) dal 2018, con un aumento significativo concomitante con l'inizio delle attività del MOB (Tab. 6). Le principali revisioni (superiori al 6% del totale) hanno riguardato i campi: Violenza e istigazione alla violenza, Prodotti o servizi con restrizioni, Incitamento all'odio e Immagini forti e violente.

Tabella 6. *Variazioni degli Standard della Community dal 2018 al 2024*

Standard della Community	Modifiche (2018-2024)	% sul totale
Violenza e comportamenti criminali	96	32,88%
Sicurezza	75	25,68%
Contenuti deprecabili	75	25,68%
Integrità e autenticità	44	15,07%
Rispetto della proprietà intellettuale	0	0,00%
Richieste e decisioni relative ai contenuti	2	0,68%
	292	100%

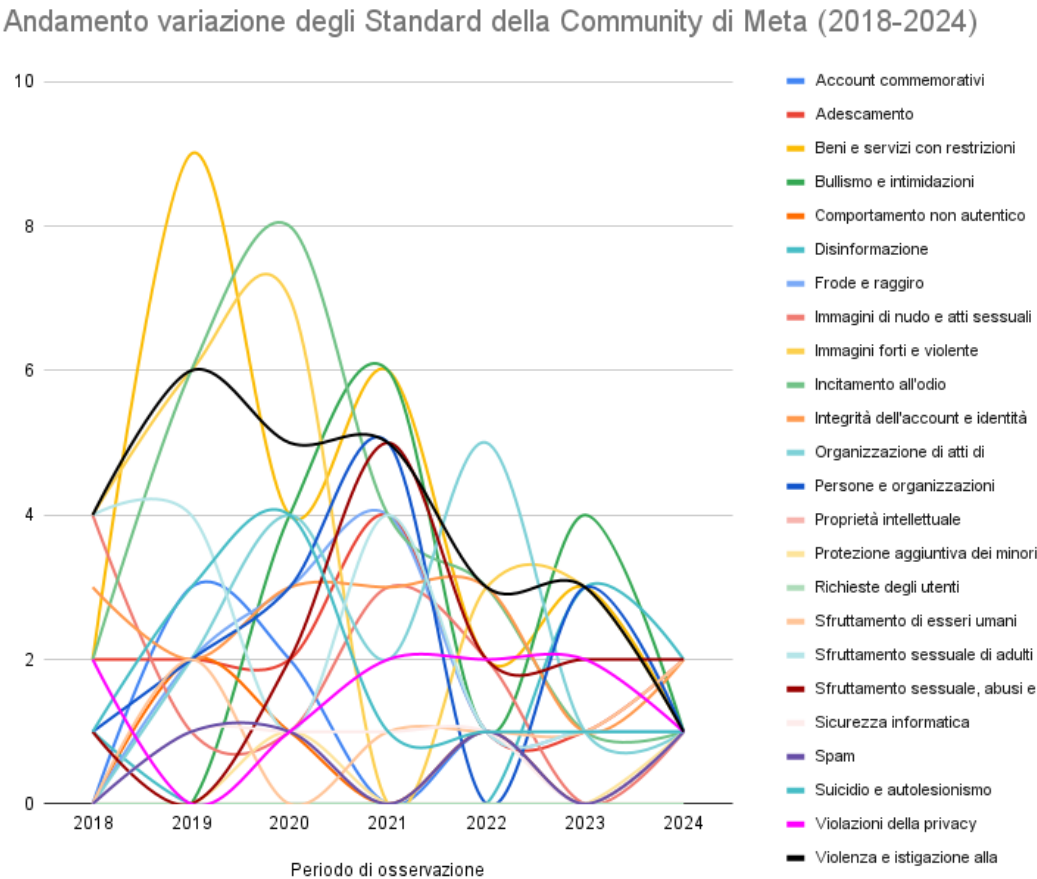
Fonte: Elaborazione a cura dell'autrice (agg. al 29 aprile 2024).

Tra il 2020 e il 2024, l'Oversight Board ha rivisto le precedenti scelte di moderazione 96 volte e ha emesso più di 100 raccomandazioni al fine di sollecitare la modifica delle “norme”. Nell’82% dei casi, il MOB ha ribaltato le decisioni originali degli algoritmi o dei revisori di Meta. Alcune di queste sono "decisioni sommarie" (*bundled*), approvate da un gruppo di membri del Consiglio, che forniscono trasparenza sulle correzioni di Meta e mettono in evidenza le aree di potenziale miglioramento nell'applicazione delle norme.

Nel 2022, il Comitato per il controllo (2023) ha ricevuto circa 1,3 milioni di richieste di appello da parte degli utenti per ripristinare o rimuovere contenuti su Facebook o Instagram, con un aumento di circa un quarto rispetto al 2021. Durante l'intero periodo 2020-2024, più di due terzi dei casi selezionati riguardavano post rimossi in ambiti come "Incitamento all'odio" (30), "Persone e organizzazioni pericolose" (25) e "Violenza e istigazione alla violenza" (19).

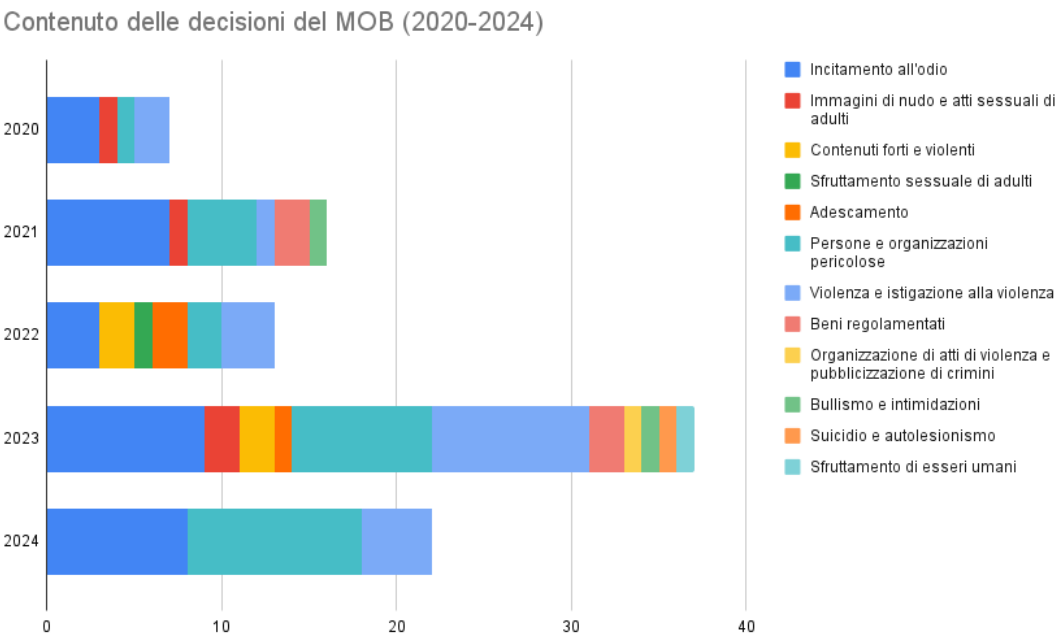
Confrontando i dati raccolti e presentati nei Grafici 1 e 2 è possibile osservare che, mentre le voci "Violenza e istigazione alla violenza" (9,25%) e "Incitamento all'odio" (8,56%) coincidono con un'elevata percentuale di modifiche agli Standard della Community effettuate dal 2018 a oggi, la voce "Persone e organizzazioni pericolose" (5,14%) mantiene una percentuale mediamente simile sia sul Capitolo che sul totale. Dunque, nonostante il MOB abbia riscontrato problematicità inerenti alla moderazione della libertà di espressione in quel campo (che corrisponde al secondo contenuto per numero di decisioni pubblicate, pari al 26%), l'azienda non ha apportato modifiche ai suoi Standard della Community con i medesimi ritmi riscontrati per le altre voci. Nei primi mesi del 2024, 5 delle 22 decisioni pubblicate hanno riguardato il conflitto in atto in Palestina. Come si vedrà nel Capitolo 4, i dati celano implicazioni politiche più profonde.

Grafico 1. *Andamento variazione degli Standard della Community di Meta (2018-2024)*



Fonte: Elaborazione a cura dell'autrice (agg. al 29 aprile 2024)

Grafico 2. *Contenuto delle decisioni dell'Oversight Board di Meta (2020-2024)*



Fonte: Elaborazione a cura dell'autrice (agg. al 29 aprile 2024)

3.3 Presentare o ricevere una segnalazione su X

Twitter è stata a un passo dal lanciare il *Twitter Moderation Research Consortium* nel 2022, poco prima di essere acquisita da Elon Musk. Al momento, un iscritto alla piattaforma X può segnalare un post, una lista, un profilo, un messaggio diretto, un momento, uno spazio X o un prodotto. Il procedimento è eterodiretto: si seleziona la spunta “Segnala” e si sceglie il tipo di contenuto da segnalare. Se la scelta ricade sulla categoria “È offensivo o dannoso”, la piattaforma chiederà un approfondimento con informazioni aggiuntive o ulteriori pubblicazioni provenienti dall’account segnalato per prove di contesto. Il contenuto viene poi trasmesso nelle notifiche e nelle e-mail di aggiornamento (da cui sarà possibile fare *opt out*), in linea con le norme in materia di informazione tra piattaforma e utente. A seguito del ricevimento della segnalazione, l’azienda fornisce “consigli su altre azioni” che possono essere intraprese per migliorare l’esperienza su X. Tipologie specifiche di violazione possono essere segnalate direttamente attraverso il Centro assistenza: a) uso non autorizzato del marchio; b) uso non autorizzato del materiale coperto da copyright; c) vendita o promozione di merci contraffatte; d) informativa sulla privacy nei confronti dei bambini; e) sfruttamento sessuale minorile; f) pornografia; g) impersonificazione di un individuo o brand; h) informazioni private pubblicate su Twitter; i) comportamento offensivo e minacce di violenza; l) spam e abuso del sistema; m) violazione delle norme di X Ads⁴⁶.

In presenza di contenuti sensibili, X analizza le segnalazioni per determinare se il contenuto deve essere corredato da un messaggio di avviso senza che ciò comporti automaticamente una restrizione o una rimozione del contenuto. Nella sezione “Privacy e sicurezza”, l’esperienza può essere decentralizzata dal punto di vista dell’utente che, accedendo a “Contenuti che vedi”, può decidere di spuntare la casella “Mostra contenuti potenzialmente sensibili”.

Il Centro assistenza fornisce delucidazioni in merito alle “richieste legali” che provengono da forze dell’ordine, autorità pubbliche, avvocati in un processo penale o parte di una controversia civile in Europa (secondo quanto stabilito dagli artt. 9-10 DSA). Le richieste possono essere di due tipi: di informazioni specifiche, quando sono correlate a indagini penali o a cause civili (divorzio o dispute commerciali); di rimozione, quando si sospetta che il contenuto violi le norme in uno o più Stati da cui proviene la segnalazione (come nei casi di diffamazione, attività illegali o sicurezza nazionale).

L’esame preliminare avviene mediante il controllo delle prescrizioni normative: se la richiesta soddisfa i requisiti, un agente di X la esamina per verificare che non limiti o soffochi la libertà

⁴⁶ Report violations (Platform Use Guidelines | Help Center), <https://help.twitter.com/en/rules-and-policies/x-report-violation#violations> (29 aprile 2024).

di espressione, non violi le Regole di X (per esempio, account di giornalisti, abbonati/verificati o comunicazioni di natura politica) e non riscontri problematiche tecniche (nel caso in cui l'account risultasse non disponibile). L'azienda si riserva di rifiutare la richiesta se sussistono circostanze quali vaghezza, inconsistenza, assenza di firma o vizi di forma nel procedimento di presentazione della richiesta («Cerchiamo sempre di avvertire gli utenti dell'esistenza di richieste legali che riguardano i loro account, ma qualche volta la legge ci proibisce di farlo»), come nel caso in cui le richieste siano generiche o se la natura dell'indagine non appaia chiara, ma solitamente il richiedente viene indirizzato verso procedure indicate da un accordo di mutua assistenza legale (MLAT) o da una rogatoria scritta.

La piattaforma comunica all'utente l'esistenza della richiesta legale, salvo che la disciplina non rientri in emergenze quali il pericolo di vita, lo sfruttamento sessuale dei minori o il terrorismo. Il contenuto sarà rimosso integralmente se in conflitto con i Termini di Servizio e/o le Regole di X o ristretto per area geografica se considerato illegale in una specifica località. La piattaforma preferisce lasciare, quando possibile, che sia l'utente a decidere se eliminare il contenuto, rispondere in prima persona a chi ha inviato la richiesta o cercare assistenza legale per contestarla, prima che sia la piattaforma a intervenire sul contenuto:

«La nostra notifica non rappresenta un consiglio legale che invita alla rimozione di determinati contenuti. Viene inviata solo per informarti che un certo contenuto presente sul tuo account X è stato identificato in una richiesta legale, in modo che tu possa decidere se e come rispondere. Inoltre, serve a renderti consapevole del fatto che, a seguito di tale richiesta, potremmo dover intraprendere un'azione sul contenuto segnalato» (Centro assistenza di X).

Stando alle Regole di X, la contestazione della decisione di oscurare un contenuto può essere presentata rispondendo all'e-mail inviata per notificare l'azione intrapresa (se avvenuta a seguito di una richiesta legale) o contattando il Centro assistenza (se scaturita da leggi locali). Il contenuto viene formalmente ripristinato se il ricorso di X o dell'utente viene accolto in appello, se il periodo di validità del provvedimento giudiziario è terminato o se un'autorità giudiziaria stabilisce che la richiesta presentata da un'autorità amministrativa non era valida. Nell'ultimo report precedente all'acquisizione⁴⁷, i dati evidenziano come i paesi con il maggior numero di richieste fossero India, Francia, Stati Uniti, Giappone e Germania.

Fare ricorso contro una decisione sui “contenuti illegali” è una pratica prevista su X grazie al DSA (artt. 17, 20-21). Un'apposita sezione prevede la compilazione di un modulo online accessibile gratuitamente in cui è possibile selezionare il motivo del ricorso contro una

⁴⁷ Information Requests (Transparency | X), <https://transparency.x.com/en/reports/information-requests.html#2021-jul-dec> (29 aprile 2024).

decisione se presa su segnalazione presentata o per il contenuto pubblicato. Infine, un box a risposta aperta chiede all'utente di argomentare le ragioni dell'infondatezza.

Secondo X, l'errore può dipendere da tre tipologie di decisione⁴⁸ di:

- a) rimuovere informazioni, disabilitarne l'accesso o limitarne la visibilità;
- b) sospendere o disabilitare, integralmente o parzialmente, account;
- c) sospendere, terminare o limitare la monetizzazione.

Ciononostante, si ricorda che X non è vincolata alle decisioni prese dagli organismi certificati per la risoluzione extragiudiziale delle controversie. Il reclamo può essere, altresì, presentato presso un tribunale in uno Stato membro dell'UE o ricorrere ad altre opzioni legali. Tra ottobre e aprile 2024 nessun reclamo è stato documentato (*X Transparency Report 2024*).

3.4 Fare appello su TikTok

ByteDance designa otto pilatri nel rispetto della sicurezza e dei Diritti Umani, basati su: bilanciamento tra espressione creativa e prevenzione del danno, sostegno alla dignità umana e garanzia di azioni eque⁴⁹. Queste considerazioni scaturiscono da quadri giuridici internazionali, tra cui si menzionano i Principi guida delle Nazioni Unite su imprese e Diritti Umani, la Dichiarazione universale dei Diritti Umani, la Convenzione internazionale sui diritti dell'infanzia e i Principi di Santa Clara. I *Community Principles* godono del coinvolgimento degli utenti, di profili professionali (esperti di sicurezza e salute pubblica) e dei Consigli consultivi (*Advisory Councils*), mediante cui l'azienda riesce a "espandere la sua presenza regionale". I Consigli consultivi sulla sicurezza e i contenuti riuniscono circa 50 esperti indipendenti a livello internazionale, le cui competenze si esplicitano in materia di giovani, libertà di espressione e incitamento all'odio:

«Concentrarci sulle regioni ci permette di comprendere meglio le sfide e le tendenze locali e di sviluppare soluzioni informate che tengano conto del contesto e delle culture locali. Questo impegno per la localizzazione ci permette di creare un approccio più mirato e reattivo alla sicurezza e ci consente di rimanere aggiornati sugli ultimi sviluppi in ogni regione» (TikTok, Coinvolgimento dei consigli consultivi⁵⁰).

La risoluzione delle controversie trova su TikTok applicazione in linea con la strategia regionale già descritta nel Capitolo 2. La piattaforma chiarisce che «i termini sono regolati dalle

⁴⁸ Digital Services Act (Country-specific resources | Help Center), <https://help.twitter.com/it/rules-and-policies/digital-services-act> (29 aprile 2024).

⁴⁹ Community Principles (TikTok), <https://www.tiktok.com/community-guidelines/it/community-principles> (26 aprile 2024).

⁵⁰ Engaging our Advisory Councils (Transparency Center), <https://www.tiktok.com/transparency/it-it/advisory-councils/> (26 aprile 2024).

leggi del Paese in cui risiede l'utente»⁵¹ e in prima istanza l'operazione sarà svolta al fine di giungere a una risoluzione in via amichevole.

Qualora le parti non raggiungano un accordo, entrambi potranno adire a tribunali locali (data la residenza dell'utente) o ai tribunali in cui sono presenti le società associate TikTok Technology Limited (Repubblica d'Irlanda), TikTok Information Technologies UK Limited (Inghilterra e Galles) con giurisdizione non esclusiva. I residenti nello Spazio Economico Europeo possono ricorrere a un ulteriore organismo attraverso la piattaforma europea di risoluzione delle controversie online a tutela dei consumatori.

I video caricati su TikTok vengono inizialmente esaminati da una tecnologia di MAC, che mira a identificare i contenuti in violazione delle *Community Guidelines*, esaminando una serie di segnali nei contenuti, come parole chiave, immagini, titoli, descrizioni e audio, e procede con la rimozione del contenuto quando le violazioni sono più evidenti (per esempio, nudità o sicurezza dei giovani). Quando i contenuti entrano “in fase di controllo”, vengono trasmessi al dipartimento *Trust & Safety* per determinare se necessitano di rimozione o oscuramento (non idonei al *feed* Per te). Ciò accade ogni volta che vengono caricati contenuti, se guadagnano popolarità o vengono segnalati. Il motivo della decisione può essere consultato ed è possibile inviare un riscontro, qualora l'utente attivi le metriche negli “Strumenti per i creator” della propria sezione “Menù”, interagendo con la scheda “Prestazioni”.

Al contempo, dal punto di vista del segnalante, TikTok chiede di selezionare il Paese in cui si ritiene che il contenuto sia illegale, i motivi per cui dovrebbe essere attenzionato e i dettagli sulla presunta legge violata (non vincolante ai fini della notifica). Nel caso in cui la violazione riguardi le *policy* della piattaforma, il contenuto viene rimosso a livello globale.

Se gli utenti ritengono che sia stato commesso un errore, possono impugnare la rimozione e l'oscuramento del loro contenuto (*Our commitments*, 2023) o contribuire con un *feedback* a migliorare la moderazione dei contenuti della piattaforma. Per presentare ricorso, l'utente seleziona “Altre opzioni” e “Metriche” o “Maggiori informazioni” e “Prestazioni” selezionando l'opzione “Ricorso”. L'invio di riscontri non preclude la possibilità di presentare un ricorso. Qualora il ricorso vada a buon fine, il video verrà ripristinato nei suggerimenti e sarà più semplice da trovare.

Coloro che si trovano in UE o nello Spazio Economico Europeo (SEE) possono presentare richieste di appello (artt. 17, 20-21 DSA) per impugnare le decisioni della piattaforma entro 180 giorni dalla data di notifica⁵². Le opzioni di ricorso saranno rese disponibili in un modulo

⁵¹ Terms of Service (TikTok), <https://www.tiktok.com/legal/page/eea/terms-of-service/it> (26 aprile 2024)

⁵² 113 giorni a livello globale e 23 giorni negli USA se l'account è stato bloccato. Underage appeals (TikTok), <https://support.tiktok.com/it/safety-hc/account-and-user-safety/underage-appeals-on-tiktok> (29 aprile 2024).

incluso nella notifica, dove l'utente troverà una casella di testo in cui inserire la motivazione dell'appello. Per ogni singola decisione l'utente può presentare ricorso una sola volta, includendo tutte le informazioni pertinenti a comprendere le ragioni del disaccordo (salvo che queste non contengano un linguaggio offensivo o inappropriato, per il quale TikTok può decidere di rifiutarlo). La motivazione che viene trasmessa in prima istanza non potrà essere successivamente modificata e dovrà essere volta a dimostrare la liceità del contenuto.

Se la decisione non può essere revocata, i ricorsi vengono resi indisponibili (trascorsi 180 giorni, in caso di divieto temporaneo o di decisione con pertinenza specifica, come per le Storie o le Dirette). Le persone che abusano del procedimento di ricorso, quando chiaramente infondato, potranno essere sospese dalla possibilità di presentarne ulteriori per un periodo di tempo. Se il ricorso ha esito positivo, l'azione intrapresa viene annullata⁵³.

Specifiche modalità di ricorso vengono contemplate per la fattispecie della minore età (per coloro che hanno meno di 13 anni, o meno di 14 anni in Indonesia, Quebec e Corea del Sud). Nel caso in cui il profilo venga bloccato, l'utente deve fornire la foto di un documento di identità, un *selfie* in cui abbia in mano il medesimo documento e un foglio in cui sia leggibile il codice associato al ricorso. Per gli utenti con età compresa tra i 13 e i 17 anni, il ricorso deve essere presentato mediante una foto che includa l'utente proprietario dell'account e un genitore o un tutore con età superiore ai 25 anni (o un "adulto fidato", identificato come un parente o un insegnante). L'adulto raffigurato deve avere tra le mani un documento su cui siano inseriti: a) Verifica dell'età di TikTok; b) data di nascita dell'utente; c) codice univoco associato al ricorso. A partire dai 18 anni, l'utente può procedere all'autorizzazione della carta di credito in autonomia o alla stima dell'età facciale. In questo secondo caso, il ricorso viene presentato attraverso un *selfie*, che verrà inviato a un fornitore esterno per l'accertamento biometrico in base alle caratteristiche facciali.

Le attività di moderazione sono le stesse per gli account appartenenti a entità governative, personaggi o partiti politici ma, a causa dell'interesse pubblico e delle dinamiche civili in cui sono coinvolti, le azioni di restrizione sugli account sono diverse. Se raggiungono il limite di avvertimenti previsto dalle Linee guida, scompariranno temporaneamente dai *feed* Per te e Seguiti per 90 giorni. Le restrizioni potrebbero aumentare in concomitanza con disordini civili, elezioni o ambienti sociali ad alto rischio, se l'account promuove violenza, odio o disinformazione, per i quali la piattaforma impedisce la pubblicazione dei contenuti per un

⁵³ Appeals and other redress possibilities (TikTok), <https://www.tiktok.com/legal/page/global/compliant-handling-eea/en> (29 aprile 2024).

periodo che varia dai 7 ai 30 giorni. TikTok non esclude la possibilità di intraprendere azioni legali al di fuori della propria piattaforma.

3.5 La dura legge dell'interoperabilità

Sul tema della concorrenza si manifestano le principali differenze tra USA e UE, da cui originano gli approcci in materia di interventi economici e *welfare* (Brown, 2020): mentre i primi si concentrano sul calmieramento dei prezzi (con ovvie contraddizioni per quei servizi *freemium* forniti dalle piattaforme), il quadro normativo europeo dedica maggiore attenzione a qualità, selezione e inclusione. L'importanza sociale di connettere tutti i membri della società è stata un perno della regolamentazione delle telecomunicazioni (già affrontato nel Capitolo 1 in riferimento alle *Essential Facilities*), caratterizzata da obblighi di interconnessione (consentendo la negoziazione dei prezzi) e interoperabilità ai fornitori di Rete - attraverso la direttiva quadro (2002/21/CE) e la direttiva sull'accesso (2002/19/CE) - e dalla garanzia di un servizio universale accessibile anche agli utenti con disabilità, tra gli obiettivi dichiarati anche dal *Digital Services Act* con esplicito rimando a “Codici di condotta per l'accessibilità” (art. 47) per migliorare l'accesso ai servizi online attraverso una progettazione o un successivo adattamento che risponda alle particolari esigenze delle persone.

L'interoperabilità consente di offrire agli utenti una maggiore possibilità di scelta anche tra diversi regimi di moderazione (Brown, 2020, p. 39). Mastodon, social network *open source* e decentralizzato, funziona come una federazione di server gestiti da privati, associazioni o collettivi – il suo caso è emblematico: ciascuna “istanza” può scegliere le proprie regole di moderazione sulla base di mere preferenze facilitate dai *software* disponibili, professandosi «senza algoritmi o pubblicità per farvi perdere tempo». Su questa piattaforma le comunità di interesse si formano attorno a specifiche istanze che hanno una elevata variabilità e la scelta a volte dipende dal rigore delle regole di moderazione. È stato osservato che la tipologia di utente attivo su Mastodon appartiene alla comunità LGBTQIA+ perché la sua composizione favorisce l'autonomia e la «struttura orizzontale sposta l'attenzione del sito dal numero di utenti alla qualità del coinvolgimento e alle comunità di nicchia» (*ivi*, p. 42). In parallelo, anche gruppi più inclini all'*hate speech* trovano spazio con propri server, come nel caso di Gab - piattaforma di *microblogging* affine all'estrema destra statunitense – a cui il *provider* originario ha negato il servizio di *hosting* in seguito alla sparatoria di Pittsburgh nel 2018.

BlueSky è un social network costruito sul Protocollo AT, una base decentralizzata per i social media, inizialmente sviluppato da Twitter. Più *provider* operano in ogni parte del sistema e gli utenti possono scegliere e cambiare quando preferiscono, controllare i contenuti e farlo in un

ambiente digitale semplice da utilizzare. Chiunque può creare il proprio *feed* su una selezione manuale, per *hashtag* o *emoji* e modificarlo, permettendo anche ad altri di aderirvi. La natura *open source* consente di contribuire alla moderazione dei contenuti e alla gestione della comunità (Kleppmann *et al.*, 2024), ispirando future ricerche sul tema dei beni comuni digitali (*Digital Commons*).

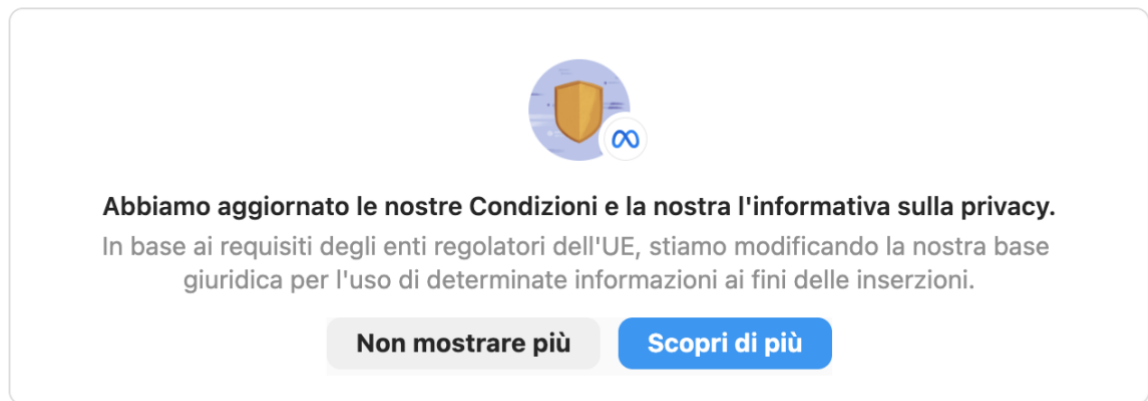


Figura 9. Disclaimer diffuso sulle piattaforme di Meta nel maggio 2023

Capitolo 4

Alienazione e identità nello spettro algoritmico

«*I want to do the imagining.
I don't want to be the idea*»
Barbie

4.1 *Code and control*. L'ipotesi di una standardizzazione della cittadinanza

Nella sua lezione inaugurale al *Collège de France* il 2 dicembre 1970, Michele Foucault avanzò un'ipotesi (2004 [1972]):

«Suppongo che in ogni società la produzione del discorso è insieme controllata, selezionata, organizzata e distribuita tramite un certo numero di procedure che hanno la funzione di scongiurare i poteri e i pericoli, di padroneggiarne l'evento aleatorio, di schivarne la pesante, temibile materialità».

Norme e pratiche regolano il funzionamento del discorso all'interno di una società, determinando chi può parlare, cosa può essere detto e come interpretarlo. Oltre a essere uno strumento di comunicazione, è un dispositivo estetico del potere che contribuisce alla costruzione della realtà, alla definizione della verità e alla sua influenza. Questa tendenza impone la delimitazione argomentativa su temi sensibili e la distinzione tra ciò che è razionale e irrazionale, escludendo o censurando discorsi che minacciano l'ordine sociale. La repressione, pertanto, si manifesta anche per mezzo del linguaggio e del sapere.

L'ordre du discours identifica tre procedure di controllo del discorso:

- Interdizione, a cui spettano divieti (es. tabù sociali);
- Esclusione, divisione (tra ragione e follia) e rifiuto, con conseguente delegittimazione;
- Opposizione del vero e del falso, ovvero la definizione dell'autorità e la creazione di gerarchie tra discorsi accettabili e inaccettabili.

Il discorso, per Foucault, non è mai neutro, anzi è inserito in rapporti di disciplina, aprendo la strada alle riflessioni sul concetto di governamentalità, ovvero il controllo politico che si manifesta attraverso la gestione della condotta delle persone.

La volontà di sapere ha le sue radici in una “partizione storica” (*ibidem*) che si è trasformata varie volte fino a diventare una nuova forma di volontà di verità: tra il XVI e il XVII secolo, in Inghilterra, si è sviluppato un approccio alla conoscenza che privilegia campi di studio e

tecniche possibili, misurabili e catalogabili e impone al soggetto di vedere e verificare, non coincidente con la volontà di astrazione sviluppatasi nella cultura classica.

Le procedure che ne derivano fungono da principi di classificazione e ordinamento, per padroneggiare la dimensione dell'evento e del caso, distribuite in discipline in cui coesistono verità ed errore, quest'ultimo con funzioni positive e non residuali, fino a formulare una proposizione. L'identità assume la forma di una «permanente riattualizzazione delle regole» (Foucault, 2004 [1972]). Il problema etico consente di analizzare la pratica riflessa della libertà, eminentemente politica proprio come lo è la verità.

«Siamo quello che le aziende come Google e Facebook vendono agli inserzionisti» sostiene Astra Taylor (2014, p. 14). Il potere dei social media si fonda sulla commercializzazione delle conversazioni: l'innovazione più dirompente della Rete è la possibilità di inviare e-mail e non di fruire di negozi online. Gli utenti creano i contenuti e i proprietari delle Corporation si appropriano del valore generato nella “mezzadria digitale” (*Digital Sharecropping*) e senza che vi siano stati definiti meccanismi di tutela (*ivi*, p. 17), come se fossero gli utenti stessi a essere posseduti dai proprietari delle piattaforme e non in maniera anonimizzata.

Helen Kennedy (2016) scrisse che gli analisti dei negozi Target, con un volano di 25 prodotti, sarebbero stati in grado di predire la gravidanza di una adolescente e che le compagnie creditizie potevano stimare un imminente divorzio dei clienti in base al punteggio determinato dai loro acquisti. Sui social media, a ogni tipologia di attore (nel campo dell'educazione, della politica, delle arti o dell'intrattenimento) è richiesto di agire secondo una logica che dipende dalla struttura della piattaforma, di cui il *data mining* costituisce una norma intrinseca e ordinaria, come pratica che avviene convenzionalmente, in maniera naturale, ogni giorno e data per scontata. Le persone usano le piattaforme come spazi di “interiorità” per interagire e condividere pensieri che corrispondono a tracce ben diverse rispetto a quelle lasciate in ambienti digitali meno “intimi” (*ivi*, p. 20). Persino l'immaginario della condivisione non è casuale: il termine, preso in prestito dall'informatica (John, N. 2013 citato da Kennedy, 2016, p.24), è considerato versatile, adatto a descrivere la distribuzione dei contenuti, oltre trasmettere a un senso di eguaglianza, altruismo e relazioni sociali positive, ma risulta calzante anche per illustrare l'azione di trasferimento e a terze parti dei dati personali.

Nel 2010, Mark Zuckerberg sosteneva che la privacy non fosse più una norma sociale, tant'è che i social media sono vissuti come luoghi per il benessere psicologico in cui viene chiesto di esporre il proprio vissuto; non sono considerati media e neanche tecnologie. Il corto circuito avviene quando l'essere del soggetto e la sua azione si scontrano con il *design*

dell'infrastruttura. È la manifestazione compiuta del *bios* che si relaziona con il suo governo, direbbe Bazzicalupo (2010).

La moderazione dei contenuti è una pratica definitoria di adeguatezza (Roberts S. T., 2017). E conviene, in quanto la pubblicazione e il mantenimento di contenuti inappropriati sconfina nella corresponsabilità. Rimuovendo o riducendo si garantisce all'utente un'esperienza positiva in conformità con i requisiti previsti dalla legge, la reputazione del marchio non viene compromessa e i ricavi dalle attività pubblicitarie possono prosperare (Roberts S. T., 2019).

Nel corso del tempo, la retorica della libertà e dell'apertura della Rete è stata spesso utilizzata per sopprimere le preoccupazioni, classificando le incursioni come "censura", e razionalizzare la mancata partecipazione di alcune componenti della società basandola sulle preferenze individuali. Si è presto osservato che i medesimi soggetti, vittime di discriminazioni e discorsi d'odio, non venivano supportati nell'azione di denuncia per il trattamento che veniva loro riservato dalla comunità digitale (Taylor, 2014, p. 119). I volti dei proprietari delle piattaforme si assomigliano (*Male, Pale and Priviledged*) e le statistiche nel campo ICT tracciano il profilo di chi detiene la proprietà (e la destinazione d'uso) dei sistemi con cui ci interfacciamo su Internet. Alcune decisioni possono segregare intere categorie di persone, altre creare nuove connessioni, così come gli algoritmi possono escludere voci e nascondere punti di vista o le interfacce espandere il senso della possibilità umana o limitarla al familiare (*ivi*, p. 116). I "motori di predizione" sanno già a cosa stai pensando. Google, per esempio, è in grado di costruire un profilo tridimensionale (*ivi*, p. 132): *the knowledge person*, chi siamo in base a cosa cerchiamo sui motori di ricerca; *the social person*, chi siamo in base a come comunichiamo e ci colleghiamo, scrivendo e-mail o sui social media; *the embodied person*, la localizzazione rivelata da dispositivi come computer, cellulare e orologio.

Le informazioni raccolte sui social media non sono finestre sul mondo, ma il risultato di precise scelte di *data mining* su dove cercarle e come ottenerle (Kennedy, 2016, p. 37), così come le previsioni che saranno elaborate a partire da quelle tracce⁵⁴. I sistemi di IA non sono mai separati dagli assetti gestionali. L'interdipendenza stratificata tra amministrazione, tecnologia e ideologia non agisce nel vuoto. Incarna una logica "necropolitica", presa in prestito dal filosofo Aquille Mbembe, di definizione tra chi può vivere o perire, un quadro teorico che descrive come i Governi o le Amministrazioni assegnano un valore differenziale alla vita umana.

⁵⁴ Un passo dall'introduzione al libro di Numerico "Big Data e Algoritmi" (2021, p. 13): «Pensare che le tracce che lasciamo siano la migliore interpretazione significa inchiodarci al passato. Se ritenessimo che gli algoritmi fossero in grado di leggere la vera fonte del nostro desiderio, non avremmo nessun motivo per prendere decisioni né per cambiare, non potremmo realizzare la variabilità e molteplicità che ci caratterizza come specie e che ci ha garantito la sopravvivenza».

Il campionamento di secondo livello ha riguardato l'analisi del rapporto tra la piattaforma e i suoi utenti per approfondire l'influenza della regolamentazione interna sulla libertà di espressione. Dal 24 al 28 giugno 2024 sono stati raccolti 229 documenti di *policy* (Appendice B) sui canali liberamente accessibili delle sezioni Trasparenza, *Help Center* e *Frequently Asked Questions* (FAQ) appartenenti a Meta⁵⁵, X e ByteDance (Tab. 7). Ciascun *item* è stato consultato e archiviato in lingua inglese come garanzia di accesso a versioni più aggiornate, che ne ha consentito una migliore comparazione nelle medesime finestre temporali; inoltre, questa scelta ha successivamente agevolato il procedimento di codifica.

Per un'organizzazione coerente e sistematica dei materiali, il *software* di analisi qualitativa NVivo ha contribuito a gestire i documenti in maniera efficiente.

Tabella 7. *Campionamento policy su Meta, X e ByteDance*

Corporation	Meta	X	ByteDance
Documenti	57 (+2)	102	68
Lingua	EN	EN	EN
Area geografica	UE	UE	UE, EEEA, Svizzera
Colore	Blu/Rosa	Nero	Giallo

Fonte: *Community Standards, Terms of Service e User Agreements Policies* (agg. Al 06/2024)

Ciascun documento è stato esaminato, etichettato e collocato in una categoria (Tab. 8): *Platform Policy* è la categoria dove sono presenti più *item* (33%), seguita da *Community Guidelines* (30%) e *Misuse Policy* (23%).

Tabella 8. *Classificazione delle policy di Meta, X e ByteDance*

Categoria	Contenuto	N° <i>item</i>	Rilevanza
Terms of Service	Linee guida, accordi, condizioni	13	0,06%
Privacy Policy	Privacy, sicurezza, minori	8	0,03%
Cookies Policy	Archiviazione e raccolta dati	7	0,03%
Platform Policy	IA, sistemi di raccomandazione, pubblicità, premi	75	0,33%
Misuse Policy	Restrizioni, contenuti proibiti e terroristici	53	0,23%
Community Guidelines	Marchi, proprietà intellettuale, reclami e appelli	68	0,30%

Fonte: *Community Standards, Terms of Service e User Agreements Policies* (agg. Al 06/2024)

⁵⁵ L'analisi dei documenti include Facebook, Instagram e Threads.

La fase di codifica delle fonti è stata condotta utilizzando la funzione *Code*, attraverso la quale i documenti sono stati allocati in categorie tematiche predefinite, basate sulle domande di ricerca e su nessi teorici. I nodi “Genere”, “Razza” e “Classe” – le categorie logiche utilizzate per rappresentare concetti, temi o idee - sono stati ulteriormente suddivisi per consentire un’analisi dettagliata di sottotemi emergenti. In particolare, è stato utilizzato un approccio di analisi tematica, attraverso cui sono stati individuati iterativamente temi e parole chiave applicando gli strumenti *query* di testo e analisi delle frequenze (inclusa la visualizzazione della densità di parole e temi), offerti dal *software*. NVivo consente, altresì, la creazione di mappe concettuali e grafici che facilitano la visualizzazione delle relazioni tra i temi, migliorando al contempo la comprensione dei fenomeni esaminati. Il *software* ha reso possibile la gestione di un grande volume di dati in modo organizzato e tracciabile, garantendo la trasparenza del processo di codifica e verifica durante tutte le fasi della ricerca.

L’analisi prosegue con la presentazione dei risultati per ciascun nodo, supportata da fonti secondarie (articoli, pubblicazioni scientifiche e inchieste, nonché le decisioni *dell’Oversight Board* di Meta) che hanno garantito l’accesso a ulteriori dati e ampliato le possibilità di interpretazione, talvolta mettendo in discussione lo sguardo privilegiato di chi ha condotto la ricerca, affinché fosse possibile una lettura intersezionale del fenomeno osservato.

4.2 Il Genere tra corpi, violenza e identità

Nelle “regioni” della sessualità e della politica, dove si articola il legame con il desiderio e il potere, si sviluppa un reticolo complesso tra il tabù dell’oggetto, il rituale della circostanza e il diritto privilegiato o esclusivo del soggetto che parla. La procedura più evidente e familiare di esclusione diventa l’interdizione, dice Foucault (2004 [1972]), e si esercita per mezzo del discorso nelle forme della manifestazione (o dell’occultamento) e della lotta (o del sistema di dominazione). Che il genere sia un campo di battaglia su cui possono verificarsi casi di standardizzazione è una considerazione quasi storica. Le tematiche legate al “Genere” (Fig. 10 e 11) sono state individuate in 126 documenti per un totale di 1726 *references*.

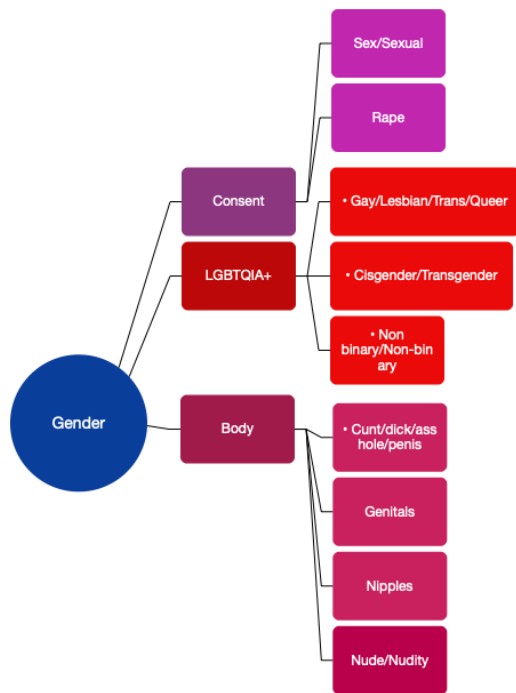


Figura 10. Mappa della ricerca per il nodo “Genere”

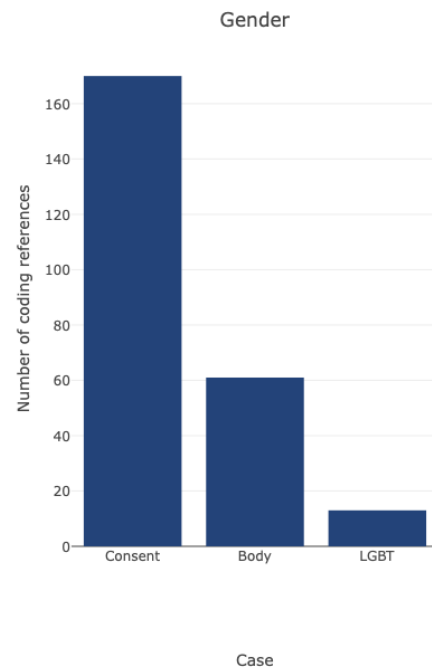


Figura 11. Distribuzione delle query per casi

Come spesso accade in ambito tecnologico, il consenso viene inteso come “consenso informato” (assenso) e “parentale” (permesso da adulto o tutore) e non come autodeterminazione del corpo, piena e libera disposizione degli attributi personali, nonché delle informazioni che riguardano la persona interessata, oltre i termini del rapporto sessuale. Il termine va inteso come politicamente intersecato da oppressioni di genere, di classe, razziali e abiliste. Pertanto, “Consenso” è stato inquadrato come un caso contenente le occorrenze relative al processo continuo tra la comunicazione (volontà), la sessualità (sesso) e la sua violazione (stupro).

Nelle Linee Guida per la Community di Instagram (Meta) si invita a postare contenuti (foto, video o altri contenuti digitali) che siano appropriati per il pubblico, escludendo aprioristicamente la nudità, ove siano mostrati rapporti sessuali, genitali (capezzoli femminili), primi piani di natiche nude, ma sono consentite immagini di allattamento, del parto e dei momenti successivi al parto, situazioni legate alla salute (ad esempio, post-mastectomia, sensibilizzazione sul cancro al seno o interventi chirurgici di conferma del genere) o di un atto di protesta. Dipinti e sculture sono generalmente accettati, mentre nessuna tolleranza è prevista per contenuti di natura sessuale che coinvolgono i minori e le minacce di diffusione di immagini intime (*revenge porn*). Gillespie (2018) dedica un intero capitolo a “*Facebook, breastfeeding,*

and living in suspension”: il *Nipplegate* avvenuto nel 2007⁵⁶ ha coinvolto alcune foto d'autore in cui oltre all'allattamento si mostrava l'esperienza della nascita, rimosse o sospese a causa di una *policy* troppo vaga che vietava “nudità, uso di droghe e altri contenuti osceni”; gruppi come “*Hey Facebook, Breastfeeding Is Not Obscene!*” (da cui partì una protesta) e una campagna di sensibilizzazione a cura dell'artista Micol Hebron (“*This is a male nipple*”), comparsa nel 2014, denunciavano l'erotizzazione, esercitata dalla piattaforma, di un'attività innocua. Sull'argomento è intervenuto più tardi anche l'Oversight Board, per segnalare la sproporzionata discriminazione di trattamento verso uomini e donne nell'esposizione del proprio corpo:

«I sistemi automatici di Facebook non hanno riconosciuto le parole “Cancro al seno”, che comparivano sull'immagine in portoghese, e il post è stato rimosso per errore. Poiché le regole di Facebook trattano i capezzoli maschili e femminili in modo diverso, l'uso di automatismi imprecisi per far rispettare queste regole colpisce in modo sproporzionato la libertà di espressione delle donne. Anche un'applicazione che si affida esclusivamente all'automazione senza un'adeguata supervisione umana interferisce con la libertà di espressione» (*Breast Cancer Symptoms and Nudity*, Decisione 20-004, 28 gennaio 2021).

Facebook (Meta) impedisce l'uso dei suoi servizi, tra gli altri, a persone condannate per reati sessuali e ritiene inaccettabile la riproduzione di sfruttamento sessuale, abusi e nudità dei minori nel contesto pubblicitario (denunciando l'evento al Centro nazionale per i bambini scomparsi e sfruttati, NCMEC), sanzionando anche rappresentazioni di persone in posizioni esplicite/ammiccanti, o attività allusive.

L'offerta di servizi sessuali⁵⁷, anche laddove sussista il consenso delle parti, non è consentita e ciò a discapito delle persone impegnate nel *sex working* (Are, 2022). Lo “Sfruttamento sessuale” (Capitolo “Sicurezza” degli Standard della Community di Facebook) è presente con due differenziazioni, la prima più generica e la seconda rivolta ai minori. Lo “Sfruttamento sessuale di adulti”⁵⁸, nella sua versione inglese, contiene una dichiarazione rivolta direttamente a chi pratica *sex working* («*We also allow for the discussion of sex worker rights advocacy and sex work regulation. We draw the line, however [...]*»), non presente nella medesima sezione italiana. In entrambe le versioni, la parola “vittima” è stata recentemente sostituita con “sopravvissuto”. Sul tema della condivisione di immagini intime non consensuali (NCII), spesso legate a *revenge porn*, Meta informa dell'adesione al *Cyber Civil Rights Initiative*

⁵⁶ Una data emblematica perché, fino al 2006, la platea che aveva accesso alla piattaforma era circoscritta alle studentesse e agli studenti della Ivy League, e chi si occupava della gestione dei contenuti non prevedeva di porsi il dubbio etico sull'ammissibilità di un contenuto contenente scene di allattamento (Gillespie, 2018, p. 146-147).

⁵⁷ Gli emendamenti FOSTA/SESTA (2018) hanno introdotto un'eccezione di corresponsabilità alla Sezione 230 del CDA, per cui le Corporation digitali cancellano e censurano i post per non incorrere in sanzione come promotori o facilitatori di prostituzione.

⁵⁸ Adult sexual solicitation and sexually explicit language (Security, Facebook Community Standards), <https://transparency.meta.com/en-gb/policies/community-standards/sexual-exploitation-adults/> (10 ottobre 2024).

(CCRI)⁵⁹ e ne prevede la rimozione, previa segnalazione. Dal 2019 un sistema di *photo-matching* registra il contenuto audiovisivo e si accerta che non vengano pubblicate nuovamente. L'azienda ha dichiarato che l'implementazione del ML permette di rilevare in modo proattivo le immagini o i video di quasi nudo che vengono condivisi senza autorizzazione su Facebook e Instagram⁶⁰, a cui è stato affiancato un dipendente del Dipartimento *Community Operations* per il controllo dei risultati, nonché una procedura di appello. Bassi livelli di interazione tra la vittima e la segnalazione erano dovuti anche ai tempi e alle modalità di risposta delle tecnologie di Meta, come è emerso da una ricerca condotta sul campo⁶¹:

«Le vittime con cui abbiamo parlato, ad esempio, hanno detto di essere state scoraggiate da quella che ritenevano essere la risposta robotica di Facebook all'NCII: semplicemente togliere le immagini senza trovare un modo per riconoscere il trauma subito dalle vittime. Altre vittime non avevano familiarità con le denunce online e non erano sicure di come orientarsi nelle nostre procedure. Abbiamo anche sentito forte e chiaro che il tempo che intercorre tra la segnalazione a Facebook e la risposta da parte nostra è pieno di ansia e impotenza. Le vittime vogliono risposte rapide ma personalizzate, perché il danno che subiscono aumenta quanto più a lungo le immagini rimangono online».

X si riserva la possibilità di mettersi in contatto con le persone coinvolte (anche per interposta persona) prima di prendere provvedimenti e la potenziale vittima può segnalare fino a 5 post. Su Meta la promozione di eventi e luoghi in cui si svolgono attività a sfondo erotico è consentita, ma non la pubblicazione di video, performance e messaggi tantrici. È concessa la divulgazione e l'educazione sessuale, anche per mezzo di disegni, a meno che non offra un adescamento sessuale (fanno eccezione la penetrazione, l'autoerotismo e l'esercizio di scenari fetish). In caso di racconto esplicito sui fatti legati alla violenza sessuale, condivisi dal soggetto interessato e da terzi per condannare l'atto o per una consapevolezza generale da determinare in base al contesto/alla didascalia, la piattaforma si riserva di inserire un *disclaimer* per segnalare il contenuto sensibile. Lo stesso trattamento è riservato ai video di immersione violenta in acqua che ritraggono minori durante riti religiosi. Nella versione di TikTok disponibile in Giappone e Taiwan, sono vietati gli annunci a sfondo religioso (*Branded Content*).

⁵⁹ Meta, "Not without my consent": <https://about.fb.com/wp-content/uploads/2017/03/not-without-my-consent.pdf>

⁶⁰ Davis, A., Global Head of Safety (15 marzo 2019). "Detecting Non-Consensual Intimate Images and Supporting Victims". Newsroom: <https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/> (12 ottobre 2024).

⁶¹ Iyengar, R. Head of Product Policy Research and Nain, K., Global Safety Policy Programs Manager (15 marzo 2019). "A Research-Based Approach to Protecting Intimate Images". Newsroom: <https://about.fb.com/news/2019/03/protecting-intimate-images/> (12 ottobre 2024).

L'etichetta è autonomamente apposta dagli utenti su X per avvertire della presenza di contenuti violenti⁶², pornografia e altre forme di contenuti per adulti prodotti in modo consensuale, tollerati sulla piattaforma («chi non desidera vederlo potrà scegliere di non procedere con la visione»⁶³). Ciò permette ai sistemi di raccomandazione di non mostrare il contenuto ai minori con età inferiore ai 13 anni (o ai 16 anni in Australia⁶⁴). L'*eSafety Commissioner* australiano, istituito nel 2015, è il regolatore indipendente per la sicurezza online che subentra qualora le norme e i provvedimenti di X non siano sufficienti.

Chi inneggia alla violenza o minaccia di stupro sulle piattaforme di Meta potrebbe subire una sospensione (*ban*) dell'account, così come previsto anche dalle norme di TikTok (*Content violations and bans*) e X (*X Rules*).

È prevista la chiusura permanente dell'account identificato come autore originale del post con contenuti multimediali intimi creati o condivisi senza consenso su X, mentre si richiede la rimozione (a rischio blocco momentaneo dell'account) a chi condivide il contenuto incriminato per esprimere shock e incredulità o per denunciare il caso. In un'altra *policy*⁶⁵, si chiarisce che fa parte dell'esperienza degli utenti decidere se condividere o meno un contenuto e ciò significa che non è previsto controllo proattivo come scelta di *design*.

L'inclusione del chiarimento⁶⁶ sui *deepfake* non ha impedito, nel gennaio 2024, la circolazione di contenuti sintetici pornografici con gli attributi fisici della cantante Taylor Swift e di altra persona non meglio identificata, minando alla reputazione della prima e violando i diritti di riproduzione della seconda. La piattaforma potrebbe anche stabilire un periodo di “sola lettura” per gli account che commettono violazioni considerate lievi⁶⁷.

Nella sezione “Contenuti fuorvianti e falsi”⁶⁸ di ByteDance è prevista una specifica disciplina per il contenuto sintetico (*deepfake/cheapfake*) in ambito pubblicitario, prodotto con IA generativa o significativamente manipolato (con le tecniche *face-swapping*, *voice-cloning*,

⁶² The X rules: safety, privacy, authenticity, and more (Rules and Policies), <https://help.x.com/en/rules-and-policies/x-rules> (27 giugno 2024).

⁶³ Non-consensual nudity policy (Safety and Cybercrime), <https://help.x.com/en/rules-and-policies/intimate-media> (27 giugno 2024).

⁶⁴ Australia Online Safety Information (X Help Center), <https://help.x.com/en/rules-and-policies/australia-resources> (27 giugno 2024).

⁶⁵ X's private information policy and doxxing (Rules and Policies), <https://help.x.com/en/rules-and-policies/personal-information> (27 giugno 2024).

⁶⁶ Sono vietati «immagini o video che sovrappongono o manipolano in altro modo digitalmente il volto di una persona sul corpo nudo di un'altra persona» in X's private information policy and doxxing (Rules and Policies), <https://help.x.com/en/rules-and-policies/personal-information> (27 giugno 2024).

Sono vietati «immagini o video che sovrappongono o manipolano in altro modo digitalmente il volto di una persona sul corpo nudo di un'altra persona» in *ibid*.

⁶⁷ Private content (Safety and Cybercrime), <https://help.x.com/en/rules-and-policies/personal-information> (27 giugno 2024).

⁶⁸ Misleading and False Content (TikTok Advertising Policies), <https://ads.tiktok.com/help/article/tiktok-ads-policy-misleading-and-false-content> (30 giugno 2024).

ecc.), accettato solo dietro consenso della persona raffigurata, inclusa la conformità alle norme su diritto d'autore, protezione dei dati e della privacy.

Ciò è contemplato anche su X, dove la piattaforma si riserva di etichettare i post contenenti media sintetici e manipolati per aiutare gli utenti a comprenderne l'autenticità e per fornire un'informazione di contesto aggiuntiva. Indipendentemente dal consenso, in Giappone non è possibile promuovere contenuti (film, programmi TV, videogiochi) a sfondo sessuale o destinati agli adulti (*Branded Content*) su TikTok.

Nella *policy* “Comportamenti offensivi e molestie” (*Abuse and Harrassment*, Regole di X) è inserita la tutela sull'utilizzo di pronomi diversi da quelli con cui il soggetto si identifica e *deadname* che, una volta accertata la violazione o l'intenzionalità, comporta una riduzione della visibilità dei post. La piattaforma preferisce adottare parole non segnate dal genere per descrivere una mansione o un ruolo (citando Whoopi Goldberg: «*An actress can only play a woman. I'm an actor – I can play anything*»)⁶⁹. “*Queer*” e “omosessuale” sono considerati termini dispregiativi in alcuni contesti geografici (*markets*) e si prega di usarli con cautela. Si chiarisce l'uso corretto dei termini sulla transizione pur senza fare a meno di citare formule:

«Evitiamo di usare termini come “transessuale”, “pre/post-operatorio”, “nato nel corpo sbagliato” e “cambio di sesso”. Generalmente usiamo “*crossdresser*”/“*cross-dresses*” al posto di “travestito” e “intersex” piuttosto che “ermafrodito”. Evitiamo i termini che implicano che qualcuno stia “cambiando sesso”».

Nel documentario *Facebookistan* (2015) veniva documentato il caso che ha riguardato l'eliminazione dell'account di alcune *Drag Queen*, che per ragioni di privacy e riconoscibilità preferivano usare il nome d'arte sulla piattaforma, promuovendo eventi e campagne di *fundraising*. Tra i requisiti di Facebook figurava, infatti, l'inserimento del vero nome (*Real Name Policy*), giudicato ingiusto e discriminatorio nei confronti degli attivisti perché li rendeva riconoscibili e, in taluni casi, perseguibili. Grazie alle pressioni della stampa, fu concesso l'utilizzo dello pseudonimo, ma poco tempo più tardi gli utenti ricevettero richieste di riconoscimento tramite documento di identità, al seguito delle quali migliaia di persone sono state segnalate e sospese.

⁶⁹ Our style guide for curation of Moments (Platform Use Guidelines, febbraio 2021), <https://help.x.com/en/rules-and-policies/curationstyleguide>

Secondo la sezione “Incitamento all’odio”⁷⁰ (Capitolo “Contenuti Deplorevoli”, *Community Standards*) di Facebook sono vietate le esternazioni rivolte a persone o gruppi con caratteristiche protette, che ne trasmettano l’inferiorità, sotto forma di generalizzazioni, affermazioni ed espressioni di disprezzo:

«Per carenze mentali si intendono quelle relative a:
Capacità intellettive (stupido, imbecille, idiota);
Istruzione (analfabeta, ignorante);
Salute mentale (malato di mente, ritardato, pazzo, folle).
Per carenze morali si intendono quelle relative a:
[...] Termini dispregiativi correlati ad atti sessuali (puttana, troia, pervertito).
[...] Imprecazioni, tranne che nel contesto della fine di una relazione romantica, definite come:
Riferimento a qualcuno citando genitali o ano (figa, cazzo, buco di culo);
Termini o frasi volgari con l'intenzione di insultare (fottiti, puttana, figlio di puttana);
Termini o frasi che augurano l'interazione in atti sessuali o il contatto con genitali, ano, feci o urina (succhiarmi il cazzo, baciami il culo, mangia merda)».

Il testo è stato redatto con un approccio abilista che, se da un lato marginalizza specifiche categorie di persone, dall’altro adotta un atteggiamento moralizzante, ancora una volta, nei confronti di persone impegnate nelle attività di *sex working*. Peraltro, a esclusione del contesto romantico, dove sembra essere tollerato il linguaggio scurrile e il tono accusatorio. In alcuni casi di satira o critica accertata, si specifica, sarà consentita la pubblicazione.

Su Instagram (*Community Guidelines*), una specifica disciplina svincola le opere artistiche e creative (foto, video, sculture), anche se digitali, e i capezzoli femminili sono tollerati ma soltanto in specifici contesti legali alla maternità (allattamento, parto), alla salute (ad esempio, post-mastectomia, sensibilizzazione sul cancro al seno o interventi chirurgici di conferma del genere) o ad un atto di protesta. La *policy* “Immagini di nudo e atti sessuali di adulti”⁷¹ (Capitolo “Contenuti deplorevoli”) ribadisce la posizione dell’azienda già espressa sui contenuti a sfondo sessuale, aggiungendo che le fattispecie eccezionali potrebbero essere accompagnate dall’etichetta di contenuto sensibile (rientra in questa categoria il materiale audiovisivo raffigurante l’esito dell’intervento chirurgico di transizione sessuale), consentiti in particolare nel contesto di carestie, genocidi, crimini di guerra o crimini contro l'umanità. Sono vietati i contenuti che chiedono, ammettono o offrono servizi di violenza come la mutilazione genitale femminile (Capitolo Violenza e Comportamenti Criminali). Su TikTok, a quanto risulta, riferimenti ai genitali nelle campagne pubblicitarie sono vietati soltanto in Giappone.

⁷⁰ Hate Speech (Objectionable Content, Facebook Community Standards), <https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/> (9 ottobre 2024).

⁷¹ Adult nudity and sexual activity (Objectionable Content, Facebook Community Standards), <https://transparency.meta.com/en-gb/policies/community-standards/sexual-exploitation-adults/> (10 ottobre 2024).

Su X la diffusione di nudità di adulti e comportamenti sessuali di tipo pornografico o finalizzati a eccitare sessualmente (inclusi primi piani di genitali, natiche o seni) sono tollerati quando consensuali e debitamente etichettati come un contenuto sensibile (Contenuti per adulti)⁷²:

«[...] Sosteniamo l'autonomia degli adulti nel creare e interagire con contenuti affini alla loro fede, ai loro desideri e alle loro esperienze, inclusi i contenuti relativi alla sessualità. Per controbilanciare questa libertà, evitiamo di mostrare i contenuti per adulti ai bambini e agli adulti che non desiderano vedere tali contenuti. Vietiamo anche i contenuti che promuovono lo sfruttamento, la mancanza di consenso, l'oggettificazione, la sessualizzazione dei minori o l'abuso dei minori e i comportamenti osceni. Inoltre, non ammettiamo la condivisione di contenuti per adulti in punti molto visibili come la foto del profilo o l'immagine d'intestazione».

Peraltro, la pubblicazione recidiva di un contenuto privo di etichetta non comporta sanzione, bensì legittima l'intervento di inserimento manuale da parte dei moderatori di X.

Negli ultimi anni, i negozi di *sex toys* sono dovuti correre ai ripari per eludere le restrizioni della piattaforma (quivi menzionate), ricorrendo a immaginari creativi con oggetti ed espedienti per la promozione dei propri prodotti. Su TikTok, invece, sono apertamente vietati e inseriti nella voce “Prodotti sessuali” (*Adult Content*, sezione *Advertising Policies*). In Giappone, prodotti farmaceutici quali coppette mestruali o legati alla salute riproduttiva sono vietati. Anche X ne impedisce la vendita come *merchandising*⁷³, insieme a tanti altri oggetti e prodotti quali contenuti “*cause-based*” (per educare, sensibilizzare o agire in relazione all'impegno civico, alla crescita economica, alla tutela dell'ambiente o a cause di equità sociale), parti del corpo, fuochi d'artificio, NFT, dispositivi medici regolamentati (si menzionano i tiralatte), armi, alcool, tabacco e affini, accessori di e con contenuto politico (con riferimento a candidati, elezioni o provvedimenti giudiziari). Appartenendo alla categoria “Contenuti sessuali per adulti” ne riduce anche la monetizzazione per i *creator*⁷⁴.

L'approccio localizzato di ByteDance ha diviso il Medio Oriente e il Nord Africa (MENA) in più subregioni e ciascun moderatore di contenuti è assegnato ad almeno una di queste, con corredo di insulti, politiche e regole culturali; sappiamo, per esempio, che una donna che indossa maniche corte è inaccettabile nella regione MENA 2, ma non nelle regioni MENA 1 e 3 (Elsawah, 2024, p. 18). La biancheria intima legata a un marchio (*Branded Content*) è accettata nelle zone European Economic Area (EEA), Svizzera e Regno Unito, Nord America, America Latina, MENA, Israele, Australia e Nuova Zelanda, ma limitata per i minori di 18 anni; è proibita in Corea se indossata da un/a modello/a. Le Linee guida sui Prodotti creativi (*Creative*

⁷² Adult Content (X, maggio 2024), <https://help.x.com/en/rules-and-policies/adult-content> (31 ottobre 2024)

⁷³ X Shopping Policy (Platform Use Guidelines, agg. 6 settembre 2024), <https://help.x.com/en/rules-and-policies/shopping-policies> (31 ottobre 2024).

⁷⁴ X's Creator Monetization Standards (X Help), <https://help.x.com/en/rules-and-policies/content-monetization-standards> (27 giugno 2024).

Guidelines) invitano a non “promuovere o rafforzare immagini negative o malsane del corpo e non devono sfruttare le insicurezze per conformarsi a determinati ideali o standard di bellezza”. In Nord America, America Latina, Indonesia, Malesia, Vietnam, Cambogia, Filippine, Singapore, Tailandia, Australia e Nuova Zelanda sono vietati i contenuti pubblicitari su prodotti e servizi contraccettivi, in aggiunta a qualsiasi prodotto, servizio, metodo artificiale o altra tecnica - medica o non medica - per prevenire o interrompere le gravidanze, compresi i servizi legati all'aborto.

4.3 La Razza tra etnia, culto e colore della pelle

Così come inquadrata da Angela Davis (2018 [1981]), la “Razza” non è un’entità oggettiva, ma un prodotto delle dinamiche di potere e delle gerarchie di oppressione, sviluppatosi in particolare attraverso la storia della schiavitù, della colonizzazione e della segregazione, che oggi costituisce un attributo interiorizzato di differenziazione sociale. Sono stati individuati 17 documenti e un totale di 48 *references*, suddivise per occorrenze direttamente collegabili a un attributo etnico, religioso o fisico (Fig. 12 e 13).

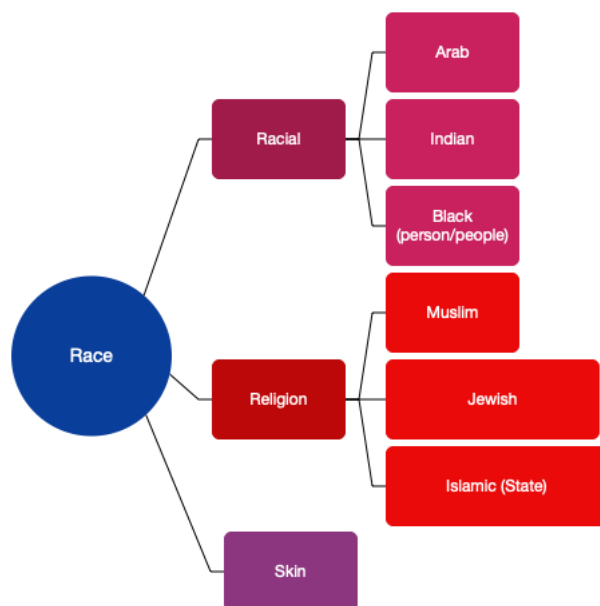


Figura 12. Mappa della ricerca per il nodo “Razza”

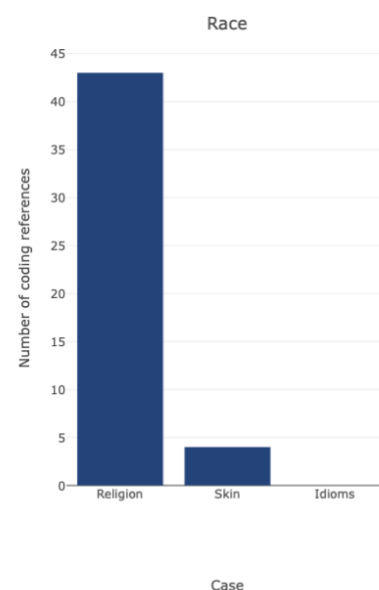


Figura 13. Distribuzione delle query per casi

La sezione “Incitamento all’odio” di Meta contiene l’invito a non pubblicare contenuti su una persona o un gruppo «sulla base della loro caratteristiche o dello status di immigrato» a cui fa seguito un elenco (a titolo esemplificativo e non esaustivo), contenente termini di paragone “disumanizzanti” da intendersi come vietati, tra cui:

«Animali in generale o tipi specifici di animali percepiti nella cultura come fisicamente o intellettualmente inferiori (persone di colore e scimmie o creature simili alle scimmie, ebrei e ratti, musulmani e maiali, messicani e vermi);
Esseri subumani (selvaggi, demoni, mostri, primitivi);
Predatori sessuali (persone di credo islamico che fanno sesso con capre o maiali);
Determinati oggetti (donne rappresentate come oggetti domestici o proprietà o oggetti in generale, persone di colore come attrezzature da fattoria, persone transgender o non binarie come "cose");
Stereotipi storicamente pericolosi collegati a intimidazione, esclusione o violenza sulla base di una caratteristica protetta, come il *blackface*, la negazione dell'olocausto, le dichiarazioni secondo cui gli ebrei controllano le istituzioni finanziarie, politiche o mediatiche o i riferimenti ai Dalit come lavoratori di bassa estrazione [...]».

Le leggi degli stati vengono costruite secondo fattispecie generali e astratte, a cui il giudice rinviene nel bilanciare i principi rilevanti e individuare la norma adattabile alle circostanze del caso concreto o facendo ricorso allo *stare decisis*, ovvero alle sentenze che costituiscono precedente vincolante. Meta, invece, inserisce specifiche casistiche in forma di aggravanti.

Dal punto di vista definitorio e linguistico, l’esercizio non convince per diverse ragioni: nel tentativo di specificare cosa non dovrebbe essere pubblicato, il testo compie una divisione sociale nominando gruppi di persone sulla base del culto, dell’estrazione sociale, dell’origine; l’enumerazione nel dettaglio (presumibilmente uno storico di violazioni registrate) riproduce una logica di rappresentazione in cui, chi differisce dal “normale” rispondente agli attributi dell’uomo abile, bianco, cisgender, occidentale e benestante, viene vittimizzato una seconda volta come categoria identitaria protetta («sulla base delle caratteristiche»); “Esseri subumani⁷⁵” è una classificazione discriminatoria che presenta l’accostamento tra selvaggi (coloro che vivono lontani dai centri urbani) e primitivi (appartenenti alle età o alle popolazioni preistoriche o privi delle tecnologie proprie del mondo moderno) e figure fantastiche del mito e della fantasia (demoni e mostri); infine, la traduzione italiana ha usato l’espressione “persone di colore” per tradurre “*Black people*”, dando per scontato che esista un colore primario e un colore che differisce dall’ordinario.

Il recente studio di Cotter (2023) ha esaminato le proteste pubbliche di alcuni *influencer* attivi nel movimento *Black Lives Matter* che avevano riscontrato una diminuzione della visibilità dovuta a pregiudizi razziali su Instagram, spingendo l’azienda a fornire dichiarazioni sul fenomeno. Grazie all’analisi del caso, l’autrice ha spiegato che il *black box shadowbanning*,

⁷⁵ Dove “subumano” per Oxford Languages significa: «Al di sotto del livello minimo che rende accettabile o possibile l'esistenza, e quindi offensivo dei diritti e della dignità dell'uomo».

tendente strategicamente a inserire gli errori sistemici tra i casi isolati, è un modo per mettere a tacere le critiche minando la credibilità della protesta od ostacolando la possibilità di parlare.

Un'ulteriore osservazione merita l'utilizzo discriminatorio della lingua. L'inglese e le lingue parlate nel Nord globale dominano la scena della conoscenza e della tecnologia, causando la marginalizzazione, ovvero l'assenza di un adeguato supporto linguistico, nel Sud Globale. Quando si introducono sistemi automatizzati di moderazione, la mediazione accentua le sue lacune: l'assenza di adeguata letteratura e ricerca scientifica, la sottorappresentazione in fase di *training*, l'incapacità di cogliere e interpretare le sfumature culturali. Secondo uno studio condotto dal *Center for Democracy & Technology* (Elswah, 2024), i sistemi automatizzati su Instagram rimuovono i commenti in cui è citato "Allah", non comprendendo il contesto della preghiera e dell'esclamazione di gioia. Inoltre, il modello di moderazione è lo stesso in tutto il mondo, fatta eccezione per TikTok che adotta un approccio regionale.

Nelle Linee Guida sull'Utilizzo di X (*X Moments Guidelines and Principles*) si dice esplicitamente che i "curatori" lavorano su 16 mercati con 5 lingue a disposizione (inglese, giapponese, arabo, spagnolo e portoghese). All'interno del documento è presente un chiarimento linguistico sui termini che coinvolgono le forze di polizia (*Police Shooting Terminology*) dove anche l'utilizzo degli esempi è discriminatorio:

«Approccio sbagliato: Un nero è morto giovedì in seguito a una sparatoria con la polizia in Wisconsin.

Approccio corretto: Un agente di polizia del Wisconsin ha sparato e ucciso un nero giovedì».

Mentre il paragrafo su Razza ed Etnicità⁷⁶ recita:

«Dobbiamo diffidare da una copertura che si traduce in una narrazione ristretta e singolare di diversi gruppi di persone, che può verificarsi in ampie questioni sociali. Ciò potrebbe includere la copertura dei neri e degli ispanici negli Stati Uniti, o di alcune popolazioni di immigrati in Europa, o la generalizzazione delle reazioni in base a classificazioni razziali o regionali, ad esempio "le donne asiatiche amano questa nuova serie televisiva"».

L'assenza di *diversity* nello sviluppo di NLP, unitamente all'insufficienza di *dataset* idiomatici e il reclutamento di personale non madrelingua esacerba l'inaccuratezza della moderazione automatizzata. L'idea di uniformare un intero continente alla visione occidentale, demandando a pochi moderatori di adattare le medesime decisioni a prescindere dalle sfumature semantiche, aumenta in maniera esponenziale la possibilità di incorrere in errore. Oltre all'impatto sulla lingua nativa, la colonizzazione - combinata all'importazione tecnologica - ha influenzato il modo in cui le persone scrivono in arabo, causando la nascita di una peculiare forma di comunicazione informale e sui social media chiamata "Arabizi" (3rabizi), che impiega numeri

⁷⁶ Our style guide for curation of Moments (Platform Use Guidelines, febbraio 2021). Disponibile qui: <https://help.x.com/en/rules-and-policies/curationstyleguide>

e lettere latine in assenza di supporto in *script* e rappresenta un ulteriore fattore di sfida nel fenomeno del “*Code-switching*”, ovvero il passaggio da una lingua all’altra nella medesima frase, che cambia, a seconda dell’utilizzo, anche l’orientamento di lettura (Elsawah, 2024, p. 7). Secondo i risultati raccolti nello studio del CDT⁷⁷, il 57% degli intervistati non crede che le Corporation digitali si stiano impegnando per colmare le lacune dei sistemi automatizzati, il 58% non prova fiducia nei loro confronti, il 62% teme che i propri post verranno rimossi e ciò li inibisce dal creare contenuti, ma si fidano più di TikTok che di Meta (*ivi*, p. 10).

4.4 Questioni di Classe: mansioni, salario e coscienza

La discriminazione di “Classe” (Fig. 14 e 15) è stata la più complessa da indagare perché non esplicitamente riscontrabile nei testi. Nonostante ciò, sono stati codificati 90 documenti e 216 *references*.

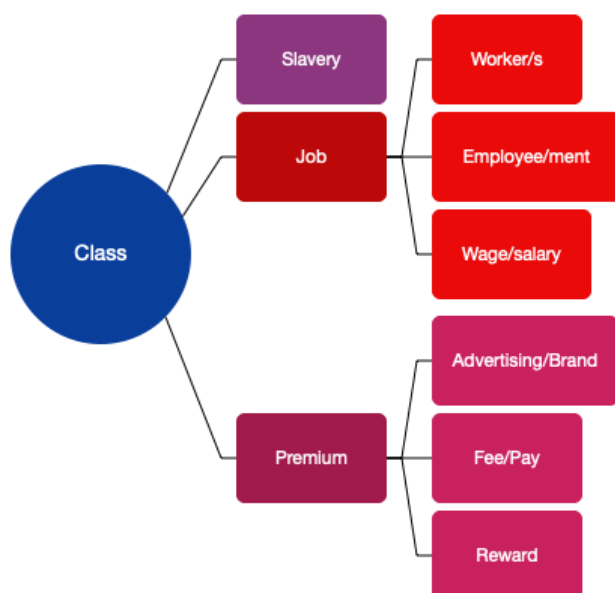


Figura 14. Mappa della ricerca per il nodo “Classe”

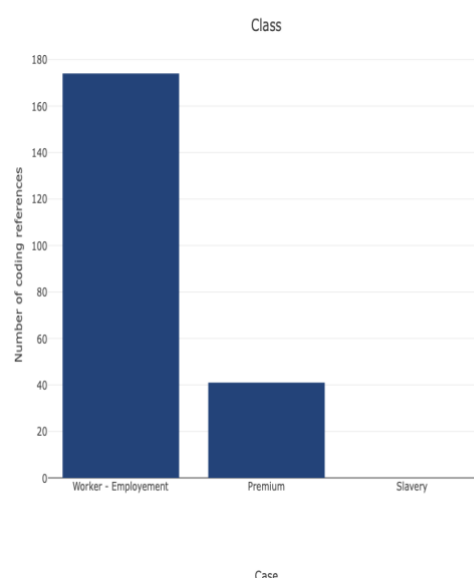


Figura 15. Distribuzione delle query per casi

Tra i documenti analizzati emerge il “*TikTok UK Modern Slavery Statement 2023*”, in cui si afferma che la società con sede nel Regno Unito proibisce la schiavitù e, nel rispetto dei Diritti Umani, «si impegna a garantire che non si verifichi all'interno delle proprie attività o della propria catena di fornitura»⁷⁸. La dichiarazione invita a prevenire, individuare e segnalare in forma di responsabilità collettiva da parte di dipendenti (inclusi dirigenti e direttori) o aziende

⁷⁷ Sono stati coinvolti 14 moderatori e dipendenti, in servizio e non, del settore *Trust & Safety* (T&S) e *Policy* di Meta, TikTok e YouTube, un centinaio di utenti e diversi ricercatori impegnati in NLP.

⁷⁸ TikTok UK Modern Slavery Statement 2023, approvato dal Comitato di Direzione e sottoscritto dal direttore Adam Michael Presser il 20 giugno 2024, <https://www.tiktok.com/legal/page/global/modern-slavery-statement/en> (9 ottobre 2024).

esterne (appaltatori indipendenti e collaboratori di terze parti), a cui si chiede di rispettare il Codice di Condotta. Con valenza di requisito contrattuale, al suo interno sono inclusi i divieti al lavoro minorile, forzato o involontario e alla tratta di esseri umani, a discriminazioni o molestie illegali, a ripercussioni per chi costituisce o aderisce a sindacati e associazioni di categoria, al superamento dei limiti orari e il mancato rispetto del salario minimo (previsto dalle leggi locali), e monitorando il loro verificarsi in aree come le pari opportunità di impiego, la lotta alla corruzione e concussione e la responsabilità sociale d'impresa. L'implementazione del piano avviene attraverso una piattaforma per lo *Speak Up* dove dipendenti e partner commerciali possono segnalare violazioni al Dipartimento *Ethics & Compliance*, senza che vi siano ritorsioni per chi denuncia o partecipa a indagini interne. TikTok dichiara di aggiornare periodicamente la *Global Recruitment Procedure*, che definisce le fasi principali del processo di assunzione e selezione e gli standard di condotta, e il suo programma *Environment, Health and Safety* (EHS), attraverso cui mira a garantire un luogo di lavoro sano, sicuro e rispettoso dell'ambiente. All'interno del documento è presente un passaggio rivolto alla *community* dove ByteDance dichiara di reindirizzare la ricerca (in Francia, Albania e Regno Unito) di alcune parole o frasi a pagine informative di “*Stop the traffik*” (organizzazione di beneficenza a cui sono stati concessi crediti pubblicitari), dove gli utenti possono fare segnalazioni o consultare risorse. Comunica, inoltre, di aver aderito a BSR⁷⁹, organizzazione globale senza scopo di lucro, e di collaborare con l'*International Center for Missing and Exploited Children* (ICMEC).

Rientrano tra i lavoratori *tech* di ByteDance i dipendenti del comparto *Security & Privacy* impegnati in *offensive adversary*, simulazione di un attacco informatico per introdursi nei sistemi e scoprirne le vulnerabilità con lo scopo di implementare migliori strumenti di difesa, e coloro che monitorano le operazioni di influenza clandestine⁸⁰. La *Global Security Organization* rileva e reagisce agli incidenti informatici critici in tempo reale a Singapore, Dublino e Washington D.C.⁸¹. Inoltre, l'azienda incoraggia “*hacker etici*” attraverso il programma *Bug bounty HackerOne*.

Come affermato nel documento *Our approach to Content Moderation* (2023), TikTok inquadra gli addetti alla moderazione dei contenuti nel Dipartimento *Trust & Safety* che conta 40.000

⁷⁹ Di cui fanno parte Google, Microsoft, Amazon, Meta, Spotify e Paypal.

⁸⁰ «Comportamenti coordinati e non autentici in cui delle reti di account collaborano in modo strategico allo scopo di ingannare le persone che utilizzano i nostri sistemi e influenzare le discussioni pubbliche» in *Countering influence operations* (TikTok Transparency Center, <https://www.tiktok.com/transparency/en-us/countering-influence-operations>) (30 giugno 2024).

⁸¹ *Strengthening security and privacy* (TikTok Transparency Center), <https://www.tiktok.com/transparency/en-us/security-privacy/> (30 giugno 2024).

unità⁸². I gruppi *Integrity & Authenticity* che si occupa di disinformazione ed *Election Integrity*, previste dal *Global Fact-Checking Program*, classificano i contenuti come accurati/falsi/non verificabili mediante *database* per sottoporli alla valutazione indipendente di 20 organizzazioni accreditate dall'International Fact-Checking Network (IFCN)⁸³ – in 50 lingue per 100 mercati - affinché ne valutino l'adequatezza. L'azienda si riserva di consultare esperti, accademici, nonché il Consiglio di sicurezza delle Nazioni Unite, per questioni legate a odio ed estremismo con un “approccio internazionale” basato sui Diritti Umani⁸⁴:

«A tale scopo, abbiamo un *team* dedicato a garantire l'equità dei nostri prodotti e delle nostre politiche per assicurare che le diverse comunità siano rappresentate e incluse. Questo *team* interdisciplinare è composto da responsabili di politiche e programmi incentrati su comunità specifiche (come LGBTQ+, BIPOC, persone con disabilità, e altre ancora) e da coloro che lavorano in modo olistico sul rispetto dei diritti umani» (*Combating hate and violent extremism*).

Tutti sono responsabili dei processi di *privacy by design*, in ogni fase del ciclo di vita dei prodotti e dei sistemi (dalla progettazione allo sviluppo, fino alla manutenzione). Il reindirizzamento al compartimento di competenza è facilitato da pagine di richiesta contatto⁸⁵. Nell'ambito della monetizzazione (*Coins Policy*), per esempio, l'utente può compilare un *form* online, far pervenire all'azienda un modulo di recesso o rivolgersi all'assistenza clienti.

Procedendo con la ricerca per *query*, e riscontrata l'assenza del termine “*worker/s*”, è stato necessario indagare la presenza più specifica di parole come “*employee*” e “*team*” per risalire alla struttura aziendale, con l'intento di comprendere la suddivisione delle mansioni.

Meta dichiara di avere tra le 15.000⁸⁶ e le 40.000 unità⁸⁷ impegnate nella revisione umana nel Dipartimento *Safety & Security*, con competenze in oltre 80 lingue. Spesso si fa riferimento a un *team* appropriato per la rifinitura dei difetti di piattaforma, ma l'azienda manca quasi sempre di menzionarlo, impedendo così ogni possibilità di indirizzare l'utente all'ufficio competente. Un organigramma di Instagram era presente nella *Legacy Instagram Platform Policy* (poi sostituito il 31 agosto 2020), in cui veniva espressamente comunicato all'utente (svincolato da interessi e relazioni datoriali) che funzionari, affiliati, agenti, concessionari, *co-brander* e dipendenti erano esenti da qualsiasi responsabilità «per reclami, danni (effettivi e/o conseguenti), costi e spese (compresi i costi di contenzioso e le spese legali) di ogni tipo e

⁸² Combating hate and violent extremism (TikTok Transparency Center), <https://www.tiktok.com/transparency/it-it/combating-hate-violent-extremism> (30 giugno 2024).

⁸³ Combating harmful misinformation (TikTok Transparency Center), <https://www.tiktok.com/transparency/en-us/combating-misinformation/> (30 giugno 2024).

⁸⁴ Combating hate and violent extremism (TikTok Transparency Center), <https://www.tiktok.com/transparency/it-it/combating-hate-violent-extremism> (30 giugno 2024).

⁸⁵ Segnala un problema (TikTok), <https://www.tiktok.com/legal/report/feedback>

⁸⁶ Additional protection of minors (Transparency Centre), <https://transparency.meta.com/en-gb/policies/community-standards/additional-protection-minors> (31 ottobre 2024).

⁸⁷ About community content reviews (Business Help Center), <https://www.facebook.com/business/help/373506759931554> (31 ottobre 2024).

natura, derivanti da o in qualsiasi modo correlati all'uso della Piattaforma e dei dati derivati attraverso tale Piattaforma»⁸⁸. Nel successivo accordo unificato *Developer Policies* il termine “employee” scompare e viene sostituito da sviluppatori (*developers*) e *web operators*, con cui Meta dialoga prevalentemente via notifiche (*alerts*) ed e-mail, e i cui accordi con i fornitori non devono confliggere con le condizioni di servizio. Viene riutilizzato nella *Privacy Policy* per specificare che lo scambio di dati avviene a livello globale tra utenti, partner (fornitori e terze parti) e dipendenti, a seguito della loro raccolta e archiviazione presso i *data center*. Successivamente, Meta ha dichiarato di aver ampliato il suo organico con figure dalla formazione interdisciplinare inserite nel Dipartimento *Responsible AI*⁸⁹:

«Per affrontare le preoccupazioni su trasparenza e inclusione associate all'implementazione dell'IA nelle tecnologie di Meta, abbiamo creato il *team Responsible AI*, un team dedicato e multidisciplinare composto da esperti di etica, ricercatori in ambito politico e sociale, esperti di normative, ricercatori e ingegneri di IA. L'obiettivo generale del *team* è sviluppare linee guida, strumenti e processi per gestire le problematiche legate alla responsabilità dell'IA e diffondere queste risorse sistemiche all'interno di Meta».

Infine, i moderatori di contenuti (*reviewers*) promuovono l'integrità del prodotto, lavorano a supplemento della tecnologia che individua le violazioni o forniscono la base di addestramento degli algoritmi per migliorare il procedimento automatizzato. L'azienda dichiara che queste persone lavorano per Meta, Meta Company o per un “fornitore di servizi di fiducia”⁹⁰.

Quanto alla formazione, di cui abbiamo traccia grazie ai *Transparency Report*, in Meta a ciascun addetto spetta una formazione specifica comprendente tre fasi: pre-formazione (reperimento fonti), formazione pratica (minimo 80 ore affiancati da un istruttore), consulenza continua (documentazione aggiornata e kit di approfondimento).

In formula simile anche su X, l'utente deve accettare i servizi “COSÌ COME SONO” e “COME DISPONIBILI” (citazione letterale “AS-IS”) dalle *X Entities* che fanno capo a X Corp., alle sue società affiliate, amministratori, dipendenti, agenti, rappresentanti, partner e concessionari, esenti da qualsiasi responsabilità per danno causato e arrecato dalla piattaforma. Sono presenti diversi *team* nell'azienda: Supporto, a cui rivolgersi per quesiti di varia natura⁹¹; Revisione delle segnalazioni, mentre i sistemi di IA rimuovono proattivamente i contenuti in violazione («*No humans look at reported posts, what's the point?*»⁹²); Sicurezza, composto da esperti a

⁸⁸ Legacy Instagram Platform Policy (Meta Development), <https://developers.facebook.com/docs/development/terms-and-policies/legacy-instagram-platform-policy> (26 ottobre 2024).

⁸⁹ Meta (26 gennaio 2022). In che modo Meta assegna la priorità ai contenuti da controllare, Transparency Center: <https://transparency.meta.com/it-it/policies/improving/prioritizing-content-review/> (29 ottobre 2024).

⁹⁰ How Meta collects and uses user data (Meta Privacy Policy), <https://www.facebook.com/privacy/policy> (24 giugno 2024).

⁹¹ Contattaci. Possiamo aiutarti, Centro assistenza (X Help) <https://help.x.com/it/forms> (4 novembre 2024).

⁹² Debunking X myths (X Help), <https://help.x.com/en/using-x/debunking-twitter-myths> (27 giugno 2024).

cui segnalare le vulnerabilità del sito; infine, i membri dell'*X Experiments Program* e dell'*Our Market Research*. X cita i *curators* nella sezione Momenti⁹³, generata da un algoritmo che permette di cogliere la conversazione durante lo svolgimento dell'evento, a cui il comitato redazionale (già menzionato nel precedente paragrafo) contribuisce, cercando ed evidenziando i migliori contenuti, secondo imparzialità (diversi punti di vista, come appaiono in piattaforma – formazione continua), accuratezza (doppio controllo prima della promozione – formazione continua), correttezza (informazioni esatte) e standard (rispetto delle Linee Guida). Se la piattaforma diventa essa stessa un argomento di tendenza (situazione per cui si configurerebbe un conflitto di interesse), i moderatori promuovono un riepilogo comprendente i fatti e una eventuale risposta da parte dell'azienda. L'allineamento valoriale è ritenuto il punto cardine della formazione⁹⁴:

«I dipendenti di X in tutta l'azienda tengono conto dei nostri valori fondamentali ogni giorno in tutti gli aspetti del loro lavoro, dalla progettazione dei prodotti al modo in cui gestiamo le richieste legali di informazioni sui conti e oltre. I nuovi assunti cominciano ad acquisirli dal primo giorno in X, durante il corso di orientamento, e li ritrovano continuamente nelle nostre comunicazioni interne ed esterne, oltre che nelle nostre azioni. Questi valori sono essenziali per il prodotto e l'azienda stessa, a tutti i livelli, inclusi la dirigenza e il consiglio di amministrazione».

È il Dipartimento *Trust & Safety* («a tempo pieno e fin dalla nascita nel 2008») ad avere il compito di “proteggere” persone e organizzazioni presenti su X. Al gruppo è affidato un ruolo centrale nel processo decisionale perché ha il compito di analizzare gli impatti sugli utenti di nuovi mercati, funzioni e modifiche ai termini contrattuali, nonché la responsabilità di facilitare acquisizioni, integrazioni e lo sviluppo di nuove regole, in dialogo costante con i settori prodotto, supporto utenti, vendite, monitoraggio legislativo e legale.

A causa di clausole stringenti su anonimato e discrezione contenute negli accordi di riservatezza, è molto complicato entrare in contatto con i *reviewers*. La figura comparve per la prima volta a seguito delle proteste per la rimozione di contenuti innocui negli USA, Adrian Chen realizzò l'articolo “*Inside Facebook's Outsourced Anti-Porn and Gore Brigade*” (2012), in cui spiegava l'esternalizzazione della forza lavoro impegnata nella revisione in Turchia, Messico, India e nelle Filippine⁹⁵. Due anni più tardi, il giornalista curò l'articolo di approfondimento “*The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*” (2014) sul funzionamento e la distribuzione delle mansioni su Facebook e Twitter. L'incarico si svolge principalmente a distanza, per cui la maggior parte del personale può essere

⁹³ Curation style guide (Platform Use Guidelines, febbraio 2021), <https://help.x.com/en/rules-and-policies/curationstyleguide>

⁹⁴ X's free speech and rights of people (X Help), <https://help.x.com/en/rules-and-policies/defending-and-respecting-our-users-voice> (27 ottobre 2024).

⁹⁵ Uno degli stabilimenti più grandi si trova a Manila e vi lavorano circa 1.000 persone (Roberts S. T., 2017).

tenuto fuori dai *business process outsourcing* (BPO) da parte di aziende terze (come oDesk, Covalen e Cognizant) ed essere soggetto a reclutamenti a cottimo digitale, come nel caso di *Amazon Mechanical Turk* o *Upwork* (Roberts S. T., 2017). Anche se integrati nell'organigramma aziendale, questi operatori sono destinatari di contratti a termine e percepiscono uno stipendio sensibilmente più basso se confrontato con altri dipendenti nel settore tecnologico, sebbene l'esposizione a materiale disturbante sia condizione di assunzione (*ibidem*). Molte delle informazioni disponibili provengono da vere e proprie inchieste giornalistiche che si sono concentrate sullo stile di vita, sulle mansioni e sul flusso di lavoro, nonché sulle condizioni di deterioramento della salute psicologica dei moderatori di contenuti, su cui ByteDance dichiara di avere un programma attivo basato sull'evidenza (Tik Tok, 2023):

«Queste misure possono includere strumenti e funzioni che consentono ai dipendenti di controllare l'esposizione ai contenuti grafici durante la revisione o la moderazione dei contenuti, tra cui il *grayscale*, il *muting* e l'offuscamento; formazione per i manager per aiutarli a identificare quando un membro del *team* potrebbe aver bisogno di ulteriore supporto per il benessere; e supporto clinico e terapeutico».

Ai dipendenti viene, inoltre, offerta l'iscrizione alla *Trust and Safety Professional Association* (TSPA) per accedere a risorse per lo sviluppo di carriera, partecipare a workshop ed eventi ed entrare in contatto con una rete di colleghi.

«Ogni minuto di ogni giorno, 500 ore di filmati sono caricati su YouTube, 450.000 tweet compaiono su Twitter, 2,5 milioni di post vengono caricati su Facebook. Ma non tutto ciò che va in rete ci resta» recitano i titoli di testa del documentario “Quello che i social non dicono” (2018). Un tipico “*cleaner*” deve osservare e valutare, allenando mediante la sua scelta i sistemi di IA alla decisione («Gli algoritmi non possono fare ciò che facciamo noi», sosteneva uno degli intervistati). Il controllo qualità avviene sul 3% del rapporto di un moderatore da parte di un caposquadra. I video più problematici possono essere scartati ma, se scoperto, l'azione viene conteggiata come errore e possono essere commessi un massimo di tre errori al mese. Queste persone sconosciute prendono importanti decisioni editoriali.

In “Facebookistan” (2015), la *reviewer* intervistata racconta che la paga dipende dalla quantità di contenuti analizzata: 1 dollaro/ora di base, 4 dollari/ora superata la soglia delle 4.000 foto (l'equivalente di circa 0,9 secondi per ciascuna foto). I moderatori del Maghreb hanno dichiarato di percepire uno stipendio equivalente a 400-600 dollari, con un potenziale bonus di 100 dollari al raggiungimento di alcuni parametri di rendimento, considerandolo inadeguato al trauma e al rigido sistema di valutazione (che oscilla tra il tempo di visione e l'accuratezza della decisione) a cui sono soggetti (Elsawah, 2024, p. 16).

È qui che si interseca la discriminazione etnica e di classe: accettare un lavoro usurante diventa una missione per ostacolare la diffusione dei peccati nel mondo e pulire la Rete dallo sporco, dall'immorale. Alcune persone si sentono come forze di polizia o eroi che combattono i crimini online (difendono i minori, agiscono contro il terrorismo, riducono il cyberbullismo). Il ritmo, la natura dei contenuti, la mole di materiale da esaminare e la segretezza possono portare a sentimenti di isolamento, *burnout* e depressione in alcuni operatori (Roberts S. T., 2017). Erin Elder e Selena Scola sono state le prime impiegate di Facebook a chiedere un risarcimento per danni alla salute mentale. Alberto Cuandra è il protagonista di una storia illustrata per il Washington Post (Lockwood, 2024) che racconta gli effetti di una prolungata esposizione a contenuti deprecabili. Dopo aver lavorato per una piattaforma di *streaming* per meno di un anno a contatto con video di omicidi, suicidi, maltrattamenti e abusi, violenza e bullismo, la sua salute mentale risulta danneggiata con livelli di rabbia e ansia che lo accompagnano in diversi momenti della giornata. L'azienda aveva sottoposto ai dipendenti una guida su come affrontare già nei primi secondi l'asprezza di alcuni contenuti e poterli classificare come "abuso su animali"; fatto che scatenò isteria di gruppo tra i moderatori, con il terrore di dover esaminare un contenuto appartenente a quella categoria. Un lavoro, che non esisteva fino a una generazione fa, ben pagato per gli standard del suo paese, diventa come una porta girevole tra sussistenza e sopravvivenza.

All'indomani dell'acquisizione di Twitter nel 2022, Elon Musk ha dato vita a una catena di licenziamenti. La nuova amministrazione ha tagliato drasticamente i numeri del personale impegnato nel rendere i sistemi più trasparenti ed equi, tra cui Rumman Chowdhury, direttrice del dipartimento *ML Ethics, Transparency, and Accountability*, e tutto il suo *team*, che lavorava a un progetto sui pregiudizi politici che avrebbe fatto scuola tra i social network, in grado evitare la penalizzazione di alcuni punti di vista⁹⁶ (Knight, 2022). La notizia fece scandalo soprattutto perché Twitter era considerato un modello virtuoso di congiunzione tra la ricerca etica e tecnologica in ambito aziendale, pronto ad esternalizzare il giudizio - interagendo con il pubblico - per migliorare la qualità del prodotto. La stessa sorte è toccata ai dipendenti che si stavano occupando di Diritti Umani, Accessibilità e Comunicazione interna, mentre sono stati decimati i gruppi *Public Policy* e Verifica account (Hatmaker, 2022). Poco tempo più tardi anche il *Trust & Safety Council*⁹⁷, gruppo consultivo di circa 100 organizzazioni indipendenti per i diritti civili, fu sollevato dall'incarico senza preavviso e via e-mail, diventando pretesto

⁹⁶ I ricercatori avevano dimostrato come fonti di notizie orientate a destra fossero promosse più di quelle orientate a sinistra in Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A. & Hardt, M. (2022). "Algorithmic amplification of politics on Twitter", *Proceedings of the National Academy of Sciences*, 119(1), e2025334119.

⁹⁷ Insieme al MOB, un altro esempio di *network governance* (Caplan, 2023).

per pubblici attacchi alla precedente dirigenza da parte di Musk. L'attuale prassi del CEO di X è consultare i suoi *followers* piuttosto che gli esperti prima di apportare modifiche alle *policies* sulla moderazione dei contenuti (The Associated Press, 2022).

A ottobre 2024, ByteDance ha annunciato che circa 500 dipendenti sono stati licenziati (via e-mail) in Malesia, come parte di una strategia di riduzione della forza lavoro – oltre 110.000 unità in 200 città del mondo - che sarà sostituita da sistemi di IA (ora già impiegati all'80%) nella moderazione dei contenuti. Ciò è avvenuto all'indomani di una stretta normativa del Governo malese che ha chiesto alle Corporation digitali di esibire entro gennaio 2025 una licenza operativa e di intensificare il controllo sui contenuti (Latiff, 2024).

In “La Rete non ci salverà” (2022), Lilia Giugni racconta la storia della lavoratrice di Covalen (intermediaria di Meta) Isabella Plunkett che, seguita dall'avvocata Cori Crider, ha presentato al Parlamento irlandese alcune raccomandazioni per regolare la moderazione dei contenuti, tra cui vietare gli NDA facendo leva sul “*duty of care*” (dovere di cura) nei confronti della forza lavoro. Giugni suggerisce di spingere le Corporation a mantenere *in-house* le attività al centro del loro *business* (ivi, p. 159). Imprescindibile è rimettere al centro i diritti dei lavoratori e delle lavoratrici e non lasciare che questa funzione venga svolta soltanto da attori privati: ciò passa per il riconoscimento delle organizzazioni sindacali dei *tech worker*⁹⁸.

4.5 Cos'è politico? Tra etica e ordine sociale

Nei termini in cui Laura Bazzicalupo (2015) intende le modalità biopolitiche di governare il presente, la moderazione dei contenuti è un “dispositivo di cattura” e gestione dell'abbondanza, che genera fattori di incertezza, rischio e “necessità” di controllo, caso per caso, dell'imprevisto. Le Corporation digitali si avvalgono della facoltà di ammettere il contenuto, che esso sia o meno consentito dalla regolamentazione esterna. Intervistato nel 2018, David Kaye, inviato speciale ONU per la libertà di espressione, dichiarò che l'ingerenza nel dibattito pubblico avrebbe interferito sulla capacità di pensiero critico delle persone, impoverendo la società («Limitando la libertà di espressione, limitiamo la sfida e lo sforzo», Quello che i social non dicono, 2018). Mentre i contenuti “sensibili” (come episodi di violenza e droghe) sono potenzialmente visibili in modalità standard su Meta, i “contenuti di natura politica”, che menzionano “governi, elezioni o argomenti sociali che interessano un gruppo di persone e/o la società in generale”, vengono limitati a meno che venga spuntata l'opzione contraria⁹⁹ (*Political Content Approach*).

⁹⁸ Un'operazione che, lontano dai riflettori, svolge a livello globale Tech Workers Coalition (TWC).

⁹⁹ «We use AI systems to personalize the content you see based on the choices you make; we aim to avoid making recommendations on Facebook, Instagram, and Threads that could be political in nature» (settembre 2024).

Il rapporto di Accountable Tech (luglio 2024) dimostra che la portata degli account (*creator* e organizzazioni) che pubblicavano assiduamente contenuti politici¹⁰⁰ è calata del 65% nel giro di pochi mesi, con impatti negativi sull'educazione, l'attivismo e la mobilitazione elettorale.

Altre voci sostengono che il provvedimento intende arginare la democratizzazione dei social media che, in assenza di *gatekeeper*, ha dato visibilità a coloro che non fanno capo ai media tradizionali e, pertanto, sono portatori di disinformazione (Stewart, 2024). Intanto, a causa di un errore tecnico, riscontrato pochi giorni prima del dibattito tra i candidati alla Presidenza USA, Instagram ha continuato a limitare i contenuti politici anche dopo che gli utenti avevano optato per la visualizzazione, a ogni riapertura dell'app (*ibidem*).

Parallelamente, è indubbio che una simile disposizione stia creando una discriminazione ageista nei confronti degli utenti non alfabetizzati alle modalità di modifica delle impostazioni.

Secondo l'*European Media Freedom Act* (2024), i benefici del mercato interno si manifestano quando c'è garanzia delle libertà fondamentali e certezza giuridica. È così che la fruizione di notizie e contenuti di attualità diventa un antidoto contro la disinformazione e la manipolazione:

«Tale aspetto è particolarmente importante per le notizie e i contenuti di attualità, che comprendono un'ampia gamma di contenuti di interesse politico, sociale o culturale a livello locale, nazionale o internazionale. Le notizie e i contenuti di attualità possono svolgere un ruolo importante nel plasmare l'opinione pubblica e hanno un impatto diretto sulla partecipazione democratica e sul benessere della società» (Considerando 14, EMFA, 2024).

In combinato disposto con il DSA, l'EMFA prevede tutela contro la rimozione ingiustificata per quei contenuti mediatici prodotti secondo standard professionali, destinata a giornalisti (dipendenti e *freelance*) e responsabili editoriali. Se non sussistono rischi sistemici (come la disinformazione), le VLOP che intendono rimuovere contenuti – secondo i propri ToS – sono tenuti a informare i fornitori di servizi mediatici delle ragioni, prima che la sospensione/limitazione diventi effettiva (art. 18.4 EMFA) ed eventuali reclami dovranno essere trattati con priorità, nonostante sussista opacità nelle regole di revisione, soprattutto nei contesti politici e sociali più sensibili (Nikiforov, 2024, p. 5).

Il regolamento, in previsione del rischio di concentrazione del mercato dei media, avverte della possibile riduzione delle opinioni divergenti. Fornitori e operatori devono essere posti nelle condizioni di conoscere l'entità del pubblico per garantire parità di condizioni e valutare il successo dell'offerta (Considerando 71, EMFA).

¹⁰⁰ Argomenti come il cambiamento climatico, la libertà riproduttiva, i diritti LGBTQIA+, le informazioni sul voto e la giustizia razziale.

Il Regolamento sulla trasparenza e al targeting della pubblicità politica (2024) configura il messaggio «che può ed è inteso a influenzare l'esito di un'elezione o referendum, un comportamento di voto o un processo legislativo o regolamentare a livello dell'Unione, nazionale, regionale o locale» come pubblicità politica, mentre è “comunicazione pubblica” quell’attività svolta a favore di o per conto di un'autorità pubblica di uno Stato membro o dell'Unione. In occasione delle Elezioni Europee 2024, gli enti pubblici si sono ritrovati nella condizione di pubblicare contenuti che informassero i cittadini delle nuove disposizioni di Meta, spiegando come completare il procedimento di spunta, affinché essi stessi non venissero penalizzati con cali dell’*audience* dalla nuova *policy*.

Le «opinioni politiche espresse a titolo personale costituiscono una particolare manifestazione del diritto alla libertà di espressione e di informazione» e si escludono dal campo di applicazione le attestazioni a titolo personale che derivano dall’autodeterminazione individuale quando non remunerate. Il considerando 54 del Regolamento chiarisce che l’intenzione del legislatore non è precludere apriori la possibilità di pubblicare:

«Pur prevedendo requisiti specifici, gli obblighi di cui al presente regolamento non dovrebbero essere interpretati come se imponessero ai prestatori di servizi intermediari un obbligo generale di monitoraggio dei contenuti politici condivisi da persone fisiche o giuridiche, ovvero un obbligo generale di prendere provvedimenti proattivi in relazione a contenuti illegali che tali prestatori trasmettono o conservano» (Considerando 54, Regolamento 900, 2024).

Il *Political Content Approach* di Meta anche se non rende il contenuto “politico” direttamente classificabile come inammissibile, lo discrimina nel volano delle alternative possibili come l'intrattenimento, il commercio e il *lifestyle*. È veramente possibile affermare che esistano categorie di contenuto “apolitiche”? Il legislatore europeo considera i servizi di media come strumenti per soddisfare i principi democratici e contribuire ai processi decisionali, immaginando che l'incontro tra domanda e offerta di informazioni si sviluppi in una sfera pubblica dove sia la diffusione che la ricezione siano equamente accessibili.

Nel mese di maggio 2024 milioni di utenti hanno ricondiviso nelle *Stories* un’immagine generata con IA contenente la dicitura *All eyes on Rafah* tra le opzioni della funzione *Tocca a te* (Fig. 16). Il gesto di condivisione ha rappresentato una delle *call to action* di maggiore successo che Instagram abbia mai sperimentato. L’adesivo invita gli utenti a sviluppare un’azione di imitazione, usufruendo della possibilità di raggiungere cerchie più ampie del proprio bacino di followers. Il procedimento ha aggirato il *Political Content Approach* sfuggendo al controllo algoritmico del contenuto previsto per il *Feed*.

Bonini e Trerè (2024), e prima ancora Gillespie, hanno osservato che la materialità dei dispositivi e delle reti è legata al processo di riappropriazione del controllo da parte degli utenti. Ne sono una dimostrazione pratiche di elusione in cui si inserisce la *algorithmic agency* (ivi, p. 19), definita come la capacità riflessiva degli esseri umani di esercitare potere e di far funzionare gli algoritmi per soddisfare le proprie esigenze, in una relazione simbiotica e ricorsiva nelle interazioni tra società e tecnologia.



Figura 16. L'adesivo comparso nelle stories a maggio 2024

Le campagne di depoliticizzazione delle Corporation digitali sono strategie implementate per prendere le distanze dalle posizioni scomode e dai conflitti associati alle lotte per il potere (Scharlach, 2024). La campagna *TikTok For Good* intende amplificare la positività, minimizzare la negatività, concentrarsi sul benessere individuale e attribuire obblighi agli utenti, declinando questi obiettivi nella sua responsabilità sociale d'impresa (*Corporate Social Responsibility*) per dimostrare una condotta etica. Questa iniziativa di *soft governance* (ibidem) non è che un tassello nel disegno globale delle piattaforme, che vorrebbero de-responsabilizzarsi dalle tensioni geopolitiche.

Com'è noto ormai da tempo, le aziende contraggono accordi con i governi per non incorrere nel blocco delle piattaforme¹⁰¹. La Turchia, per esempio, già soggetta a importanti interventi sulla stampa, è il paese con il maggior numero di limitazioni. Non trovando spazio nei media tradizionali, le minoranze etniche hanno ritenuto Facebook e Twitter canali più appropriati di comunicazione, espressione e rappresentazione non mediata. Si è presto riscontrato che le pagine legate ai movimenti per la libertà del Kurdistan venivano censurate da Facebook, incluse le foto raffiguranti le bandiere del Partito dei Lavoratori del Kurdistan (PKK), dal 2017 dissociata dalle organizzazioni terroristiche dalla Corte d'appello belga. Il Partito Democratico dei Popoli (*Halkların Demokratik Partisi* - HDP), forza politica di opposizione, ha ricevuto diverse segnalazioni e ripetutamente la sua pagina, inclusi i profili dei deputati, è stata oscurata senza ricevere alcuna spiegazione (Facebookistan, 2015). La piattaforma riteneva i titolari degli account responsabili anche dei contenuti postati nei commenti o pubblicati in bacheca da terzi. Lo stesso è accaduto al *Syrian Network for Human Rights*.

Nel Maghreb organizzazioni della società civile e accademici si sono assunti l'onere di contattare le aziende per ridurre i tempi di risposta in caso di moderazione mendace: secondo lo studio del CDT, il 40% degli intervistati ha subito la rimozione dei contenuti e ipotizza che sia avvenuto per motivi politici; oltre il 62% delle persone che parlano arabo magrebino teme di essere messa a tacere, anche mediante *shadowban*, dalle piattaforme (un moderatore di TikTok ha dichiarato di poter classificare un contenuto come “difficile da trovare”); coloro che hanno partecipato ai focus group hanno menzionato il conflitto a Gaza tra le cause della censura (Elsawah, 2024, p. 12-13). Il ricorso a parole e frasi in codice (*algospeak*) è diventato molto comune e si sviluppa per mezzo di alcune tattiche come l'accostamento tra contenuti casuali e contenuti problematici, la sostituzione di lettere e numeri, la rimozione ortografica dei punti, l'utilizzo di simboli o *emoji* (l'anguria per riferirsi alla bandiera palestinese, per esempio) ed evitare alcune parole (*ivi*, p. 14).

Accade, però, molto spesso che i modelli securitari applichino indicatori significativi di “radicalismo” ad aspetti banali presenti nella quotidianità di comunità emarginate e sottorappresentate - che si vedono, così, escluse dal discorso pubblico per pura appartenenza culturale - a dimostrazione della presenza di un costrutto sociale nell'inquadramento della minaccia terroristica (Bloch-Wehba, 2020, p. 41-42). Pochi anni fa, Intercept scoprì che i moderatori di Facebook disponevano di una lista di 37 organizzazioni terroristiche proveniente dal Dipartimento della Sicurezza Interna USA ed era loro richiesta la memorizzazione di ogni

¹⁰¹ Gli algoritmi utilizzati per distinguere i discorsi d'odio dalle espressioni politiche legittime sono programmati per rispettare *policy* in cui la parola di élite e governi prevale su quella di movimenti dal basso e minoranze etniche (Angwin & Grassegger, 2017).

informazione utile a poterle bandire, inducendoli ad allinearsi alla politica estera del Nord Globale (USA, UK e UE). La *Dangerous Individuals and Organizations List* (DIO), contenente oltre 4.000 persone e gruppi considerati terroristici (mediorientali, sud-asiatici e musulmani) e imprese criminali violente (neri e latini), era accompagnata da una *policy* sulle punizioni, ovvero sulla censura (Biddle, 2021).

La Corporation distingue due livelli di terrorismo (*Community Standards*):

- 1) Odio organizzato, attività criminali su larga scala, tentativi di atti di violenza con più vittime, atti di violenza con più vittime, omicidi seriali e violazioni violente¹⁰²;
- 2) Soggetti non statali violenti ed entità che inducono alla violenza (incluse le reti cospirazioniste che induce alla violenza).

Come di consueto, gli esempi esprimono giudizi su alcuni soggetti e gruppi, menzionando: Al-Qaeda, Hitler, ISIS, Hizbul Mujahideen, Ku Klux Klan, Atomwaffen Division, Partito Nazista Americano e le cosiddette “ideologie d’odio” (Nazismo, Supremazia bianca, Nazionalismo bianco, Separatismo bianco). Uno studio più recente di Divon, Are & Briggs (2025) ha coinvolto *creators* ebrei che combattono l'antisemitismo, palestinesi per i Diritti Umani e attivisti LGBTQIA+, combinando i dati raccolti con le dichiarazioni pubbliche di Meta e TikTok, per dimostrare la presenza di *gaslighting* umano e automatizzato: alcuni intervistati denunciano di aver segnalato post contenenti apologia di nazismo e antisemitismo per capire quanto efficace fosse la moderazione, ricevendo una notifica che li informava della mancata violazione. La spiegazione è stata fornita in passato da Nicole Wong, già responsabile *policy* in Google e Twitter: il contenuto terroristico è contestuale perché la sua classificazione (critica, satira, minaccia, notizia, commento) dipende sia dalla piattaforma sulla quale viene pubblicato sia dal momento storico. Secondo l’azienda, mantenere online il video dell’impiccagione di Saddam Hussein era importante da un punto di vista storico («L’esecuzione di un ex dittatore»), mentre le successive immagini del corpo inerme vennero rimosse (Quello che i social non dicono, 2018). Attraverso la MAC, Facebook assegna punteggi predittivi ai post stimando la probabilità che violino le politiche sul terrorismo (Elsawah, 2024, p. 19), ma le ha allentate nei confronti della Brigata Azov, un gruppo paramilitare neonazista ucraino di estrema destra, nel contesto della guerra tra Ucraina e Russia¹⁰³ (Bouko & Rothut, 2024). Con largo anticipo, lo

¹⁰² «Organizzazioni che diffondono l'odio, organizzazioni criminali, tra cui quelle identificate dal governo degli Stati Uniti come *Specially Designated Narcotics Trafficking Kingpins* (boss del narcotraffico con designazione speciale, SDNTK), e organizzazioni terroristiche, inclusi individui ed entità identificati dal governo degli Stati Uniti come *Foreign Terrorist Organizations* (organizzazioni terroristiche straniere, FTO) o *Specially Designated Global Terrorists* (terroristi globali con designazione speciale, SDGT)».

¹⁰³ Dal 2022, anche Meta è un’organizzazione terroristica per la Russia, dopo che Facebook ha consentito per un breve periodo di inneggiare alla violenza contro i russi (Capone, 2022).

studio di Diaz e Hecht-Felella (2021) aveva già riscontrato la presenza di un doppio standard nella moderazione applicato all'ISIS rispetto ai contenuti dei suprematisti bianchi.

Accettare alti tassi di errore nella MAC è una decisione politica che si basa sul preferire l'eliminazione dei “contenuti cattivi” al «progettare sistemi con tassi di errore minimi o con la massima tutela della libertà di parola» (Bloch-Wehba, 2020, p. 44).

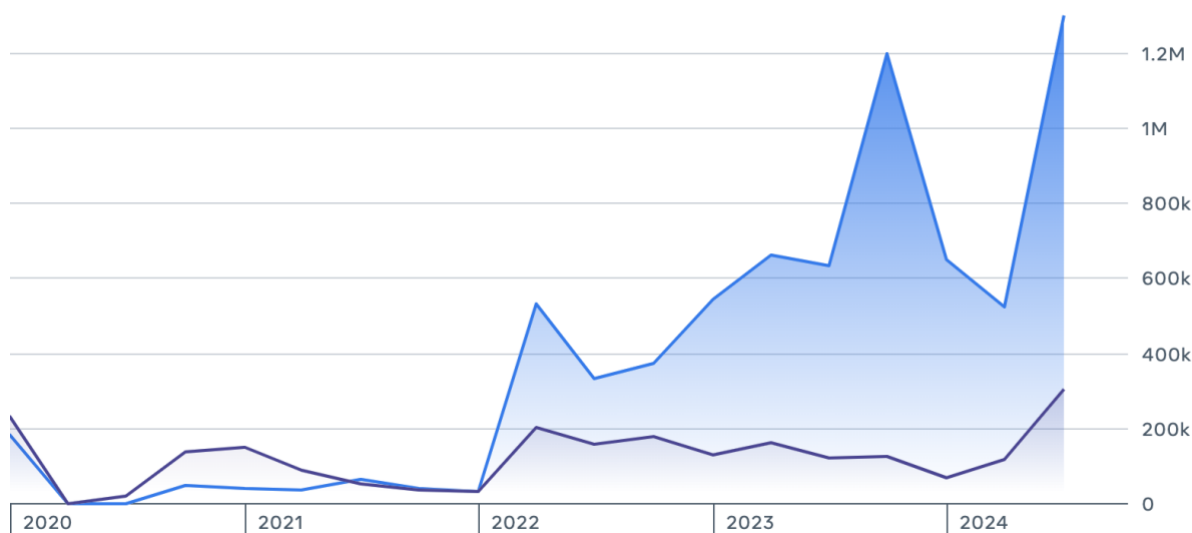


Figura 17. Quanti dei contenuti che abbiamo bloccato per organizzazioni pericolose sono stati impugnati? Fonte: Transparency Center di Meta. In blu “terrorismo”, in viola “odio organizzato”.

TikTok ha aderito al progetto europeo *Tech Against Terrorism* (TaTE) per prevenire l'estremismo violento (de Baillencourt, 2022) con un approccio di “tolleranza zero”. La MAC viene utilizzata per individuare chi promuove o esalta la violenza fuori dalla piattaforma o minaccia la sicurezza pubblica (l'1% dei video rimossi rientra in questa categoria); inoltre, si tiene conto dei comportamenti al di fuori della piattaforma per individuare individui e gruppi potenzialmente terroristici.

Sulla base delle informazioni in possesso, l'analisi della personalità fondata su interessi, conoscenze e ambizioni è un tassello cardine per la predizione dei comportamenti futuri e, anche in questo caso, il tema è eminentemente politico. Ciò che si verifica sulle piattaforme online dal punto di vista degli utenti è una libera obbedienza (Bazzicalupo, 2015, p. 36) di soggetti attivi che si autoregolano, sebbene siano controllati, e si aggregano, mentre chi sta dall'altro lato compie un atto di contenimento, di selezione ed esclusione delle eccezioni, a cui ciascuno è chiamato a rispondere singolarmente e non come parte di un insieme sociale.

Capitolo 5

Corpi, *agency* e giustizia nell'*Ethically Aligned Design*

«*For kin makers of all the oddkin*»
Donna Haraway

5.1 Le teorie della Giustizia oltre la ragione Artificiale

La giustizia è un concetto sia giuridico che filosofico, legato a nozioni di etica, moralità e stato di diritto, in continua tensione tra le libertà individuali e le esigenze della società. In termini filosofici, la giustizia è spesso discussa in relazione all'equità, ai diritti e al bene comune, con pensatori come Platone e Aristotele che offrono diverse interpretazioni di ciò che costituisce una società giusta. Le teorie normative della giustizia distributiva “riattivano” (Bazzicalupo, 2014, p. 50) la critica illuministica per interpretare i fenomeni politici in maniera diversa dalle teorie economiciste, a partire dall'imparzialità universale nel giudizio.

In “Teoria della Giustizia” (1971), John Rawls identifica criteri razionali per formulare giudizi politici, ricorrendo alla ragione pubblica, in cui Giustizia (*Fairness*) indica il fondamento trascendentale in ragione della critica al dato reale (Bazzicalupo, 2014, p. 52). In uno scenario ipotetico, le persone che hanno il compito di progettare la struttura di base della società (i contraenti, i *designer*) scelgono le regole in una posizione originaria, coperti da un velo di ignoranza che consente la scelta impersonale a favore di due principi della giustizia sociale: uguaglianza (nell'accesso alle libertà) e differenza (a condizione che avvantaggi i più deboli). In questo modello ideale, concentrato sugli aspetti istituzionali, ciò che è giusto differisce da ciò che è bene o utile, assicurando il coinvolgimento di tutte le parti negli accordi sociali.

La critica che Amartya Sen propone in “L'idea di Giustizia” (2009) si concentra sui beni primari, non garantiti a causa delle condizioni di partenza, cioè della capacità diseguale di ciascuno. La capacitazione (*Capability*) è la garanzia di uguaglianza nelle opportunità e l'impegno a perseguire i “funzionamenti” sostanziali di essere e di fare, osservando il capitale sociale e la possibilità di agire. Giusto è il sistema che valuta «gli aspetti materiali e immateriali della relazione tra persona e contesto» (Bazzicalupo, 2014, p. 53). Il modello della giustizia comparativa rifiuta la standardizzazione dei diritti e l'omologazione dei bisogni, concentrandosi sulla riduzione delle ingiustizie. Il processo di deliberazione e discussione è essenziale affinché le basi concettuali di ogni azione siano chiare e godano «di un'adesione consapevole e continua nel tempo» (*ibidem*).

Con “Giustizia sociale e dignità umana” (2002), Martha C. Nussbaum arricchisce il ragionamento sulle *capabilities* con un’analisi fondata sulla centralità della persona, intesa come fine, che propende verso la sua autodeterminazione e rigetta la riduzione a numero statistico, preferendo la varianza alla media generale (Bazzicalupo, 2014, p. 54). Le società, sostiene, dovrebbero essere giudicate in base a quanto permettono agli individui, prescindendo dalle risorse di partenza, di sviluppare ed esercitare le capacità essenziali, quali: la vita, la salute fisica, il pensiero (i sensi e l’immaginazione), le emozioni, la ragione pratica, la partecipazione alla vita sociale e politica (l’affiliazione), il rispetto per le altre specie (animali e naturali), il gioco (godere delle attività ricreative e di svago), il controllo del proprio ambiente. Il suo modello si concentra sull’idea di una vita degna fondata sull’irriducibilità degli esseri umani e sulle vulnerabilità dei gruppi emarginati, sostenendo il valore sociale di emozioni come la compassione e l’empatia. Tracce di una sintesi a cui era giunta tempo prima Iris M. Young in “Le politiche della differenza” (2024 [1996]): una società giusta deve tenere conto delle differenze e consentire l’autorappresentazione senza dominare gli esiti dello scontro politico; per farlo, ha bisogno di rivitalizzare lo spazio pubblico (e così il discorso e il linguaggio) e lasciarvi accedere i cittadini senza ridurli a consumatori (Bazzicalupo, 2014, p. 223).

Tatiana Dancy affronta le peculiarità dell’implementazione di sistemi di IA nell’ambito della giustizia penale nella sua opera dal titolo “*Artificial Justice*” (2023). In questo settore, gli algoritmi vengono utilizzati per prendere decisioni sulla distribuzione delle risorse di polizia e sugli obiettivi di sorveglianza e punizione che ricadranno su intere popolazioni. Il sistema “*Correctional Offender Management Profiling for Alternative Sanctions*” (COMPAS), sviluppato da Equivant, predice il tasso di recidiva stabilendo *pattern* basati sulla relazione tra le caratteristiche individuali e l’attività criminale; a partire dal 1998 supporta l’attività giudiziaria per facilitare la “diagnosi”. Il *software* include una serie di algoritmi di predizione del rischio “configurabili dall’utente”, tra cui uno strumento di avvicinamento al processo (*Pre-Trial Services*) e tre scale di valutazione (generale, violenza e completa) della recidività. I dati prodotti realizzano un profilo statisticamente criminale, rispondente a un punteggio consolidato sull’incrocio di indici e categorie deterministiche fondate su fattori sociali e familiari, come casi di delinquenza in età giovanile, disturbi vocazionali/educazionali, frequenza nell’assunzione di droghe. Dancy definisce scomoda (“*uneasy*”) la vicinanza tra la storia del metodo statistico e l’eugenetica: quando il risultato inferenziale viene considerato “giusto”, si applica a gruppi sempre più ampi con il rischio di culminare in ambizioni politiche¹⁰⁴, strategie di

¹⁰⁴ Charles Goring individuò le condizioni ricorrenti dei detenuti per prevedere apriori le cause dell’incarcerazione

segregazione¹⁰⁵ e regolazioni della riproduzione¹⁰⁶ (Dancy, 2023, p. 72-73). Secondo questo orientamento, un individuo edotto delle conseguenze indesiderate delle sue azioni, se agisce per principio di volontà portandole a termine, meriterebbe la marginalizzazione e ne sarebbe moralmente corresponsabile. La ricorrenza di ipotesi pregiudiziali infondate, classificazioni su *cluster* e conclusioni in base alle somiglianze è un principio di razzismo: se vale l'analogia tra entità simili e circostanze nel processo di astrazione, il comportamento del singolo sarà predeterminato da quello del gruppo (Numerico, 2021).

Un'inchiesta di ProPublica (Angwin *et al.*, 2016) ha scoperto che la predizione di COMPAS era corretta soltanto nel 61% dei casi in ambito generale e scendeva al 20% nella recidività violenta; la possibilità di essere erroneamente classificato (falsi positivi) ad alto rischio, era più elevata, quasi doppia, per le persone nere. Tempo più tardi uno studio condotto da Dressel e Farid (2018) ha dimostrato che il livello di accuratezza dello strumento non si discostava dall'opinione pubblica generalizzata, priva di formazione sui temi della giustizia penale (62%). Lo strumento integrava fattori avulsi dalla scelta individuale nel determinare la predeterminazione, correlandola a genere, razza e classe sociale.

Le ragioni per cui vengono introdotti *statistical algorithms* sono spesso strumentali, sostiene Dancy (2023, p. 129), perché sono considerati utili a migliorare il nostro stile di vita. Anche impedire a un individuo di prendere una decisione, adottare una condotta o esprimere una preferenza equivale a esprimere un giudizio; di riflesso, l'altro penserà di non essere all'altezza o di non avere i requisiti per essere trattato come un "normale" membro adulto della società, rafforzando differenze di status che escludono intere categorie di persone dall'accesso a benefici. L'utilizzo di questi strumenti è opinabile da più punti di vista: di rilevanza, cioè della coerenza logica tra i criteri, i costi sociali e gli obiettivi; egualitario, sul medesimo trattamento destinato a tutti i membri di un gruppo; predittivo, nella determinazione statistica delle condotte future per cui le decisioni diventano inevitabili. I confini della giustizia penale, in particolare, allocano a loro volta privilegi. L'autrice suggerisce la supervisione umana sulla predizione algoritmica lungo tutto il processo di elaborazione, ricavando informazioni dalla performance alla distribuzione dell'errore, inclusa una valutazione controfattuale dei risultati, ma anche di non ignorare quei risultati in presenza di un sospetto sull'accuratezza (*ivi*, p. 131).

La statistica torna utile in situazioni di incertezza razionale, in assenza di informazioni sufficienti per dedurre una conclusione, in forma di panoramica di un fenomeno. Quando,

¹⁰⁵ Cesaro Lombroso nel 1870 introdusse la teoria atavica positivista di identificazione del presunto colpevole in base a una tassonomia fisica e psicologica.

¹⁰⁶ Leonard Darwin nel 1914 la descrisse come «una classe di criminali al cui seme non sarebbe consentito di contaminare i posterì» (citato da Dancy, 2023).

invece, vengono individuati *pattern* di prescrizione basati sull'incrocio tra l'origine, le opportunità economiche, le condizioni familiari, le condotte dei conoscenti, le esperienze di abuso sessuale, gli algoritmi vengono sistematicamente guidati verso la “diatesi criminale” con cui Charles Goring diagnosticava la correlazione tra il fatto colposo e le circostanze di appartenenza (Dancy, 2023, p. 134).

Il *gold standard* sarebbe, pertanto, ottenere l'evidenza che la *policy* basata su determinati criteri eviti un danno o ne riduca gli effetti sulla popolazione, ma non tutte le ragioni giustificano l'imposizione di sacrifici. Il valore della scelta assume, quindi, un portato significativo per avere o vedersi riconosciuta l'opportunità di incidere sul processo e di modificarlo. Dancy torna sui punti di Sen e Nussbaum (*ivi*, p. 137): la differenza non è data dalle scelte che prendiamo ma dalle alternative che ci vengono rese disponibili e da quanto siamo pronti ad affrontarle. L'adeguatezza delle opportunità dipende dalla posizione ricoperta, dalle condizioni di partenza e dalle garanzie che invoglieranno a prendere buone decisioni, che significa fornire le informazioni contestuali sui rischi e i passi da seguire per evitarli, assicurando che soltanto chi dispone delle risorse incorrerà nel rischio di ricevere un danno.

La puntata “*Joan is awful*” di Black Mirror (2023) riproduce una dimensione distopica in cui Streamberry, popolare piattaforma di *streaming* (parodia di Netflix), adatta in tempo reale, sotto forma di serie tv interpretata da Salma Hayek, eventi reali della vita privata di una utente (Anni Murphy). Mediante una rappresentazione enfattizzata e caricaturale dei fatti, le ripercussioni sulla reputazione di Joan non si lasciano attendere: a causa dello show, il compagno la lascia, le persone perdono stima nei suoi confronti e viene licenziata (la messa in onda dei dialoghi in qualità di manager corrisponde a una violazione degli accordi di riservatezza aziendali). La sua avvocatessa le spiega che la sottoscrizione dei termini di servizio ha concesso alla piattaforma il consenso a trasmettere i dati raccolti dai dispositivi a un computer quantistico (che elabora immagini digitali in CGI, per cui la stessa Hayek ha concesso i diritti di immagine), inclusa la creazione di personaggi e dialoghi a scopo narrativo. Joan prova a causare la rescissione del contratto, commettendo azioni ributtanti per i codici etici televisivi, e scatena la reazione della stessa Hayek che accusa la piattaforma di utilizzo improprio della sua immagine e di ricevere un compenso inferiore rispetto a quanto pattuito. Joan e Hayek si alleano, infine, per distruggere il Quanputer – computer quantistico creatore di multiversi reinderizzati che efficientia i tempi e i costi di realizzazione dei contenuti (volutamente incentrati su momenti di debolezza, egoismo e vigliaccheria per ragioni di *audience*), di cui Joan era un *beta test*. Joan-Murphy, infatti, non è altro che un ulteriore livello di rappresentazione virtuale di una persona ordinaria, Joan-sorgente. Sia Joan che Murphy finiscono ai domiciliari, come garanzia della validità del

contratto con Streamberry, a cui hanno arrecato danni. L'episodio è stato scritto dal creatore di Black Mirror Charlie Brooker e diretto da Ally Pankiw, cogliendo l'occasione per affrontare in anticipo temi quali la titolarità dell'arte, la diffusione di GenAI e la riproducibilità del lavoro nell'audiovisivo, che hanno trovato una sponda nei mesi a seguire nelle rivendicazioni della filiera cinematografica di Hollywood.

5.2 Quale apprendimento?

L'Educazione è uno dei settori maggiormente interessati dal fenomeno e ciò comporta un incremento nella produzione sia di esperimenti che di regolamentazione. L'IA è considerata un mezzo per migliorare la didattica e ha il potenziale per trasformare l'istruzione in modo significativo, agendo su diversi fronti come la personalizzazione, l'automazione e il supporto all'insegnamento. Piattaforme di *e-learning* come Coursera, EdX e Udemy utilizzano l'IA per suggerire contenuti adatti al livello e allo stile di apprendimento di ciascun utente, mentre tutor virtuali sono in grado di assistere in tempo reale studentesse e studenti con spiegazioni su misura. Se si considera l'apprendimento delle lingue, è indubbio l'apporto pratico del NLP in traduzioni simultanee e semi-contestuali che riducono i tempi di traduzione e l'asimmetria linguistica tra enormi platee di beneficiari. Nel 2018 il Ministero dell'Economia e dell'Energia tedesco ha pubblicato il *White Paper on Digital Platforms*, in cui sosteneva che investire nelle istituzioni pubbliche e offrire strumenti con cui sviluppare piattaforme e tecnologie proprie fosse cruciale per colmare il divario di conoscenza che le separa dalle aziende commerciali e mantenere in vita la competitività del settore pubblico (Van Dijck, 2019).

Il 24 gennaio 2025 l'UNESCO ha scelto il tema "IA e Istruzione: Preservare l'*agency* umana in un mondo di automazione" per incorniciare la Giornata internazionale dell'Educazione. La Direttrice generale Audrey Azoulay ha invitato gli Stati membri a investire nella formazione di insegnanti e studenti sull'uso responsabile di questa tecnologia, integrando le dimensioni umane e sociali dell'apprendimento, piuttosto che sostituirle, proprio come i bisogni essenziali.

Il motivo per cui prima la Raccomandazione sull'Etica dell'IA (UNESCO) e poi l'AI Act (UE) hanno prestato attenzione ed enfasi al settore è che la formazione è un momento decisivo per la creazione di aspettative di vita e di mobilità sociale. I divieti introdotti dall'AI Act includono il riconoscimento delle emozioni sul posto di lavoro e nelle scuole, i sistemi di categorizzazione biometrica sulle caratteristiche sensibili (come convinzioni politiche, filosofiche, religiose, orientamento sessuale e razza) e l'identificazione biometrica remota in tempo reale di individui in spazi pubblici da parte delle forze dell'ordine. Inoltre, fornitori e *deployer* hanno l'obbligo di garantire che il personale coinvolto possieda un livello sufficiente di alfabetizzazione, che

presuppone il possesso di competenze, conoscenze, consapevolezza e comprensione adeguate a procedere verso una diffusione informata dei sistemi (art. 4, AI Act), acquisite mediante percorsi di formazione e orientamento.

Ci sono diverse ragioni per voler evitare uno scrutinio algoritmico, tra cui l'isolamento sociale, l'ansia, la perdita dell'impiego, dei sussidi e di un volano di opportunità economiche (Dancy, 2023, p. 89). La valutazione standardizzata è lo strumento più diffuso per determinare gli orientamenti educativi, mentre la modalità di valutazione cambia in ogni paese del mondo, tra prove scritte e orali con diversa ponderazione. Il Regno Unito ricorse a un algoritmo nel 2020 (in sostituzione degli esami di fine anno, annullati a causa del COVID-19) per misurare la *performance* scolastica per il *General Certificate of Secondary Education* e l'*Advanced Level*. Un declassamento del 36% dei risultati – confrontati con i voti dei docenti – portò migliaia di studentesse e studenti a perdere posti utili in graduatoria per l'accesso all'università. Le scuole pubbliche sostennero di essere state penalizzate molto più degli istituti privati (frequentate dal 7% degli studenti), dove coorti più piccole impedivano la normalizzazione (Eccles & Gallardo, 2020). Il presidente di Ofqual, Roger Taylor (2020) scrisse al Parlamento britannico:

«Nei giorni successivi alla pubblicazione dei risultati del livello A, è emerso chiaramente che né le analisi sulle pari opportunità, né la prospettiva di appelli, né l'opportunità di sostenere gli esami in autunno, potevano compensare la sensazione di ingiustizia che uno studente provava quando gli veniva assegnato un voto diverso da quello che lui e i suoi insegnanti ritenevano fosse in grado di ottenere, senza aver avuto la possibilità di sostenere l'esame».

Il saggio "*Jill Watson Doesn't Care if You're Pregnant*" (Eicher, Polepeddi, & Goel, 2018) esamina questioni etiche nell'uso di un assistente didattico virtuale, progettato per supportare gli studenti di un corso online della Georgia Tech. Gli autori mostrano come un approccio empirico – basato su esperimenti sul campo – possa rivelare problemi di trasparenza, fiducia e potenziali pregiudizi nei confronti di gruppi demografici minoritari. Jill Watson, addestrata su dati che riflettono soprattutto esperienze maschili, tende infatti a ignorare problemi specifici di altre categorie, evidenziando la necessità di IA più inclusive.

In Italia, 14 istituti di istruzione superiore di primo e secondo grado tra le regioni Lombardia, Toscana, Lazio e Calabria stanno sperimentando per la prima volta un progetto di esercizi guidati, di cui Google si è fatta carico a titolo gratuito. Alta è l'attenzione affinché non sussistano episodi di discriminazione algoritmica. Alcuni casi giurisprudenziali che hanno riguardato le sentenze n°8472/8473/8474 del Consiglio di Stato sanciscono il diritto dei cittadini alla trasparenza e alla sindacabilità delle regole algoritmiche nel reclutamento scolastico, chiarendo che la decisione algoritmica deve essere ricondotta al decisore umano, che è responsabile di eventuali comportamenti discriminatori (Musella, 2022, p. 18).

La dipendenza tecnologica rischia di ridurre il coinvolgimento diretto dei discenti e di compromettere il valore umano dell'insegnamento, così come di ampliare le disuguaglianze educative tra chi dispone di queste risorse e chi ne è escluso. Un'altra sfida è la sostenibilità: il consumo di risorse/energia prodotta dall'IA espone all'aumento dei costi delle istituzioni scolastiche e universitarie (in particolare elettricità e acqua).

Quando ChatGPT è stato reso disponibile alla mercé del grande pubblico, in molti lo hanno considerato una minaccia per l'integrità e l'indipendenza degli studi accademici. La maggior parte degli studenti sta già usando la GenAI a supporto dei propri studi: per condurre ricerche, per rivedere testi, per scrivere materiali da consegnare, per rispondere alle domande di esame e, infine, per preparare codici. Se decidono di non farlo è per obiezioni morali, per difficoltà di comprensione o per paura di essere puniti in un contesto regolamentare ancora ambiguo. Secondo l'Università di Oxford la tecnologia deve essere guidata dall'istruzione. Se da un lato gli insegnanti riconoscono il potenziale per personalizzare l'apprendimento e alleggerire la pressione sul carico di lavoro, dall'altro intravedono anche il rischio di mancato riconoscimento degli errori da parte degli studenti, il plagio e l'eccessiva dipendenza dai dispositivi (Oxford University Press, 2023). Anche il rapporto UNESCO “*AI and Education*” (2021), indirizzato ai decisori politici, rimarca che l'IA non dovrebbe essere utilizzata se mina le competenze cognitive e sociali acquisite attraverso osservazioni della vita reale, interazioni con altri esseri umani e sperimentazione pratica.

L'Università di Pisa ha deciso di introdurre lo strumento nelle proprie attività ordinarie mediante una collaborazione formale con OpenAI e l'acquisto della licenza ChatGPT Edu, prima in Italia ma in compagnia già di Harvard e Oxford. Leah Belsky, General Manager per l'Educazione di OpenAI, per l'occasione ha affermato di non vedere l'ora di «supportare il *team* per offrire un apprendimento più personalizzato, nel promuovere scoperte nella ricerca e nell'aiutare a preparare gli studenti per un futuro orientato all'IA»¹⁰⁷.

Non sappiamo, infine, quanto il desiderio di automatizzare usurperà o muterà il desiderio di conoscere.

5.3 La disobbedienza del *glitch*

Nel saggio “Disobbedire”, Frédéric Gros si domanda quali sono le ragioni dell'obbedienza dinanzi a inaccettabili ingiustizie sociali e ambientali. Scrive che la democrazia rappresenta «una tensione etica nel cuore di ciascuno, l'esigenza di reinterrogare la politica, l'azione

¹⁰⁷ Università di Pisa (24 ottobre 2024). Intelligenza Artificiale: alleanza tra Università di Pisa e OpenAI, *Unipinews*, <https://www.unipi.it/index.php/news/item/28885-intelligenza-artificiale-alleanza-tra-universita-di-pisa-e-openai>

pubblica, l'andamento del mondo a partire da un *sé politico* che racchiuda un principio di giustizia universale [...]» (Gros, 2019, p. 9), e vano se non sostenuto dalla collettività in un progetto di futuro. Mentre gli esperti si esprimono a partire da statistiche anonime, disobbedire è una dimostrazione di umanità.

Nel settembre 2023, a seguito di uno sciopero durato cinque mesi, per la prima volta nella storia è stata ottenuta e ratificata un'istanza di regolazione dell'IA a opera della *Writers Guild of America*. Le proposte dei lavoratori non cercavano di impedire l'utilizzo dei sistemi ma si concentravano sulla protezione degli sceneggiatori dai danni economici, consentendo loro la libertà di utilizzo in una visione ottimistica che riteneva l'IA limitabile/sfruttabile per fini strumentali (Maddaus, 2023). In una fase iniziale, gli *studios* avevano escluso l'idea di dover rinunciare all'IA, posizione non smentita dalla *Alliance of Motion Picture and Television Producers* che sedeva sui tavoli di contrattazione. Il *Memorandum of Agreement* (MOA) 2023/26 contiene un capitolo interamente dedicato ai sistemi di IA in ambito creativo e audiovisivo (Writers Guild of America, 2023), in cui si stabiliscono alcuni principi di base: la GenAI non può scrivere o riscrivere materiale artistico e tale materiale non sarà considerato originale, ovvero non può essere utilizzato per compromettere la titolarità e i diritti di un autore; uno sceneggiatore può scegliere di utilizzare l'IA durante la stesura, previo consulto con l'azienda e a condizione che vengano rispettate le *policy* applicabili (a titolo di esempio, programmi che individuano potenziali violazioni del *copyright* o plagii), viceversa l'azienda non può richiedere a colui/colei che scrive di utilizzare *software* di IA (come ChatGPT) quando svolge le sue mansioni; l'azienda deve comunicare allo sceneggiatore se i materiali forniti sono stati generati integralmente, o in parte, da sistemi di IA; WGA si riserva il diritto di affermare che il materiale coperto dall'accordo non può essere usato per allenare l'IA.

Oltre alla flessibilità nell'uso dell'IA come strumento, l'accordo non contiene limiti alla quantità di testi redatti da IA che un autore può incorporare, restando in attesa di un aggiornamento della norma sul diritto d'autore (Maddaus, 2023). Le negoziazioni sindacali per la *Screen Actors Guild-American Federation of Television and Radio Artists* (SAG-AFTRA, 2024) sono, invece, proseguite fino alla firma del governatore della California Gavin Newsom di una legge che estende i diritti anche ad attrici e attori, dove l'AB 2602 richiede il consenso informato degli artisti prima dell'uso delle loro repliche digitali, mentre l'AB 1836 fornisce una protezione aggiornata ai diritti di voce e immagine degli artisti deceduti.

L'attrice Scarlett Johansson, doppiatrice dell'IA nel film *Her* (2013), ha accusato OpenAI di aver indebitamente utilizzato la sua voce¹⁰⁸ per il nuovo assistente vocale di ChatGPT. Contattata dall'azienda un anno prima (e pochi giorni prima del lancio), l'attrice aveva declinato l'offerta per motivi personali. OpenAI ha respinto l'accusa, sostenendo che un'altra interprete avesse prestato la voce, e ritirato il prodotto. SAG-AFTRA ritiene che questo sia un momento di insegnamento per l'industria tecnologica e appoggia il *No Fakes Act*, al vaglio al Senato USA (Milmo, 2024). In Tennessee, la legge *Ensuring Likeness Voice and Image Security* (ELVIS) Act protegge i doppiatori dallo "sfruttamento sleale delle loro voci", inclusi i cloni da GenAI, ma gli esperti legali richiamano la difesa della parodia, come forma di libertà di espressione, nella stesura dei provvedimenti legislativi (Jones, 2024).

"*Algorithms of Resistance*" di Bonini e Treré (2024) legittima le piccole esperienze di resistenza contro il potere delle Corporation digitali. Il libro esplora l'evoluzione delle pratiche che hanno modellato il nostro rapporto con app e media digitali, alcune di queste diventate normali forme di raggirio, come il confronto con *scam*, *hate speech* e linguaggio molesto, e l'adozione di strategie per evitare la penalizzazione algoritmica, adattando parole e simboli per poter continuare a comunicare efficacemente. Spesso descritti come passivi e sottomessi alle decisioni, gli utenti riacquisiscono il ruolo attivo di interazione precedente ai social media, dove l'esperienza si articolava nella creazione e produzione del contenuto in dialogo tra codici *HTML* e forum. Venuta meno la componente tecnica di intervento sul codice, anche chi non è alfabetizzato ai linguaggi di programmazione può intuire come agire: la capacità di elaborare attività volte a eludere un sistema di controllo, formando una scissione netta tra chi prende le decisioni e chi trova gli strumenti creativi per aggirarle, è una dimostrazione dell'*agency* degli utenti. La resistenza, scrivono gli autori (*ibidem*), diventa attrito e si distingue dalla ribellione. Nel caso dei rider, chi agisce rifiuta un ordine non redditizio, eterodiretto (lavora su più piattaforme) e schematizzato (lavora con ordini falsi). Culpepper e Thelen (2020) sostengono che il concetto classico di potere strumentale e strutturale non colga efficacemente l'influenza che Facebook, Google e Amazon esercitano nell'economia, mettendo in campo mobilitazione, fiducia e dipendenza. Alcune persone finiscono per allinearsi a ciò che le Corporation vogliono, diventano persino sostenitrici tacite o attive di decisioni che possono danneggiare i loro interessi materiali. Al contempo, i consumi definiscono le preferenze politiche e si intersecano con conflitti più antichi, come quello tra lavoro e capitale (*ibidem*). Le relazioni industriali

¹⁰⁸ Ricorrendo al "diritto di pubblicità", esiste un precedente nella causa "*Bette Midler vs. Ford Motor Company*" che risale al 1988. L'azienda aveva ingaggiato una voce-sosia per uno spot pubblicitario e la cantante vinse in appello. Nel 1992, la conduttrice Vanna White fece causa alla Samsung, che aveva vestito un *robot* come lei in una pubblicità. Anche la replica non esatta è perseguibile per "appropriazione" (Jones, 2024).

differiscono a seconda del Paese e la presenza di condizioni ostili si riversa nella possibilità di adesione a forme associative tradizionali (sindacati) e alternative (*alt-labor*). Negli Stati Uniti, per citare l'esempio di Amazon, il diritto alla rappresentanza sindacale può essere realizzato solo attraverso un voto a maggioranza, che è difficile ma non impossibile da ottenere, come dimostrano le catene di solidarietà a Bessemer o a Staten Island e l'emersione di realtà come Amazon Global Alliance (UNI), il movimento Transnational Social Strike (TSS) con Amazon Workers International (AWI), Amazon Employees for Climate Justice (AECJ) e le associazioni come Former and Current Employees (FACE) of Amazon, ispirando la nascita dell'Alphabet Workers Union (AWU, o Google Union) nel 2021 (Kassem, 2023, p. 157).

Inoltre, non va escluso che persone, gruppi sociali e istituzioni possano decidere di allearsi con gli algoritmi (*Algorithmic alliance*) per raggiungere obiettivi materiali/sociali o di interrompere ogni cooperazione ed eluderli: Bonini e Treré propongono l'immagine del tango (2024, p. 56), per descrivere questa tensione ("*It takes two to tango*" sottintende che nella disputa la responsabilità è di entrambe le parti), in cui emergono differenti gradi di intimità, che possono essere effimeri o temporanei; a differenza del passo a due, tuttavia, l'ambivalenza è meno chiara e più asimmetrica. All'indomani del 7 ottobre 2023, con l'inasprirsi degli attacchi in Palestina, non soltanto gruppi, organizzazioni e movimenti sociali, ma ciascuno - nell'esperienza individuale di fruizione ed esposizione mediante account - ha iniziato ad adottare condotte molto simili, condividendo un immaginario algoritmico, dalla proposta di contro-narrazione dei fatti di Gaza all'alternanza tra numeri e lettere (P4l3st1n4), fino all'elezione del cocomero a codice identificativo della bandiera, per evitare che la piattaforma attuasse *shadowban*.

Non è la prima volta che si osserva come siano le voci dissonanti quelle più abili nell'individuazione dei meccanismi di raccomandazione. Agire come l'altro ci rende parte di un soggetto collettivo, diffondendo una nuova consapevolezza sui sistemi automatizzati. La "classe operaia" delle piattaforme sembra condividere le sorti di contadini e artigiani dell'epoca della prima rivoluzione industriale, come descritto nel Capitolo 1. *Creator*, rider, attivisti, blogger, autisti, musicisti e consumatori sono stati tutti indistintamente sussunti dalle logiche delle piattaforme (Bonini & Treré, 2024, p. 123). In passato, i *creator* di YouTube lamentarono la demonetizzazione non giustificata: si accorsero che il programma *YouTube Partner Program* (YPP), che prometteva di condividere i ricavi pubblicitari, stava imponendo regole e strutture che favorivano i grandi canali rispetto a piccoli e nuovi utenti, marginalizzati, spesso a causa di argomenti politici o controversi, e puniti per contenuti non conformi alle linee guida "adatte agli inserzionisti", nel tentativo di massimizzare la distribuzione di ciò che il pubblico desidera e genera profitto (Caplan & Gillespie, 2020).

Un altro caso meritevole di menzione è quello delle persone impegnate in attività di *sex working* che hanno accusato le piattaforme di *shadowban* e sono state costrette a raccogliere evidenze empiriche dai propri profili per fornire prove (Are, 2022).

Gli attivisti esercitano pressioni sugli inserzionisti per fermare i finanziamenti ai gruppi d'odio, le agenzie di marketing inseguono il consumerismo etico scegliendo in maniera ragionata siti, gruppi e canali per evitare danni alla propria reputazione, mentre i ricercatori sviluppano indicatori di trasparenza per valutare la moderazione dei contenuti.

La società civile sta contribuendo attivamente alla moderazione e alla regolamentazione dei contenuti sulle piattaforme digitali. Ciononostante, le strategie e i mezzi d'azione rimangono ancora poco studiati, soprattutto perché i canali di accesso sono spesso sconosciuti.

Conclusioni

Tutte le piattaforme moderano e non esisterebbero senza questa complessa operazione sociotecnica, scrive Tarleton Gillespie (2018). Spesso motivate da ragioni di gestione della complessità, le Corporation digitali compiono decisioni etiche di allargamento o restringimento della libertà di espressione. Lo Stato incontra in questa contingenza un elemento di crisi della sua sovranità, storicamente posta su un *continuum* tra autorità e potere nella gestione dell'ordine e nel respingimento delle ingerenze.

Giovanni Sartori (1995 [1987]) si è occupato delle “Tecniche decisionali” per spiegare la differenza tra la decisione individuale, monocratica, e la decisione collettiva, che presuppone la partecipazione di molti alla discussione. Quando una decisione viene sottratta alla competenza dei titolari qualificati, e chi decide lo fa per gli altri, parliamo di decisione collettivizzata. «Il progresso tecnologico», diceva, «porta *naturaliter* con sé un ampliamento della collettivizzazione, dal momento che comporta servizi pubblici e beni collettivi» (*ivi*, p. 381). Ciò avviene per pura convenienza, secondo un “calcolo del consenso” che deriva dal criterio di massimizzazione dell'interesse di matrice economica. Nel mondo occidentale contemporaneo anche la politica rincorre vantaggi immediati a discapito dell'etica e del dovere, intendendo il bene comune come parte di questo scambio.

Così come il politologo a suo tempo, questa ricerca si chiede dove sta il potere e come si formano i procedimenti che ne disciplinano l'esercizio. La moderazione automatizzata dei contenuti segue questa linea di demarcazione perché, nell'ambito di applicazione, la decisione sancita è vincolante nei confronti di un'intera collettività, così come lo sono tutte le decisioni sovrane, dunque sovraordinate, *erga omnes* e corredate di “forza cogente”, la cui trasgressione presuppone sanzione. Sebbene condividano ampio terreno con le decisioni politiche, nel digitale nessuna di queste appare negoziabile.

Sartori scriveva che la convenienza trova limiti nelle esternalità, i costi di oppressione (ingiustizia, coercizione, privazione) e di spreco (incompetenza, inefficienza) che i terzi subiscono indipendentemente dalla loro volontà, e nei costi decisionali come il tempo, la fatica e la produttività, dati dal numero di persone che partecipano alla decisione, le regole elettive (successione, sorteggio, elezione a tempo determinato o a vita) e di voto (unanimità, simultaneamente, in sequenza). Questo indice di rischio equivale al calcolo di interdipendenza sociale, a cui le democrazie hanno storicamente risposto con la rappresentanza politica. Come in ogni situazione decisionale discontinua, il caso in analisi sembra coincidente con una

frammentazione che manca di qualsiasi interazione di gruppo e di costruzione «con reciproche persuasioni e transazioni» (Sartori, 1995 [1987], p. 397).

Come vengono prese le decisioni e quali ricadute avranno sulle persone sono le domande che si pone chi ha a cuore il concetto di giustizia. Seguendo Cammaerts e Mansel (2020), le deliberazioni democratiche necessiterebbero di un pluralismo radicale per valorizzare gli interessi pubblici. L'introduzione di sistemi di IA per l'efficientamento della filiera del pensiero e della sua condivisione è, invece, un campanello d'allarme tanto quanto lo slittamento di responsabilità dal pubblico al privato (i "governatori primari" in Berg & Hofmann, 2021) per errori prodotti dai processi decisionali automatizzati.

L'osservazione dei valori collettivi e la declinazione che subiscono per le scelte compiute dai CEO risultano interessanti per approfondire questo frammento di *design* del digitale: superando l'approccio funzionalista, la ricerca ha esplorato il campo per comprendere se la libertà di espressione sia diventata il margine conteso di sovranità tra pubblico e privato. L'analisi qualitativa comparata sulle Corporation digitali è stata condotta sotto la lente della regolamentazione UE, considerata come variabile indipendente. L'Europa è arbitro nello scacchiere internazionale del digitale, svincolata da interessi di parte ma sempre più dipendente dai loro servizi (Nieminen *et al.*, 2023). Inoltre, l'intervento pubblico contribuisce a invertire la rotta in un settore storicamente interessato da *self-regulation*.

Il periodo di osservazione ha riguardato variazioni e implementazioni delle *Digital Private Policy* avvenute tra il 2018 e il 2024. Meta (Facebook e Instagram), X (già Twitter) e ByteDance (TikTok) rientrano nella classificazione delle *Very Large Online Platform* (VLOP) secondo il *Digital Services Act* e, pertanto, sono tenute alla pubblicazione semestrale dei *Transparency Report* per tenere traccia delle attività di gestione.

Il punto di partenza è stato il problema di Ganton, che invita a chiedersi come mai società lontane, con culture politiche dissimili, ricorrano alle medesime strutture organizzative. La gestione proprietaria, con i dovuti distinguo, è un tratto comune delle piattaforme di social media, ma la specificità di ciascuna è la leadership: mentre in Occidente il CEO è in prima linea nella promozione del marchio, ByteDance adotta un approccio più implicito, a cui partecipa il governo cinese. Le strategie adottate sono emanazione di questa *governance* aziendale: Meta persegue una missione universale di promozione valoriale; X si considera un *player* nazionale fedele al Primo Emendamento USA; ByteDance adotta un approccio regionale in dialogo con le culture locali, assegnando a ciascuna area un *manager* (in Europa, l'*EMEA Regional Policy Team*). Una ulteriore similitudine è proprio la decisione di gestire la complessità e la mole di contenuti mediante sistemi automatizzati.

L'analisi multilivello della documentazione prodotta dalle Corporation digitali ha permesso una classificazione degli orientamenti gestionali. Meta e ByteDance hanno implementato una moderazione proattiva di account e/o contenuti e reattiva, quando scaturita da utenti e/o autorità, che revisiona e rimuove in caso di violazione dei Termini di Servizio. X, al contrario, difende una moderazione contestuale che preferisce restringere la platea piuttosto che eliminare definitivamente il contenuto.

L'ampiezza dei destinatari di tutela è misura della libertà di espressione, così come configurata nelle liberal-democrazie, dove è previsto il bilanciamento delle libertà individuali con la protezione dei diritti altrui (reputazione, privacy, *hate speech*) e la sicurezza nazionale: su Meta è il pubblico, su X è l'individuo. Meta la considera esattamente al pari della sicurezza, concretizzandosi nel processo di informazione tra piattaforma e utente, che è identico sia nello scambio sociale che commerciale. X costruisce sul tema uno *storytelling* che, dalle mobilitazioni globali alla condivisione individuale di meme, ricorda il ruolo svolto dal social media nella diffusione delle opinioni. Per ByteDance l'espressione è creatività, intrattenimento e divulgazione, dove in primo luogo è il *creator* a compiere l'azione.

Seguendo l'approccio della *Data Justice* (Dencik *et al.*, 2019), si riconosce responsabilità alla tecnologia nella produzione di rappresentazioni distorte della società. La rilevazione di corrispondenze e somiglianze, infatti, determina una categorizzazione discrezionale di non conformità con le *policy* (Termini di Servizio e Condizioni d'uso) che ha impatti dirompenti sia sulla fruizione del prodotto che sulla vita reale del soggetto interessato. Le scelte liminali delle Corporation digitali sono veri e propri esercizi retorici che trovano nella moderazione automatizzata un più efficace compimento per vagliare, classificare, filtrare e bloccare i contenuti, creando processi gerarchici di selezione e raccomandazione. Le persone non possono concordare o emendare questi documenti, impropriamente definiti accordi (*user agreements*) e meglio configurabili come contratti unilaterali. Gli utenti sperimentano da tempo sulla propria pelle le discriminazioni algoritmiche causate dai sistemi di moderazione: ai destinatari spetta un *disclaimer* informativo, a chi pubblica e/o condivide soltanto una percezione. Rimozione, *shadowban* ed etichette sono entrati a far parte della quotidianità di ciascuno sebbene le strategie censorie non venissero notificate, mentre ora la motivazione contenente i fatti, le circostanze, l'origine, la portata territoriale, la sua durata e le condizioni di ricorso per le restrizioni è prerogativa dei prestatori di servizi (art. 17 DSA). Cosa sia o meno un "contenuto illegale" (art. 3 DSA) resta un aspetto centrale che dimostra quanto l'interoperabilità abbia un fondamento nella definizione semantica della regolazione.

I *Transparency Report* non esplicitano lo *shadowban*, ma viene contemplata la “retrocessione” come misura di influenza sulla disponibilità del contenuto illegale: Meta consente ai suoi sistemi di IA di prevedere l’idoneità, attribuendo un punteggio inferiore e, dunque, mostrandolo di meno nel *feed*. Ciò suggerisce che ai modelli di apprendimento automatico venga affidata l’individuazione della notiziabilità e l’azienda afferma che la riduzione avviene anche in assenza di accertata violazione. In tempi più recenti, la stessa moderazione, implementata in nome della sicurezza (Gillespie, 2022), è diventata alquanto controversa e il *Political Content Approach* di Meta sembrerebbe un modo per evitarla del tutto, de-responsabilizzando la Corporation digitale dalla complessità dei contenuti diffusi. TikTok riduce la visibilità se il video può generare turbamento o se viola la legge del paese in cui viene pubblicato (quando non rimosso), a partire dai risultati di ricerca suggeriti e nella distribuzione *Per te*. X distingue i contenuti in base all’impatto sulla platea, restringendo il pubblico o mostrando un’etichetta. La sospensione dell’account è prevista sia in caso di diffusione di contenuti illegali che di recidività nelle segnalazioni.

Una ulteriore prospettiva sul fenomeno osservato è fornita dal diritto d’appello. I concetti di giustizia e legge hanno storicamente trovato un punto di incontro nell’indipendenza delle corti e nell’imparzialità delle autorità amministrative, soprattutto in settori sensibili e ad alto contenuto tecnico. Secondo l’art. 20 DSA, i destinatari dei servizi (persone o enti) hanno il diritto di rivolgersi a qualsiasi organismo per la presentazione di un reclamo, inclusi i sistemi interni e gli organismi di risoluzione extragiudiziale delle controversie (art. 21) che, entro 90 giorni, provvedono a un riscontro imparziale, efficace e indipendente. L’utente deve poter accedere a un sistema efficiente di gestione del reclamo (per segnalazione e/o notifica di contenuto illegale), nonché libero nell’accesso e gratuito. Il DSA chiarisce che il fornitore dei servizi non deve avvalersi esclusivamente di sistemi automatizzati in questo ambito, ovvero obbliga le piattaforme alla supervisione di personale qualificato. La Commissione europea ha reso disponibile la piattaforma ODR per gli utenti del mercato europeo (inclusi Norvegia, Islanda e Liechtenstein), disponibile nelle rispettive lingue e a disposizione dei consumatori per confronti digitali con professionisti del settore. Le informazioni sul numero di controversie, i tempi di risoluzione e il numero di sospensioni sono condivise nei report semestrali.

Le Corporation digitali condividono la pretesa di pronuncia, delegando a terzi il parere ma rimanendo all’interno della piattaforma. L’*Oversight Board* di Meta ha rappresentato un vero e proprio caso di studio sull’argomento. Sebbene sia un organismo indipendente, il suo rapporto con l’azienda madre trasmette non poche ambiguità, sia sul procedimento di selezione e retribuzione dei suoi membri che sulla conseguente percezione d’urgenza legata ai singoli casi.

È il Consiglio, infatti, a stabilire i criteri di ammissibilità degli appelli, inclusi i livelli di gravità e priorità per la riapertura del caso. Entro 90 giorni la decisione, a cui partecipa almeno un membro di giuria originario della regione coinvolta, viene motivata e pubblicata e diventa, dunque, un contributo alle linee interpretative delle *policy*. Ciò entra in conflitto con la rapidità e la modalità unilaterale con cui l'azienda modifica le singole voci delle sue condizioni di servizio (in media 41 volte in un anno), sebbene l'aumento più significativo sia stato registrato proprio a seguito delle sollecitazioni del MOB, mediante raccomandazioni. Gli Standard della Community che hanno subito maggiori revisioni hanno riguardato la violenza, beni e servizi illeciti, l'incitamento all'odio e le immagini sensibili. Tra il 2020 e il 2024, soltanto un centinaio di casi (a fronte di milioni di richieste) hanno meritato la revisione secondo gli indicatori, ma il dato significativo è che 8 volte su 10 il MOB ha ritenuto errate le decisioni algoritmiche e/o umane. I contenuti rimossi appartenevano ad aree sensibili quali incitamento all'odio, terrorismo e violenza. È curioso osservare che l'azienda è intervenuta per emendare le voci sul discorso d'odio e l'istigazione alla violenza quasi in concomitanza con le osservazioni del Comitato, mentre sul Capitolo "Persone e organizzazioni pericolose" le variazioni sono state minime. L'agenda dei casi rilevanti, che sia per scelta o per mole di contenuti pubblicati, riflette il momento storico e lo smistamento dell'opinione pubblica sui fatti di attualità. Lo dimostrano i numeri disponibili già nei primi mesi del 2024, dove 5 su 22 decisioni globali riguardavano la Palestina, a seguito dei fatti del 7 ottobre 2023.

X si calibra in base alle attitudini e ai gusti degli utenti, che possono sottrarsi (*opt out*) o segnalare per violazione al Centro assistenza su fattispecie specifiche: uso non autorizzato del marchio, diritto d'autore, beni contraffatti, privacy dei minori, pedofilia, pornografia, furto di identità, condivisione di dati sensibili, violenza, *spam* e norme pubblicitarie. L'applicazione dell'etichetta per contenuti sensibili non comporta automaticamente la restrizione o la rimozione perché la piattaforma preferisce decentralizzare l'esperienza (riprendendo la strategia FoSNR). In caso di richiesta legale di informazioni o rimozione (valutando anche l'ipotesi *geoblocking*), l'esame preliminare avviene per accertare in primo luogo che non sussista limitazione della libertà di espressione nei confronti di chi ha pubblicato il contenuto e l'azienda può rifiutarla se riscontra vaghezza, inconsistenza e/o vizi di forma, comunicando all'utente la ricezione salvo impedimenti determinati dalla condizione di emergenza (come il pericolo di vita, lo sfruttamento sessuale di minori ed episodi di terrorismo) per favorire una conciliazione tra le parti che escluda implicazioni per la piattaforma. Il DSA (artt. 17, 20 e 21) ha permesso l'accesso a un'apposita sezione per presentare ricorsi, ma il reclamo può essere

presentato anche presso tribunali europei. Tuttavia, X si dichiara non vincolata alle decisioni degli organismi certificati per la risoluzione extragiudiziale delle controversie.

Per ByteDance, la libertà di espressione si bilancia con la prevenzione del danno, resa possibile grazie al coinvolgimento di utenti, professionisti in sicurezza e salute pubblica e dei Consigli consultivi regionali formati da circa 50 esperti. Non raggiunta la conciliazione, le parti possono adire ai tribunali o alla piattaforma di ODR. Già in “fase di controllo” i contenuti vengono sottoposti al dipartimento *Trust & Safety* per una revisione e la decisione è consultabile dagli Strumenti; chi segnala può indicare anche il Paese in cui ritiene che il contenuto sia illegale o, se del caso, la *policy* violata, che ne determinerebbe la rimozione a livello globale. L’utente può inviare sia un *feedback* di miglioramento della moderazione automatizzata che un ricorso. La decisione può essere impugnata una sola volta ed entro 180 giorni dalla notifica, tramite l’apposito modulo argomentando la motivazione. La recidività del ricorso è punita con la sospensione. I minori devono essere accompagnati dai genitori o dai tutori in tutte le fasi del procedimento di verifica dell’identità. Enti pubblici, personaggi e partiti politici sono utenti che godono di restrizioni più lievi a causa dell’interesse pubblico dei contenuti postati, mentre la promozione della violenza o della disinformazione comporta l’interdizione dalla pubblicazione nell’arco di 7/30 giorni. L’azienda non esclude di intraprendere azioni legali al di fuori della piattaforma.

Le tre distinte strategie di moderazione dei contenuti hanno condotto la ricerca verso una ulteriore analisi degli impatti sociali. Foucault spiega che il controllo del discorso avviene per scongiurare da poteri e pericoli e ha l’effetto di determinare chi può parlare, su quale argomento e come le altre persone dovranno interpretarlo. La difesa dell’ordine sociale finisce per creare categorie distinte tra razionale/irrazionale, dentro/fuori e, nel nostro caso, lecito/illecito, che equivalgono a gerarchie di repressione. Le forme di governamentalità (Bazzicalupo, 2015) altro non sono che il controllo politico sulle condotte delle persone che, coincidendo con la nascita del capitalismo, privilegiano il padroneggiamento dell’evento e del caso, in cui l’identità è una costante delle regole. La raccolta dei dati per ciascun utente, e la conseguente categorizzazione, avviene *by design* e secondo le logiche strutturali della piattaforma, costruite per contenere idealtipici *buyer personas* che non corrispondono alla performatività del *bios*. La moderazione dei contenuti è, dunque, una forma di governo della vita che protende verso la conformità ai principi di legge, del marchio e dei ricavi.

Di solito le persone si accorgono di interagire con una tecnologia quando si manifesta un malfunzionamento, ma secondo la logica appena presentata la relazione tecnopolitica con l’identità trasforma gli utenti non conformi e non binari in un *glitch*. Alcune decisioni

segregano, escludono e nascondono l'anomalia, che appartiene a ciascuno in quanto esseri desideranti e non-finiti. Come avviene questa standardizzazione? Attraverso le *policy*, redatte dalle Corporation digitali sotto forma di Standard della Community, Linee Guida e Condizioni d'uso, come norme di convivenza tra pari.

Sulla base di 229 documenti di *policy*, raccolti tra il 24 e il 28 giugno 2024 e caricati nel *software* NVivo, è stato possibile codificare gli *item* in base a categorie tematiche, iterare temi e parole chiave con strumenti di *query* e analisi delle frequenze e raccogliere in Nodi i risultati. Alcune fonti secondarie, tra cui articoli, report e inchieste, hanno contribuito a una più ampia interpretazione dei dati raccolti, nel tentativo di offrire una lettura intersezionale del fenomeno osservato. La ricerca ha prodotto 1726 *references* di Genere (presente nel 55% dei documenti), 48 di Razza (7%) e 216 di Classe (39%).

Il consenso, in ambito tecnologico inteso come “consenso informato” del soggetto interessato, è il termine che definisce l'autodeterminazione, la disposizione degli attributi personali e i termini del rapporto sessuale. Le occorrenze includono risultati associati all'espressione della volontà, della sessualità e del corpo. L'incitamento alla violenza e la minaccia di stupro sono vietati e sanzionabili con il *ban* su Meta, X e TikTok.

Su Instagram (Meta) è penalizzata la nudità e l'esposizione di parti intime, citando tra i genitali i capezzoli femminili, la cui rappresentazione è culturalmente sessualizzata, a esclusione di immagini di: allattamento, parto, mastectomia, sensibilizzazione sul cancro al seno e interventi chirurgici di transizione; atti di protesta; dipinti e sculture. Nessuna tolleranza per contenuti di natura sessuale che raffigurano minori o casi di *revenge porn*. L'offerta di servizi sessuali, anche quando è supportata dalla domanda, non è consentita, a discapito delle persone che lavorano come *sex workers*.

Facebook (Meta) ha interdetto i suoi servizi a persone condannate per reati sessuali, inasprendo le sanzioni per allusioni nel contesto pubblicitario. Nella versione inglese del Capitolo “Sicurezza”, si rivolge al mondo del *sex working* assicurando una sensibilità sul tema. Nelle *policy* la “vittima” diventa “sopravvissuto” e la rimozione di immagini intime non consensuali (NCII) è prevista soltanto dietro segnalazione. Un sistema di *photo-matching* rileva proattivamente la corrispondenza di contenuto e un dipendente ne controlla i risultati, mentre l'utente può avviare una procedura di appello. In passato, scarse segnalazioni erano dovute ai tempi e alle risposte fornite da Meta, come d'altronde avviene anche nel mondo reale, dove la denuncia viene scoraggiata dalla percezione di supporto inadeguato. Persino il racconto di condanna passa al vaglio della piattaforma, che può decidere di applicare un *disclaimer* come

contenuto sensibile (come avviene anche in caso di video che ritraggono l'immersione violenza in acqua di minori durante riti religiosi).

La promozione di eventi a sfondo erotico è consentita, ma non la pubblicazione di video, performance e messaggi tantrici. L'educazione sessuale è consentita, a meno che non si traduca in adescamento (eccetto la penetrazione, l'autoerotismo e scenari fetish).

X preferisce mettersi in contatto con le persone coinvolte, anche tramite mediazione, prima di sanzionare e la vittima può segnalare fino a cinque post. Gli utenti selezionano l'etichetta di contenuto sensibile per segnalare la presenza di contenuti violenti, pornografia e altre forme di contenuti per adulti che, se prodotti in modo consensuale, sono tollerati dalla piattaforma (strategia FoSNR) e consentono ai sistemi di raccomandazione di non mostrare il contenuto ai minori di 13 anni (o 16 in Australia, dove l'*eSafety Commissioner* che subentra qualora i provvedimenti della piattaforma siano insufficienti). Il controllo proattivo non è previsto *by design*, garantendo agli utenti la scelta ultima di pubblicare un contenuto e i moderatori intervengono quando un contenuto sensibile viene reiteratamente pubblicato privo di etichetta.

Nel 2024, la gestione dei *Deepfake* sessuali raffiguranti il volto di Taylor Swift sul corpo di una attrice è risultata tardiva e insufficiente dal punto di vista della prevenzione del rischio. La condivisione e diffusione di contenuti sintetici è accettata su TikTok (ByteDance) solo con il consenso della persona raffigurata, nel rispetto del diritto d'autore e della privacy. Secondo la strategia regionale, ciò non vale in Giappone dove, sulla base delle norme locali, la promozione pubblicitaria di contenuti a sfondo religioso, sessuale, con riferimenti a genitali (e dunque anche coppette mestruali e prodotti per la salute riproduttiva) o destinata al pubblico adulto è vietata.

I capi di abbigliamento variano in base all'area interessata: una donna che veste abiti a maniche corte è inaccettabile in MENA 2, ma non in MENA 1; la biancheria intima di marca può essere promossa in Europa, ma non in Corea se indossata da un/a modello/a; i prodotti e servizi contraccettivi, così come la prevenzione o interruzione di gravidanza, sono vietati in diversi Paesi, tra cui Nord e Sud America, Australia e Nuova Zelanda.

La piattaforma di Elon Musk, invece, impedisce la vendita di prodotti *cause-based* (per la sensibilizzazione su cause sociali come impegno civico, ambiente, equità sociale) e accessori con contenuto politico, ponendoli sullo stesso piano di armi, alcool e tabacco.

Le manifestazioni dell'identità sessuale sono contemplate da X, secondo la *policy* di tutela contro l'uso improprio dei pronomi e del *deadname* con sanzione di riduzione di visibilità. La piattaforma si riserva di utilizzare parole non segnate dal genere in ambito professionale. Alcuni anni fa, Facebook aveva sperimentato un caso di discriminazione con la *Real Name Policy* che

impediva alle Drag Queen e agli attivisti la libera fruizione, la promozione degli eventi e delle campagne di *fundraising*, rendendoli perseguibili.

Si ritiene che la *policy* su “Incitamento all’odio” (Contenuti Deplorevoli, Standard della Community) di Meta sia stata redatta con un approccio abilista che marginalizza specifiche categorie di persone e ne discrimina altre (così come definite con “carenze mentali” o “sulla base delle caratteristiche o dello status di immigrato”), moralizza l’attività di *sex working*, disumanizza con termini di paragone inimmaginabili in una norma dello Stato, basate su fattispecie generali e astratte su cui è il giudice a individuare l’applicazione al caso concreto. Le caratteristiche elencate in questa formula costituiscono forma di aggravante per chi pubblica, ma il risultato è una esplicita divisione sociale basata sul culto, sull’estrazione e sulle origini, accompagnata da una duplice vittimizzazione: l’enumerazione di chi differisce dal “normale” e la circoscrizione in una categoria identitaria protetta. Inoltre, l’espressione “esseri subumani” e la traduzione di “*Black people*” in “Persone di colore” appaiono inappropriate. Esempi simili sono stati riscontrati su X dove i chiarimenti linguistici che coinvolgono le forze di polizia citano “neri”, “ispanici” e “donne asiatiche”.

Proprio su quest’ultimo appunto bisogna ricordare che la discriminazione linguistica è all’ordine del giorno nella moderazione dei contenuti sulle piattaforme di social media, dove l’inglese e le lingue del Nord Globale sono implementate universalmente anche nei servizi di supporto, causando fenomeni di sottorappresentazione (assenza di *diversity* nel NLP) e incapacità interpretativa dei sistemi di IA nel cogliere le sfumature (dove, anche in questo caso, fa eccezione ByteDance). Nel caso di X è particolarmente evidente perché i moderatori lavorano su 16 mercati con un volano di 5 lingue a disposizione (a fronte delle 80 lingue dichiarate da Meta), dove anche il mancato reclutamento di personale madrelingua esacerba l’inaccuratezza della moderazione.

Ciò ha comportato l’analisi su una ulteriore discriminazione: la classe sociale che soggiace alla moderazione umana a supporto dei sistemi automatizzati. Da ultima rilevazione, Meta dichiara di avere 15.000 addetti (di cui 2.500 “agnostici”), a fronte dei 6.354 (in UE) di ByteDance e dei 1849 di X. Stando ai documenti, le aziende forniscono un percorso di formazione e supporto psicologico, laddove Meta inserisce anche assistenza sanitaria e ByteDance la presenza di un’associazione professionale di categoria. Le sorti di queste persone, localizzate per lo più nei paesi del Sud Globale, sono precarie a causa di esternalizzazioni del servizio che le vincolano ad aziende intermediarie, impedendo la negoziazione contrattuale con l’amministrazione centrale. Il dialogo con i diretti interessati è reso impraticabile a causa della stipulazione di accordi di riservatezza (NDA). È possibile lavorare a distanza, ma la maggior parte del

personale è localizzato nelle Filippine e in India, con contratti a termine e stipendi più bassi degli altri colleghi *tech worker*, sebbene le mansioni comportino l'esposizione continuata a materiale disturbante che comporta un deterioramento della salute fisica e mentale. Un lavoro usurante che alcuni percepiscono come una missione per ostacolare la diffusione dei peccati e rimuovere l'immoralità dalla Rete. Sono sconosciute le persone che ci separano dagli algoritmi, ma rappresentano l'ennesimo caso di lavoro precario prodotto dalla *gig economy*. A questo punto dell'analisi si inserisce una prospettiva futura di indagine, in cui si rendano chiari i motivi per cui ai possessori di privilegio di status reputazionale o economico (gli utenti *premium* che ricevono o pagano per la spunta di verifica) spetti in via prioritaria la revisione umana.

I documenti *Modern Slavery Statement* rientrano nella responsabilità sociale d'impresa: le aziende autocertificano l'assenza di sfruttamento e illeciti nelle proprie attività produttive e di fornitura. In ByteDance vale come requisito contrattuale, includendo il divieto di lavoro minorile, forzato o involontario, la tratta degli esseri umani, discriminazioni o molestie, ripercussioni per chi aderisce ai sindacati, superamento dei limiti orari e mancato rispetto degli importi minimi salariali. È stata introdotta una piattaforma di *Speak Up* attraverso cui è possibile segnalare violazioni al Dipartimento *Ethics & Compliance*.

In un organigramma opaco di gestione, la ricerca per *query* ha restituito l'assenza di trasparenza sulle figure impegnate nei dipartimenti amministrativi delle Corporation digitali, che inibisce l'attribuzione di responsabilità da parte degli utenti. Tuttavia, emerge che sono i Dipartimenti di Sicurezza a monitorare gli impatti sociali delle modifiche alle regole, legate altresì al funzionamento, e che gli impiegati nella revisione assolvono mansioni di amministrazione ordinaria. Il termine Sicurezza, ancora una volta, è strettamente legato al concetto di Sovranità. X chiarisce che i suoi dipendenti sono esenti dai danni causati e arrecati dalla piattaforma, dove l'allineamento ai valori – e a chi li decide – è il cardine della formazione di ciascuno.

Il *Political Content Approach* di Meta esacerba una visione “apolitica” delle interazioni sociali, discostandosi dagli auspici dell'EMFA (2024) che inquadrano l'accesso a una varietà di notizie e contenuti di attualità come un antidoto contro la disinformazione e la concentrazione del mercato dei media come un pericolo per le opinioni divergenti, causando, inoltre, effetti di polarizzazione tra chi gode già di una *audience* più ampia. Le VLOP devono informare i servizi mediatici delle ragioni di una eventuale violazione dei termini di servizio prima di applicare qualsiasi sospensione o limitazione della visibilità. Il Regolamento 2024/900 chiarisce che le attività di comunicazione pubblica svolte da enti e istituzioni sono meritevoli di tutela. Così facendo, Meta discrimina il contenuto “politico” a favore di altre tipologie (come

l'intrattenimento, il commercio e il *lifestyle*), riducendo l'opportunità di incontro tra la domanda e l'offerta di opinioni e informazioni.

I proprietari di Meta e X hanno le 3W (*White Western Whales*) e concepiscono il mondo come uno spazio in cui le sfumature, incluse le minoranze e il dissenso, sono irrilevanti. L'uomo bianco, giovane e abile, occidentale¹⁰⁹ è il cittadino neutro che riceverà meno discriminazioni algoritmiche nella condivisione del pensiero e delle attività sulle piattaforme di Meta e X.

Questa è una forma di standardizzazione della cittadinanza che si discosta dai fondamenti presenti nelle Costituzioni delle liberal-democrazie, dove il principio di uguaglianza si accompagna alla tutela delle differenze e al sostegno di chi è privo di mezzi. Agli albori, Facebook e Twitter erano considerati canali alternativi ai media *mainstream* dalle minoranze etniche. Un ruolo che oggi viene ricoperto da TikTok proprio a causa della moderazione dei contenuti con strategia regionale (Elsawah, 2024).

È disfunzionale un sistema che alterna valori democratici e illiberali in base al contesto con pretesa di universalità e chi detiene potere economico non si limita a fruirne, ma lo investe per esercitare influenza nella sfera politica. Se i social media hanno contribuito alla materialità della libertà di espressione, come a loro tempo fecero la stampa, la radio e la televisione, la storia insegna che è la regolamentazione a impedire le concentrazioni di potere (economico tanto quanto politico), altrimenti predominanti nel settore.

La disposizione contenuta nella norma è il perno della giustizia di piattaforma (*Platform Justice*) che necessita di uno sguardo critico sui risvolti sociali che produce, recuperando i contributi di autori e autrici come Rawls, Sen, Nussbaum, Young e Dancy.

La predizione algoritmica, oltre a realizzare profili fondati su indici e categorie deterministiche, può restituire risultati inferenziali ingiusti e allocare privilegi, come dimostrano i casi di segregazione causati da COMPAS. La difesa della differenza non è basata sulle scelte intraprese ma sulle alternative rese disponibili, che sono legate all'opportunità di miglioramento delle proprie condizioni di vita. Non a caso, uno degli ambiti su cui si presta maggiore attenzione è l'Educazione perché gli orientamenti didattici incidono fortemente su ambizioni, aspettative e mobilità sociale. Le discriminazioni subite dagli studenti nel Regno Unito e dai lavoratori in Italia dimostrano che la trasparenza e la sindacabilità delle regole algoritmiche sono diventate il fondamento di diritti imprescindibili per i cittadini.

¹⁰⁹ La cerimonia di insediamento di Donald Trump nel gennaio 2025 segna un precedente nella storia degli *endorsement* USA. La presenza dei CEO delle Big Tech, inclusi Elon Musk e Mark Zuckerberg, ha spinto alcuni a definirli una *brologarchy* (termine coniato dalla crasi *brothers* + *oligarchy*), fondata sull'*ultra-wealthy tech-industrial complex*; un riferimento contenuto anche nel discorso di addio del presidente Joe Biden, diffuso pochi giorni prima: «*Today, an oligarchy is taking shape in America of extreme wealth, power and influence that really threatens our entire democracy, our basic rights and freedom*».

Nella società digitale del XXI secolo i lavoratori e le lavoratrici sono contemporaneamente la maggioranza che produce la ricchezza del capitalismo di piattaforma e il gruppo esautorato del plusvalore generato, finanche privo di identità collettiva e alienato a causa dei mutamenti avvenuti nell'ecologia industriale dei servizi, che ne ha depotenziato la capacità organizzativa (French, 2024). La contrattazione collettiva sui termini di servizio avviata nel 2023 dalla filiera cinematografica e l'emersione delle formazioni sindacali negli stabilimenti di Amazon e Google dimostra che è ancora possibile invertire questa tendenza.

Il *glitch* può diventare un momento di sovversione, l'imprevisto che sfugge alla definizione di eccezione al *datum*, l'espediente per impedire l'assorbimento dell'esperienza umana in costante trasformazione. Bonini e Tréré (2024) hanno dato risalto alle piccole esperienze di resistenza come pratiche politiche. Il raggirio e l'elusione della moderazione dei contenuti, la consapevolezza collettiva anche in assenza di notifica, ne sono una dimostrazione. Restituire agli utenti l'*agency* significa favorire l'opportunità della disobbedienza, ovvero della necessaria e creativa reazione alla fredda e granitica legge dell'algoritmo.

Appendice

A. Moderazione dei contenuti alla luce del *Digital Services Act*

Tabella 9. Comparazione tra le strategie di moderazione delle Corporation digitali (Meta-X-TikTok)

Corporation	Meta	X	TikTok
Leadership	Mark Zuckerberg / esplicita	Elon Musk / esplicita	ByteDance / implicita
Missione	Connettere persone, trovare comunità e far crescere le aziende → nicchia	Porsi al servizio della conversazione pubblica → opinione pubblica	Ispirare creatività e buonumore → <i>audience</i>
Strategia	Globale	Nazionale	Regionale
Tutela della “Libertà di Espressione”	Per il pubblico	Per chi si esprime	Per il <i>prosumer</i>
Moderazione dei contenuti	Rimuovere, ridurre e informare. Priorità a contenuti virali e con potenziale rischio offline (terrorismo, suicidio)	Lasciare, ridurre e rimuovere in base al contesto “ <i>Freedom of Speech, Not Reach</i> ” (FoSNR)	Identificare, revisionare e rimuovere in caso di violazione delle <i>policy</i>
Tecniche, Criteri e Linee Guida	Moderazione proattiva: • Rimozione account • Revisione contenuti (post, commenti, foto/video) Moderazione reattiva: • Avviso da parte di utenti o autorità	Moderazione contestuale: • Notifica (contenuto sensibile) • Restrizione in base all’età • Restrizione su base geografica	Moderazione proattiva: • Rimozione • Restrizione in base all’età • Controllo idoneità Feed “Per te” • Rafforzare la comunità con info, strumenti e risorse Moderazione reattiva: • Avviso da parte di utenti o autorità
Sospensione account (art. 23 DSA)	Limitazione misurata sull’entità del contenuto. Sospensione (periodo limitato) o disattivazione (definitiva, art. 4.2 ToS) se la violazione persiste <i>Misuse Policy</i> (art. 23 DSA)	Account di sola lettura (limitazione della possibilità di postare e interagire)	Sospensione o disattivazione permanente in caso di: • No requisiti minimi ed età • Appropriazione identità altrui • Contenuti illegali • Violazioni multiple e frequenti

Corporation	Meta	X	TikTok
Diminuzione visibilità	L'utente può chiedere una revisione ulteriore della decisione	Se il contenuto è considerato fastidioso ma legale, con un'etichetta	Riduzione della visibilità nel Paese in cui viola la legge o rimozione integrale
Obblighi di comunicazione trasparente (art. 24 DSA) al 06/24	Provvedimento: 27.7h FB; 20.4h IG Decisione: 28.1h Appello: 0.2h	Provvedimento: 2.7h Reclami su: contenuti illegali 2.8h; <i>policy</i> 0.34h Se richiesto dal Governo: rimozione: 4.1h; informazioni aggiuntive: 3 giorni	Segnalazione: 16 ore Provvedimento: 3h Se chiesto dal Governo: rimozione 6h; informazioni aggiuntive 3 giorni; riscontro 10h Appello: 1-2h
Sistemi automatizzati	Individuano con quanta probabilità il contenuto possa violare una delle <i>policy</i> , in base a segnali o <i>pattern</i> , e ne determinano la rimozione. Gli algoritmi sono allenati sulle decisioni umane in contesti simili	Strumenti: • Modelli di NLP • Modelli di <i>Image processing</i> • Modelli <i>content-related</i> (post insieme a testo, immagine, ecc...) • Rilevatore di violazioni di abuso • Modelli euristici (<i>pattern</i> di testi o parole chiave in alcune categorie)	Strumenti: • Modelli di <i>Computer Vision</i> • Sensore di Parole chiave • Sensore di duplicati
Moderazione umana (art. 42 DSA)	Classificazione della violazione in riferimento alla <i>policy</i>	Beta test sui sistemi automatizzati per la revisione. Revisione delle segnalazioni da parte dei modelli euristici o dagli utenti e delle richieste di appello. Classificazione della violazione in riferimento alla <i>policy</i> e assegnazione etichette	Approccio locale/regionale. Consulenza esterna (es. Tech Against Terrorism) e coinvolgimento esperti società civile (ONG e organizzazioni) e autorità governative

Corporation	Meta	X	TikTok
Risorse umane e condizioni di lavoro (art. 42 DSA) al 06/24	40.000 <i>Safe & Security Team</i> (15.000 revisori, di cui 2.500 “agnostici” con diverse mansioni 24/7 su oltre 70 lingue)	1849 revisori	40.000 <i>Trust & Safety Team</i> (6.354 revisori in UE)
Contratto	A tempo pieno, a contratto ed esternalizzati	Dipendenti interni e consulenti esterni (la metà con max 2 anni di anzianità)	Sede a Dublino e specificità regionali UE
Formazione & Welfare	Percorso di formazione Supporto psicologico e sanitario	Percorso di formazione continua Supporto psicologico	Percorso di formazione Supporto psicologico Associazione professionale

Fonte: Meta – X – TikTok Transparency Report 2023 e 2024

B. Fonti

Meta

Transparency Report

- a. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook (27 October 2023)
- b. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram (27 October 2023)
- c. Digital Services Act - Information on Average Monthly Active Recipients in the European Union (February 2023)
- d. Digital Services Act - Information on Average Monthly Active Recipients in the European Union (August 2023)
- e. Digital Services Act - Information on Average Monthly Active Recipients in the European Union (February 2024)
- f. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook (26 April 2024)
- g. Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Instagram (26 April 2024)

Community Guidelines

- a. Community Guidelines – Instagram Help Center
- b. Facebook Community Standards
 - Introduction
 - I. Content-related Requests and Decisions
 - o Additional protection of minors
 - o User requests
 - II. Integrity and Authenticity
 - o Account integrity and authentic identity
 - o Cybersecurity
 - o Inauthentic behaviour
 - o Memorialisation
 - o Misinformation
 - o Spam
 - III. Objectionable Content
 - o Adult nudity and sexual activity
 - o Adult sexual solicitation and sexually explicit language
 - o Hate speech
 - o Violent and graphic content
 - IV. Respecting Intellectual Property
 - V. Safety
 - o Adult sexual exploitation
 - Detecting Non-Consensual Intimate Images and Supporting Victims
 - A Research-Based Approach to Protecting Intimate Images
 - o Bullying and harassment
 - o Child sexual exploitation, abuse and nudity
 - New Technology to Fight Child Exploitation
 - o Human exploitation
 - o Privacy violations
 - o Suicide, Self-Injury and Eating Disorders
 - VI. Violence and Criminal Behaviour
 - o Coordinating harm and promoting crime
 - o Dangerous organisations and individuals
 - o Fraud and deception
 - o Restricted goods and services
 - o Violence and incitement
- c. Not Without My Consent

Cookies Policy

- a. Meta Cookies Policy

Misuse Policy

- a. Misuse Policy
- b. Dissemination of Terrorist Content Online – Instagram Help Center
- c. Meta Pay Acceptable Use Policy

Platform Policy

- a. AI at Meta
- b. Enforcement – Meta Development
- c. European Electronic Communications Code - Facebook
- d. European Electronic Communications Code - Instagram
- e. Introduction to the Advertising Standards
- f. IP Reporting API – Terms
- g. Legacy Instagram Platform Policy
- h. Our approach to explaining ranking
- i. Platform Terms – Meta for Developers
- j. Recommendations on Instagram
- k. What are recommendations on Facebook

Privacy policy

- a. Meta Privacy Policy – How Meta Collects and uses user data
- b. Meta Privacy Policy (June 2024)
- c. Threads Policies and Terms

Terms of Service

- a. Terms of Service – Facebook
- b. Terms of Use – Instagram
- c. Threads Terms of Use
- d. About professional accounts of Instagram
- e. Meta Pay Merchant Terms and Conditions
- f. Supplemental Portal Terms of Service – Meta Store
- g. Onsite Checkout Payments Features Terms and Conditions

X

Transparency Report

- a. DSA Transparency Report November 2023
- b. DSA Transparency Report April 2024

Community Guidelines

- a. General
 - X Rules
 - X's policy on username squatting
- b. Intellectual Property
 - Automated copyright claims for live video
 - X's copyright policy
 - X's counterfeit policy
 - X's trademarks policy
- c. Law Enforcement Guidelines
 - FAQ on legal requests to X
 - X's guidelines for law enforcement
- d. Platform Integrity and authenticity
 - Copypasta and duplicate content policy
 - X's civic integrity policy
- e. Country-specific resources
 - Australia Online Safety Information
 - Digital Services Act
 - European Union
 - India
 - India 2
- f. Platform Use Guidelines
 - Guidelines and Principles for X moments
 - How to report violations of X Rules and Terms of Service
 - Overview on X's rules and account behaviors
 - Public-interest exceptions to enforcement of X rules
 - Reporting false information on X for France
 - Understanding when content is withheld based on country
 - X's enforcement philosophy & approach to policy development
 - X's free speech and rights of people | X Help

Cookies Policy

- a. Platform Use Guidelines
 - How cookies are used on X
 - More information on X's data processing

Misuse Policy

- a. Account Settings
 - About blocking on X and how to know if someone blocked you
 - Contacting X about a deceased family member's account
 - Customizing your X experience
 - Debunking X myths
 - How to block accounts on X
 - How to handle X abuse and how we can help
 - How to help someone experiencing online abuse
 - How to make images accessible for people
 - How to manage your block list
 - How to mute accounts on X and what happens after muting
 - How to protect and unprotect your Posts
 - How to report abusive behavior on X

- How to use advanced muting options
- Not a Bot
- Our approach to blocking links
- Report impersonation accounts
- Reporting and viewing sensitive media
- Understanding age screening and how it works
- Understanding how X handles offensive or explicit Posts
- Understanding your media settings on X
- Verified Organizations
- What to do about self-harm and suicide concerns on X
- b. General
 - X's policy on media depicting deceased individuals or death
- c. Platform integrity and authenticity
 - Misleading and deceptive identities policy
 - Our synthetic and manipulated media policy
 - X's financial scam policy
 - X's platform manipulation and spam policy
 - X's policy on ban evasion
- d. Platform Use Guidelines
 - Misuse of Reporting Features Policy
 - Our range of enforcement options for violations
 - Specific instances when a post's reach may be limited
 - Understanding search rules and restrictions
 - Understanding X limits
 - X's fair use policy
- e. Safety and Cybercrime
 - Adult Content Policy
 - How X handles abuse and harassment
 - Our policy on child sexual exploitation
 - Our policy on illegal or certain regulated goods and services
 - Our policy on violent organizations
 - Perpetrators of violent attacks policy
 - Violent Content Policy
 - X's abusive profile and profanity policy
 - X's non-consensual nudity policy
 - X's policy on hateful conduct
 - X's private information policy and doxing
 - X's suicide & self-harm policy

Platform Policy

- a. Platform Use Guidelines
 - About our approach to recommendations
 - About X's APIs
 - Account Recommendations
 - Conversations Recommendations
 - Explore Recommendations
 - For You Home Timeline Recommendations
 - Notifications Recommendations
 - Our guidelines for running promotions
 - Our style guide for curation of Moments
 - Paid Partnerships Policy
 - Professional Account Policy
 - Reporting X security vulnerabilities and bugs
 - Search Recommendations
 - Spaces Recommendations
 - Subscriptions policy
 - Trends Recommendations
 - X account notices and what they mean
 - X shopping policies
 - X's automation development rules

- X's Creator Monetization Standards
- X's inactive account policy
- b. Research and Experiments
 - User research conducted by X
 - Welcome to the X experiments program
 - What is the X iOS Beta program
 - X Voices, X Insiders, and Creator Insiders Terms of Use

Privacy policy

- a. X Privacy Policy
- b. Platform Use Guidelines
 - X's Services, corporate affiliates, and your privacy

Terms of Service

- a. X Terms of Service
- b. Platform Use Guidelines
 - Updates to our Terms of Service and Privacy Policy
 - Vine camera terms of service and privacy policy

TikTok

Transparency Report

- a. TikTok's DSA Transparency Report 2023
- b. TikTok's DSA Transparency Report January – June 2024

Community Guidelines

- a. Branded Content Policy
- b. Branded Content Policy – Country-specific Requirements
- c. Intellectual Property Policy
- d. Intellectual Property Infringements – TikTok for Business
- e. Commercial Music Library – User Terms
- f. Community Guidelines
- g. Community Guidelines Enforcement Report
- h. Countering influence operations
- i. Engaging our Advisory Councils
- j. Protecting teens online
- k. Protecting the integrity of elections
- l. Supporting independent research
- m. TikTok Law Enforcement Guidelines
- n. TikTok UK Modern Slavery Statement 2023
- o. Upholding human rights – Our commitments

Cookies Policy

- a. Data Collection Standards
- b. Requesting your data
- c. TikTok Platform Cookies Policy
- d. TikTok Web Cookies Policy

Misuse Policy

- a. Combating harmful misinformation
- b. Combating hate and violent extremism
- c. Content violations and bans
- d. Why Some Ads Are Rejected After They Are Approved

Platform Policy

- a. Advertising Policies
 - Ad Format and Functionality
 - Ad Review
 - o Checklist Overview
 - o Ad Language
 - o Ad Messaging
 - o HTTPS Link
 - o Landing Page
 - Adult Content
 - Advertising & Ad Review
 - Age Targeting Restrictions
 - Alcohol
 - Animals and Environment
 - Anti-Discrimination Policy
 - Common Reasons Ads Fail Review
 - Dangerous Products or Services
 - Deceptive Practices
 - Business Help Center
 - Gambling and Games

- Healthcare and Pharmaceuticals
- Housing, Employment, and Credit (HEC)
- Intellectual Property Infringement
- Misleading and false content
- Other products and services
- Politics, Religion and Culture
- TikTok Ad Serving Policy
- Violence and Dangerous Activities
- Weight Management
- Youth Safety
- Actor Policy
- b. Coins Policy
- c. Introduction to the TikTok recommendation system
- d. Labelling State-affiliated Media Entities
- e. Open Source Software Notices
- f. Our approach to content moderation
- g. Reward Policy
- h. Supporting responsible, transparent AI-generated content
- i. System Guides
- j. TikTok Ad After Conversion Experience (ACE) policies
- k. Virtual Items

Privacy policy

- a. Privacy policy
- b. Children's Privacy Policy
- c. Strengthening Security and Privacy

Terms of Service

- a. Terms of Service
- b. Business Terms of Service
- c. Music Terms of Service

Bibliografia

- Abba, L., Lazzaroni, A., & Pietrangelo, M. (2022). *La Internet Governance e le sfide della Trasformazione digitale*. Napoli: Internet Society Italia, Editoriale Scientifica.
- Accountable Tech. (luglio 2024). *Meta's Political Content Limit Causes Steep Drop in Reach for Accounts*.
- Amato, S. (2021). *Biodiritto 4.0. Intelligenza artificiale e nuove tecnologie*. Torino: G. Giappichelli Editore.
- Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.
- Angwin, J., & Grassegger, H. (2017, giugno 28). *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children*. Tratto da ProPublica: <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, maggio 23). *Machine Bias*. Tratto da ProPublica. There's software used across the country to predict future criminals. And it's biased against blacks: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Are, C. (2022). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22, 2002-2019.
- Babalola, O. (2023, febbraio). Data Protection Legal Regime and Data Governance in Africa: An Overview. *AERC Working Paper*, DG-003, 1-32.
- Barassi, V. (2021). *I figli dell'algoritmo: Sorvegliati, tracciati, profilati dalla nascita*. Luiss University Press.
- Barrett, B., & Kreiss, D. (2019). Platform transience: changes in Facebook's policies, procedures, and affordances in global electoral politics. *Internet Policy Review*, 8(4).
- Bazzicalupo, L. (2010). *Biopolitica. Una mappa concettuale*. Roma: Carocci Editore.
- Bazzicalupo, L. (2014). *Politica. Rappresentazioni e tenciche di governo*. Roma: Carocci Editore.
- Bazzicalupo, L. (2015). Biopolitica come governamentalità: la cattura neoliberale della vita. *La Deleuziana - Rivista online di Filosofia*, 1, 27-39.
- Beck, U. (2000). *La società del rischio. Verso una nuova modernità*. (W. Privitera, A cura di) Roma: Carocci Editore.
- Berg, S., & Hofmann, J. (2021). Digital democracy. *Internet Policy Review*, 10(4).
- Biddle, S. (2021, ottobre 12). *Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations"*. Tratto da The Intercept: <https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous/>
- Bloch-Wehba, H. (2020). Automation in moderation. *Cornell Int'l LJ*, 53(41).
- Blum, A. (2012). *Tubes: A Journey to the Center of the Internet*. New York: HarperCollins Publishers Inc.
- Blunt, D., Coombes, E., Mullin, S., & Wolf, A. (2020). Posting Into The Void. A community report by Hacking//Hustling. 1-87.
- Bobbio, N., Matteucci, N., & Pasquino, G. (2004). *Il dizionario di politica*. Utet.
- Bonini, T., & Treré, E. (2024). *Algorithms of resistance: The everyday fight against platform power*. MIT Press.
- Bouko, C., & Rothut, S. (2024, maggio 29). *Définir et modérer le contenu terroriste en ligne : trois défis majeurs, Catherine Bouko et Sophia Rothut*. Tratto da blogdroiteuropeen: <https://blogdroiteuropeen.com/2024/05/29/definir-et-moderer-le-contenu-terroriste-en-ligne-trois-defis-majeurs-catherine-bouko-et-sophia-rothut/>
- Bradford, A. (2023). *Digital empires: The global battle to regulate technology*. Oxford University Press.

- Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.
- Brown, I. (2020). *Interoperability as a tool for competition regulation*. OpenForum Academy.
- Browne, J., Cave, S., Drage, E., & McInerney, K. (2023). *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford: Oxford University Press.
- Bucher, T. (2018). *If... then: Algorithmic power and politics*. Oxford University Press.
- Calise, M., & Lowi, T. J. (2010). *Hyperpolitics. An Interactive Dictionary of Political Science Concepts*. University of Chicago Press.
- Calise, M., & Musella, F. (2019). *Il principe digitale*. Gius. Laterza & Figli Spa.
- Calise, M., Lowi, T. J., & Musella, F. (2021). *Concetti chiave: manuale di scienza politica*. Società editrice Il Mulino.
- Cammaerts, B., & Mansell, R. (2020). Digital Platform Policy and Regulation: Toward a Radical Democratic Turn. *International Journal of Communication*, 14, 135-154.
- Caplan, R. (2018). *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. Data & Society.
- Caplan, R. (2023). Networked Platform Governance: The Construction of the Democratic Platform. *International Journal of Communication*, 17, 3451–3472.
- Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, 1–13.
- Capone, E. (2022, ottobre 11). *La Russia ha inserito Meta fra le organizzazioni "terroristiche ed estremiste": perché è successo e cosa significa*. Tratto da La Repubblica: https://www.repubblica.it/tecnologia/2022/10/11/news/la_russia_ha_inserito_meta_fra_le_organizzazioni_terroristiche_ed_estremiste_perche_e_successo_e_cosa_significa-369551522/
- Caramani, D. (2023). *Comparative politics*. Oxford: Oxford University Press.
- Celeste, E., Heldt, A., & Iglesias Keller, C. (2022). *Constitutionalising Social Media* (Vol. Hart Studies in Information Law and Regulation). Oxford: Hart Publishing.
- Center for AI and Digital Policy . (2023). *Artificial Intelligence and Democratic Values*. Tratto da CAIDP: <https://www.caidp.org/reports/aidv-2023/>
- Chen, A. (2012, febbraio 16). *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads'*. Tratto da Gawker: <https://www.gawkerarchives.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads>
- Chen, A. (2014, ottobre 23). *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*. Tratto da Wired: <https://www.wired.com/2014/10/content-moderation/>
- Commissione europea. (2022, ottobre 17). *European Digital Rights and Principles*. Tratto da European commission: <https://digital-strategy.ec.europa.eu/en/policies/digital-principles>
- Commissione europea. (2022, settembre 16). *The Digital Economy and Society Index (DESI)*. Tratto da European commission: <https://digital-strategy.ec.europa.eu/en/policies/desi>
- Commissione europea. (2022, ottobre 12). *The Digital Services Act package*. Tratto da European Commission: <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>
- Commissione europea. (2023, aprile 25). *DSA: Piattaforme online molto grandi e motori di ricerca* . Tratto da Commissione europea: <https://digital-strategy.ec.europa.eu/it/policies/dsa-vlops>
- Cotter, K. (2023). “Shadowbanning is not a thing”: black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, 6(26), 1226-1243.

- Cowls, J., Darius, P., Santistevan, D., & Schramm, M. (2022). Constitutional Metaphors: Facebook's «Supreme Court» and the Legitimation of Platform Governance. *New Media & Society*.
- Crawford, K. (2016, giugno 25). Artificial intelligence's white guy problem. *New York Times*.
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Crawford, K., & Gillespie, T. (2014). What is a Flag for? Social media Reporting tools and the Vocabulary of Complaint. *New Media & Society*, doi: 10.1177/1461444814543163.
- Creemers, R., Webster, G., & Toner, H. (2022, gennaio 10). *Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022*. Tratto da DigiChina: <https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/>
- Culpepper, P. D., & Thelen, K. (2020). Are We All Amazon Primed? Consumers and the Politics of Platform Power. *Comparative Political Studies*, 53(2), 288–318.
- Dancy, T. (2023). *Artificial Justice*. Oxford Technology Law and Policy, Oxford University Press.
- Davis, A. Y. (2018 [1981]). *Donne, razza e classe*. Roma: Alegre.
- de Baillencourt, J. (2022, settembre 12). *Collaboriamo per prevenire l'estremismo violento*. Tratto da TikTok Newsroom: <https://newsroom.tiktok.com/it-it/tiktok-prevenire-extremismo-violento>
- De Tena, C. L., & De Viedma, D. G. (2023). *European Tech Insights 2023. Mapping European Attitudes towards Technological Change and its Governance*. Center for the Governance of Change at IE University.
- Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring Data Justice: Conceptions, Applications and Directions. *Information, Communication & Society*, 22(7), 873-881.
- Diaz, Á., & Hecht-Felella, L. (2021). *Double standards in Social Media Content Moderation*. Brennan Center for Justice at New York University School of Law.
- D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press.
- Dini, A. (2023, ottobre 14). *La “guerra fredda” dell'AI*. Tratto da Guerre di Rete: <https://www.guerredirete.it/la-guerra-fredda-dell-ai/>
- Divon, T., Are, C., & Briggs, P. (2025). Platform gaslighting: A user-centric insight into social media corporate communications of content moderation. *Platforms & Society*, 2, 1-17.
- Doda, I. (2024). *L'utopia dei miliardari*. Roma: Edizioni Tlon.
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 10.1126/sciadv.aao5580.
- Eccles, M., & Gallardo, C. (2020, agosto 14). *UK government faces elitism row after algorithm downgrades pupils*. Tratto da Politico: <https://www.politico.eu/article/uk-government-faces-elitism-row-after-algorithm-downgrades-final-year-pupils/>
- Eicher, B., Polepeddi, L., & Goel, A. (2018). Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, (p. 88-94).
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models.
- Elsawah, M. (2024, settembre). *Moderating Maghrebi Arabic Content on Social Media*. Tratto da Center for Democracy & Technology (CDT): <https://cdt.org/wp-content/uploads/2024/09/2024-09-26-CDT-Research-Global-South-Moderating-Report-English-Arabic-final.pdf>

- Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Ethics, Governance, and Policies in Artificial Intelligence*, 81-90.
- Floridi, L. (2022). *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*. Raffaello Cortina Editore.
- Fossa, F., Tamburrini, G., & Schiaffonati, V. (2021). *Automi e persone. Introduzione all'etica dell'intelligenza artificiale e della robotica*. Roma: Carocci.
- Foucault, M. (2004 [1972]). *L'Ordine del discorso e altri interventi*. Torino: Einaudi.
- Franchi, J. (2021). *Gli obsoleti. Il lavoro impossibile dei moderatori di contenuti*. Agenzia X.
- Freeman, P., & Aspray, W. (1999). The Supply of Information Technology Workers in the United States. *Computing Research Association (CRA)*.
- French, N. (2024, settembre 14). *No, Liberalism Hasn't Buried Marxism*. Tratto da Jacobin: <https://jacobin.com/2024/09/liberalism-marxism-cohen-rawls-workers/>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.
- Gillespie, T. (2022, giugno - settembre). Do Not Recommend? Reduction as a form of Content Moderation. *Social Media + Society*, p. 1-13.
- Giugni, L. (2022). *La rete non ci salverà: perché la rivoluzione digitale è sessista (e come resistere)*. Longanesi.
- Greenleaf, G., & Cottier, B. (2018). Data privacy laws and bills: Growth in Africa, GDPR influence. *Privacy Laws & Business International Report*, 152(11), 11-13.
- Gros, F. (2019). *Disobbedire*. Torino: Einaudi.
- Haggart, B., & Tusikov, N. (2023). *The New Knowledge. Information, Data and the Remaking of Global Power*. London: Rowman & Littlefield Publishers.
- Hatmaker, T. (2022, novembre 4). *Elon Musk just axed key Twitter teams like human rights, accessibility, AI ethics and curation*. Tratto da Tech Crunch: <https://techcrunch.com/2022/11/04/elon-musk-twitter-layoffs/>
- Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information polity*, 23(4), 429-434.
- Hildebrandt, M. (2018). Algorithmic Regulation and The Rule of Law. *Philosophical transactions of the Royal Society A: Mathematics, Physical and Engineering Sciences*, 376(2128), 1-11.
- Hobsbawm, E. J. (1972). *La rivoluzione industriale e l'impero. Dal 1750 ai giorni nostri*. Torino: Giulio Einaudi editore.
- Human Rights Watch. (2018, Febbraio 14). *Germany: Flawed Social Media Law*. Tratto da HRW: <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law>
- Internazionale. (2019, aprile 12). *Hans Block racconta una scena di Quello che i social non dicono - The cleaners*. Tratto da Internazionale: <https://www.internazionale.it/video/2019/04/12/hans-block-quello-social-non-dicono-the-cleaners>
- Istituzione dell'elenco di Organismi ADR, ai sensi dell'art. 141-decies del Codice del Consumo, e approvazione del Regolamento sulla tenuta degli Organismi ADR nel settore delle comunicazioni elettroniche e postali, Delibera n. 661/15/CONS (AGCOM dicembre 1, 2015).
- Jones, N. (2024, maggio 29). *Who owns your voice? Scarlett Johansson OpenAI complaint raises questions*. Tratto da Nature: <https://www.nature.com/articles/d41586-024-01578-4>
- Kaplan, J. (2016). *Le persone non servono: lavoro e ricchezza nell'epoca dell'intelligenza artificiale*. Luiss University Press.
- Kassem, S. (2023). *Work and Alienation in the Platform Economy. Amazon and the Power of Organization*. Bristol University Press.

- Keane, D. (2020, luglio 10). *Lobbying success in EU digital policy*. Tratto da <https://thesis.eur.nl/pub/56292>
- Kennedy, H. (2016). *Post, mine, repeat. Social media data mining becomes ordinary*. London: Palgrave Macmillan (Springer Nature).
- Kleppmann, M., Frazee, P., Gold, J., Graber, J., Holmgren, D., Ivy, D., . . . Volpert, J. (2024). Bluesky and the AT Protocol: Usable Decentralized Social Media. *DIN '24 Proceedings (MOBICOM)*, 1-7.
- Klonick, K. (2020). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *The Yale Law Journal*, 129(2418).
- Klonick, K. (2021, febbraio 12). *Inside the making of the Facebook's Supreme Court*. Tratto da NewYorker: www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebookssupreme-court
- Knight, W. (2022, novembre 4). *Elon Musk Has Fired Twitter's 'Ethical AI' Team*. Tratto da Wired: <https://www.wired.com/story/twitter-ethical-ai-team/>
- Lahoti, S. (2019, luglio 3). *Facebook fined \$2.3 million by Germany for providing incomplete information about hate speech content*. Tratto da Packt Hub: <https://hub.packtpub.com/facebook-fined-2-3-million-by-germany-for-providing-incomplete-information-about-hate-speech-content/>
- Latiff, R. (2024, ottobre 11). *ByteDance's TikTok cuts hundreds of jobs in shift towards AI content moderation*. Tratto da Reuters: <https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai-moderation-sources-say-2024-10-11/>
- Lessig, L. (1999). *Code and other laws of Cyberspace* (2006 ed.). New York: Basic Books.
- Lijphart, A. (1971, September). Comparative Politics and the Comparative Method. *The American Political Science Review*, 65(3), 682-693.
- Lockwood, B. (2024, settembre 9). *'I quit my job as a content moderator. I can never go back to who I was before.'* Tratto da The Washington Post: <https://www.washingtonpost.com/opinions/2024/09/09/shifts-content-moderator-comic/>
- Longo, E. (2023). *Giustizia digitale e Costituzione: riflessioni sulla trasformazione tecnica della funzione giurisdizionale*. Torrossa.
- Longo, G. O. (2017). Prospettive del post-umano. *Futuribili – Rivista di studi sul futuro e di previsione sociale*, XXII(2), 175-190.
- Maddaus, G. (2023, maggio 23). *How the WGA Decided to Harness — but Not Ban — Artificial Intelligence*. Tratto da Variety: <https://variety.com/2023/biz/news/wga-ai-writers-strike-technology-ban-1235610076/>
- Manancourt, V. (2022, luglio 7). *Europe faces Facebook blackout*. Tratto da Politico: <https://www.politico.eu/article/europe-faces-facebook-blackout-instagram-meta-data-protection/>
- Mazzucato, M. (2014). *Lo Stato Innovatore*. Bari: Gius. Laterza & Figli Spa.
- Mazzucato, M. (2018). *Il valore di tutto. Chi lo produce e chi lo sottrae nell'economia globale*. Bari: Editori Laterza.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955, agosto 31). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Tratto da <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- Meaker, M. (2022, settembre 23). *A Danish City Built Google Into Its Schools - Then Banned It*. Tratto da Wired: <https://www.wired.co.uk/article/denmark-google-schools-data>
- Menabrea, L. F. (1842, ottobre). *Sketch of The Analytical Engine Invented by Charles Babbage. With notes upon the Memoir by the Translator Contless Ada A. Lovelace (1843), Scientific Memoirs*, 3, pp. 666-731. Tratto il giorno gennaio 2024 da <https://www.fourmilab.ch/babbage/sketch.html>

- Meta. (2022, marzo). *The Code of Conduct*. Tratto da About Facebook: <https://about.facebook.com/code-of-conduct>
- Meta. (marzo, 2022). *Facebook Community Standards*. Tratto da Transparency Center: <https://transparency.fb.com/it-it/policies/community-standards>
- Mill, J. S. (1859). *On Liberty*. Chicago and New York: Belford, Clarke & Company.
- Milmo, D. (2024, maggio 27). *Scarlett Johansson's OpenAI clash is just the start of legal wrangles over artificial intelligence*. Tratto da The Guardian: <https://www.theguardian.com/technology/article/2024/may/27/scarlett-johansson-openai-legal-artificial-intelligence-chatgpt>
- Musella, F. (2022). Digital regulation: come si cambia la Pubblica Amministrazione. *Rivista Di Digital Politics*, 3-31.
- Nieminen, H., Padovani, C., & Sousa, H. (2023). Why Has the EU Been Late in Regulating Social Media Platforms? *Javnost - The Public*, 30(2), 174-196.
- Nikiforov, L. (2024). The Digital Services Act, content moderation and elections. *Brussels Privacy Hub*, 1-9.
- Numerico, T. (2021). *Big Data e Algoritmi: Prospettive critiche*. Carocci.
- Nussbaum, M. C. (2002). *Giustizia sociale e dignità umana. Da individui a persone*. ilMulino.
- Ofqual. (2020, settembre 2). *Written statement from Chair of Ofqual to the Education Select Committee*. Tratto da UK Government: <https://www.gov.uk/government/news/written-statement-from-chair-of-ofqual-to-the-education-select-committee>
- O'Neil, C. (2017). *Armi di distruzione matematica*. Bompiani.
- OSTP. (2022, ottobre). *Blueprint for an AI Bill of Rights*. Tratto da The White House: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
- Oversight Board. (2023, giugno). *Annual Report: Oversight Board reviews Meta's changes to bring fairness and transparency to its platforms*. Tratto da Oversight Board: <https://oversightboard.com/news/560960906211177-2022-annual-report-oversight-board-reviews-meta-s-changes-to-bring-fairness-and-transparency-to-its-platforms/>
- Oxford University Press. (2023, ottobre 18). *AI in education: where we are and what happens next*. Tratto da OUP: <https://corp.oup.com/feature/ai-in-education-where-we-are-and-what-happens-next/>
- Palazzani, L. (2020). *Tecnologie dell'informazione e intelligenza artificiale: Sfide etiche al diritto*. Edizioni Studium Srl.
- Parlamento europeo e Consiglio. (2024, aprile 11). Regolamento (UE) 2024/1083 che istituisce un quadro comune per i servizi di media nell'ambito del mercato interno e che modifica la direttiva 2010/13/UE. *European Media Freedom Act (ELI: http://data.europa.eu/eli/reg/2024/1083/oj)*. Testo rilevante ai fini del SEE.
- Parlamento europeo e Consiglio. (2024, marzo 13). Regolamento (UE) 2024/900 relativo alla trasparenza e al targeting della pubblicità politica. (ELI: <http://data.europa.eu/eli/reg/2024/900/oj>).
- Pasquinelli, M. (2023). *Pasquinelli, M. (2023). The eye of the master: A social history of artificial intelligence*. Verso Books.
- Polanyi, K. (1974). *La grande trasformazione. Le origini economiche e politiche della nostra epoca*. Torino: Piccola Biblioteca Einaudi Ns.
- Project Lovelace. (1842). *Ada Lovelace's Note G*. Tratto il giorno gennaio 2024 da Project Lovelace: <https://projectlovelace.net/problems/ada-lovelaces-note-g/>
- Rawls, J. (1971). *A Theory of Justice*. Cambridge, Massachusetts: Belknap Press of Harvard University Press.
- Rivista Africa. (2021, ottobre 4). *L'Africa nel mirino della criminalità informatica*. Tratto da Rivista Africa: <https://www.africarivista.it/africa-il-continente-nel-mirino-della-criminalita-informatica/192488/#:~:text=%E2%80%9D%2C%20ha%20sottolineato.->

- ,La%20convenzione%20di%20Malabo%20sulla%20sicurezza%20informatica%20e%20la%20protezione,ratificato%20l'accordo%20d
- Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, *iaae073*, <https://doi.org/10.1093/ia/iaae073>.
- Roberts, S. T. (2017). Content Moderation. In L. Schintler, & C. McNeely, *Encyclopedia of Big Data* (p. https://doi.org/10.1007/978-3-319-32001-4_44-1). Cham: Springer.
- Roberts, S. T. (2019). *Behind The Screen*. Yale University Press.
- Sadin, É. (2019). *Critica della ragione artificiale. Una difesa dell'umanità*. Luiss University Press.
- SAG-AFTRA. (2024, settembre 17). *Gov. Newsom Signs Union-Championed A.I. Bills at SAG-AFTRA Plaza*. Tratto da SAG-AFTRA: <https://www.sagaftra.org/gov-newsom-signs-union-championed-ai-bills-sag-aftra-plaza>
- Santaniello, M. (2021). La regolazione delle piattaforme e il principio della sovranità digitale. *Rivista di Digital Politis*, *1*(3), 579-600.
- Santoriello, S. C. (2022). Al di là dei confini. Ripensare il paradigma della frontiera in una prospettiva interculturale. (M. Durante, A cura di) *Quaderni OCSM*, *1*, 297-319.
- Santoriello, S. C. (2022). Il Meta-diritto dell'Oversight board. *Rivista di Digital Politics*, *I-II*, 285-302.
- Santoriello, S. C. (2024, febbraio 2). *Book Review: "L'utopia dei miliardari" di Irene Doda (2024)*. Tratto da Game of Tech: <https://www.gameoftech.eu/2024/02/07/book-review-lutopia-dei-miliardari-di-irene-doda-2024/>
- Santoriello, S. C., & Franchi, J. (2023). Se la giustizia diventa un affare privato. *Emma. Culture e pensieri libertari*, 70-74.
- Sartori, G. (1995 [1987]). *Elementi di teoria politica*. Bologna: Il Mulino.
- Schaake, M. (2024). *The Tech Coup: How to Save Democracy from Silicon Valley*. Princeton University Press.
- Scharlach, R. (2024). How to Spark Joy: Strategies of Depoliticization in Platform's Corporate Social Initiatives . *Social Media + Society*, doi.org/10.1177/20563051241277601.
- Schyns, C. (2023). *The lobbying ghost in the machine*. Brussels: Corporate Europe Observatory.
- Sen, A. (2009). *The Idea of Justice*. London: Penguin.
- Simon, F. M. (2022). Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism*.
- Spangler, T. (2024, febbraio 9). *Instagram and Threads Will Quit Recommending Political Content*. Tratto da Variety: <https://variety.com/2024/digital/news/instagram-political-content-recommendations-1235906219/>
- Staglianò, R. (2022). *Gigacapitalisti*. Einaudi.
- Stewart, G. (2024, novembre 1). *Meta's content preferences raise concerns amid upcoming 2024 election with recommended political content default settings*. Tratto da The Columbia Chronicle: <https://columbiachronicle.com/metasp-content-preferences-raise-concerns-amid-upcoming-2024-election-with-recommended-political-content-default-settings>
- Tarrant, A., & Cowen, T. (2022). Big Tech Lobbying in the EU. *The Political Quarterly*, *93*(2), 218-226.
- Taylor, A. (2014). *The people's platform: Taking back power and culture in the digital age*. Metropolitan Books.
- The Associated Press. (2022, dicembre 12). *Musk's Twitter has dissolved its Trust and Safety Council*. Tratto da NPR: <https://www.npr.org/2022/12/12/1142399312/twitter-trust-and-safety-council-elon-musk>

- Tik Tok. (2023, giugno). *Engaging our Advisory Councils*. Tratto da Our commitments: <https://www.tiktok.com/transparency/en-us/advisory-councils>
- Tik Tok. (2023, giugno). *Our Approach to Content Moderation*. Tratto da Our commitments: <https://www.tiktok.com/transparency/en-us/content-moderation/>
- Treves, T. (2008). *United Nations Convention on the Law of the sea*. Tratto il giorno gennaio 2024 da United Nations Audiovisual Library of International Law: http://untreaty.un.org/cod/avl/pdf/ha/uncls/uncls_e.pdf
- Tufekci, Z. (2018, marzo 10). *YouTube, the Great Radicalizer*. Tratto da The New York Times: <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236).
- UNESCO. (2021). *AI and Education. Guidance for policy-makers*. Tratto da <https://unesdoc.unesco.org/ark:/48223/pf0000376709/PDF/376709eng.pdf.multi>
- Van Dijck, J. (2019). Reflections on creating responsible digital societies in Europe. Knowledge for Change Lecture series #4. (T. b. Benecchi, A cura di) *Studies in Communication Sciences*, 19(2), 161–171.
- Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.
- Veutro, M. F. (2008). Qualità e quantità: una sintesi del dibattito. *L'analisi qualitativa. Teorie, metodi, applicazioni*, 29.
- Wiener, N. (1950). *The Human Use of Human Beings*. The Riverside Press (Houghton Mifflin Co.).
- Writers Guild of America. (2023, ottobre). *Summary of the 2023 WGA MBA*. Tratto da WGA Contract 2023: <https://www.wgacontract2023.org/the-campaign/summary-of-the-2023-wga-mba>
- Wu, T. (2003). When code isn't law. *Virginia Law Review*, 89 (4), 679-752.
- Wu, T. (2011). *The master switch: The rise and fall of information empires*. Vintage.
- Wu, T. (2018). *The curse of bigness: Antitrust in the New Gilded Age*. Faculty Books, 63.
- Young, I. M. (2024 [1996]). *Le politiche della differenza*. Edizioni Società Aperta.
- Zhang, B., & Barata, J. (2020, marzo 1). *Provisions on the Governance of the Online Information Content Ecosystem*. Tratto da World Intermediary Liability Map: <https://wilmap.stanford.edu/entries/provisions-governance-online-information-content-ecosystem>
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Filmografia

Blade Runner (1982) di Ridley Scott
Videodrome (1983) di David Cronenberg
Brainstorm (1983) di Douglas Trumbull
Decoder (1984) di Muscha
Hardware (1990) di Richard Stanley
The Net (1995) di Irwin Winkler
Conceiving Ada (1997) di Lynn Hershman Leeson
The Matrix (1999) di Lana e Lilly Wachowski
Bread and Roses (2000) di Ken Loach
A.I. Artificial Intelligence (2001) di Stephen Spielberg
Slm0ne (2002) di Andrew Niccol
Surrogates (2009) di Jonathan Mostow
The Social Network (2010) di David Fincher
The Congress (2013) di Ari Folman, da “Il congresso di futurologia” di Stanisław Lem
Her (2013) di Spike Jonze
The Imitation Game (2014) di Morten Tyldum
Code: Debugging the Gender Gap (2015) di Robin Hauser
Facebookistan (2015) di Jakob Gottschau
Ex Machina (2015) di Alex Garland
The Facebook Dilemma (2015) di Frontline PBS
The Cleaners (2018) di Moritz Riesewieck e Hans Block
Elon Musk: The Real Life Iron Man (2018) di Sonia Anderson
Upgrade (2018) di Leigh Whannell
Alita: Battle Angel (2019) di Robert Rodriguez
The Great Hack (2019) di Karim Amer e Jehane Noujaim
Coded Bias (2020) di Shalini Kantayya
The Social Dilemma (2020) di Jeff Orlowski
The Billion Dollar Code (2021) di Robert Thalheim
The Playlist (2022) di Per-Olav Sørensen e Hallgrim Haug
Joan Is Awful (Black Mirror 6x01, 2023) di Charlie Brooker, diretto da Ally Pankiw
Barbie – The Movie (2023) di Greta Gerwig
Poor Things! (2023) di Yorgos Lanthimos
Union (2024) di Brett Story e Stephen Maing



UNIONE EUROPEA
Fondo Sociale Europeo



La borsa di dottorato è stata cofinanziata con risorse del
Programma di finanziamento PON REACT-EU
Azione IV.4 – Dottorati su tematiche dell’Innovazione

Questa tesi riflette solo i punti di vista e le opinioni degli autori, né l’Unione Europea né la
Commissione Europea possono essere considerate responsabili per essi.

*The doctoral scholarship was co-funded with resources in the frame of the
PON REACT-EU financing Program
Action IV.4 – Doctorates on Innovation topics*

*This thesis reflects only the authors’ views and opinions, neither the European Union nor the
European Commission can be considered responsible for them.*

*The copyright of this thesis belongs to the author. Unless otherwise indicated, this document
is distributed under the license “Creative Commons CC BY, attribution”*