I cattivi maestri digitali: come gli algoritmi creano nuovi outsiders

Master of (digital) puppets: how algorithms create new outsiders

Rosario Palese, Giovanna Truda*

DOI: 10.14664/rcvs/268

Riassunto

Il contributo propone un riesame della *labelling theory* al fine di interpretare gli algoritmi di moderazione delle piattaforme video come attori classificatori che, attraverso pratiche di etichettamento automatizzato, contribuiscono alla produzione di outsider digitali. Integrando critical algorithm studies, studi sulla governance delle piattaforme e teoria sociologica classica, il lavoro mette in relazione questo quadro teorico con evidenze empiriche riguardanti YouTube, TikTok e Douyin, con particolare riferimento alla marginalizzazione algoritmica di minoranze LGBTQ+, persone con disabilità e gruppi etnici. L'analisi suggerisce che tali sistemi possano configurare forme specifiche di controllo sociale caratterizzate da elevata scala operativa, opacità strutturale e pratiche riconducibili a ciò che la letteratura definisce *algorithmic camouflage*. In questa prospettiva, le scelte distributive incorporate nei sistemi di moderazione tendono a presentarsi come esiti tecnici piuttosto che come decisioni normative. Il contributo applica sistematicamente la *labelling theory* ai contesti digitali, introducendo le nozioni di devianza primaria e secondaria algoritmica. Sulla base della letteratura esaminata, si propone di interpretare le piattaforme come dispositivi di stratificazione che possono contribuire all'amplificazione di disuguaglianze attraverso processi opachi e difficilmente contestabili.

Parole chiave: Teoria dell'etichettamento, controllo sociale algoritmico, imprenditori morali

Abstract

This article revisits labelling theory in order to interpret video platform moderation algorithms as classificatory actors that contribute to the production of digital outsiders. By integrating critical algorithm studies, platform governance research, and classical sociological theory, the study connects this framework to empirical evidence from YouTube, TikTok, and Douyin, focusing on the algorithmic marginalization of LGBTQ+ minorities, people with disabilities, and ethnic groups. The analysis suggests that these systems may configure specific forms of social control characterized by large-scale operation, structural opacity, and practices described in the literature as algorithmic camouflaging. Within this perspective, distributive choices embedded in moderation systems tend to appear as technical outcomes rather than as normative decisions. The article systematically applies labelling theory to digital contexts, introducing the notions of algorithmic primary and secondary deviance. On the basis of the reviewed literature, platforms are interpreted as stratification devices that may contribute to the amplification of inequalities through processes that are opaque and structurally difficult to contest.

Keywords: Labelling theory; Algorithmic social control; Moral entrepreneurs

* Rosario Palese: Universidad de Salamanca. E-mail: rosario.palese@usal.es Giovanna Truda: Università degli Studi di Salerno. Riconoscimenti: il contributo nasce da un lavoro congiunto degli autori. Rosario Palese ha curato i paragrafi 3, 4 e 5, oltre la nota metodologica; a Giovanna Truda si devono i paragrafi 1 e 2. Le conclusioni sono state redatte in collaborazione.

© The Author(s) 2026

Résumé
Cet article réexamine la théorie de l'étiquetage (*labelling theory*) afin d'interpréter les algorithmes de modération des plateformes vidéo comme des acteurs classificatoires qui, par des pratiques d'étiquetage automatisé, contribuent à la production d'outsiders numériques. En articulant les *critical algorithm studies*, la recherche sur la gouvernance des plateformes et la théorie sociologique classique, l'étude relie ce cadre théorique à des données empiriques portant sur YouTube, TikTok et Douyin, en accordant une attention particulière à la marginalisation algorithmique des minorités LGBTQ+, des personnes en situation de handicap et des groupes ethniques. L'analyse suggère que ces systèmes peuvent configurer des formes spécifiques de contrôle social caractérisées par une échelle opérationnelle élevée, une opacité structurelle et des pratiques que la littérature désigne sous le terme d'*algorithmic camouflaging*. Dans cette perspective, les choix distributifs intégrés dans les systèmes de modération tendent à se présenter comme des résultats techniques plutôt que comme des décisions normatives. L'article applique systématiquement la théorie de l'étiquetage aux contextes numériques en introduisant les notions de déviance primaire et secondaire algorithmique. Sur la base de la littérature examinée, les plateformes sont interprétées comme des dispositifs de stratification susceptibles d'amplifier les inégalités à travers des processus opaques et structurellement difficiles à contester.

Mots-clés: Théorie de l'étiquetage, Contrôle social algorithmique, Entrepreneurs de morale

---

1. Obiettivi del contributo e nota metodologica

Il lavoro si colloca all'intersezione tra sociologia classica della devianza e critical algorithm studies, proponendo una rilettura della *labelling theory* di Howard S. Becker alla luce delle trasformazioni connesse alla governance algoritmica delle piattaforme video. Si tratta di un contributo teorico-critico fondato su una revisione ragionata della letteratura. Non vengono presentati dati etnografici originali né analisi quantitative primarie; l'argomentazione si basa sull'integrazione critica di studi empirici già pubblicati, ricondotti a un quadro interpretativo unitario.

Gli obiettivi sono tre. In primo luogo, rielaborare le categorie di Becker (1963), Lemert (1967) e Goffman (1963) per l'analisi della moderazione algoritmica, discutendo in che misura gli algoritmi delle piattaforme video possano essere interpretati come imprenditori morali contemporanei e quali elementi li distinguano dai precedenti storici, in termini di scala operativa, opacità e modalità di legittimazione delle decisioni. In secondo luogo, introdurre le nozioni di devianza primaria e secondaria algoritmica come strumenti concettuali per descrivere i processi attraverso cui la classificazione automatizzata produce effetti identitari e traiettorie di adattamento tra creator appartenenti a gruppi marginalizzati. In terzo luogo, esaminare, attraverso la letteratura disponibile su YouTube, TikTok e Douyin, la presenza di pattern ricorrenti di marginalizzazione che coinvolgono minoranze sessuali e di genere, persone con disabilità e minoranze etniche, valutando l'applicabilità del framework teorico in contesti culturali differenti.

La selezione delle fonti privilegia contributi provenienti dai *critical algorithm studies*, dagli studi sulla *platform governance* e dalla sociologia della devianza, con particolare attenzione a ricerche empiriche su *shadowbanning*, demonetizzazione e riduzione algoritmica della visibilità. L'esposizione segue tale articolazione: (§1) ricostruzione del quadro teorico classico; (§2) transizione dal controllo sociale tradizionale a quello algoritmico; (§3) analisi della moderazione come forma di *governance* privata; (§4) discussione delle evidenze sul *bias* algoritmico; (§5) confronto *cross-platform*.

2. Introduzione alla teoria dell'etichettamento: fondamenti classici

Negli anni Sessanta la teoria dell'etichettamento introdusse una svolta decisiva negli studi sulla devianza, spostando l'attenzione dalle caratteristiche degli individui ai processi sociali che definiscono atti e attori come devianti. In contrasto con gli approcci positivisti centrati su patologia individuale, disorganizzazione

© The Author(s) 2026

sociale o deficit culturale, questa prospettiva interpreta la devianza come esito di pratiche definitorie e reazioni sociali (Becker, 1963; Lemert, 1967; Goffman, 1963). L'opera più influente in questo campo è *Outsiders: Studies in the Sociology of Deviance* di Howard Becker (1963). Attraverso studi etnografici su consumatori di marijuana e musicisti jazz, Becker mostrò che gli atti diventano devianti non per una loro qualità intrinseca, ma attraverso l'applicazione di regole sociali a determinate persone e comportamenti. Secondo Becker, la devianza non dipende dagli atti in sé, ma dal modo in cui gruppi sociali definiscono e applicano regole che etichettano alcuni individui come outsider.[16] Questa impostazione evidenzia che la devianza esiste solo in relazione a specifici pubblici, contesti sociali e rapporti di potere. Un concetto centrale del quadro teorico di Becker è quello di *imprenditori morali*: attori che promuovono e sostengono con successo la creazione e l'applicazione delle regole. Becker distingue tra due tipi di imprenditori morali: i creatori di regole, che si impegnano affinché le proprie preoccupazioni morali vengano tradotte in norme vincolanti, e gli applicatori di regole, che danno esecuzione a tali norme nei confronti di casi e individui specifici. I creatori di regole sono generalmente mossi da preoccupazioni morali su presunti problemi sociali, cercando di imporre la propria visione del comportamento appropriato agli altri. Gli applicatori di regole, come agenti di polizia, insegnanti o assistenti sociali, occupano posizioni che conferiscono loro l'autorità per identificare e gestire i trasgressori. Entrambe le categorie esercitano un potere distribuito in modo diseguale, con un'applicazione che colpisce in misura sproporzionata popolazioni già marginalizzate o stigmatizzate (Becker, 1963).

Becker descrive la devianza come un processo graduale in cui gli individui passano dalla conformità a modelli stabili di comportamento deviante in seguito alle reazioni sociali che ricevono. Riprendendo Lemert (1951), distingue tra devianza primaria, composta da trasgressioni iniziali con scarso impatto sull'identità, e devianza secondaria, che emerge quando l'individuo viene scoperto, etichettato e sanzionato, assumendo così un'identità deviante. L'etichetta funziona come uno status maestro che orienta la percezione altrui e l'autoconcezione del soggetto. La transizione alla devianza secondaria avviene attraverso un meccanismo di amplificazione (Lemert, 1967): le risposte stigmatizzanti, l'esclusione sociale e la mancanza di opportunità legittime spingono l'individuo verso sottoculture e carriere devianti, producendo una sorta di profezia che si autoavvera in cui il controllo sociale finisce per rafforzare ciò che tenta di prevenire.

Erving Goffman, con *Stigma: Notes on the Management of Spoiled Identity* (1963), ampliò ulteriormente la teoria dell'etichettamento analizzando i processi microsociologici attraverso cui gli individui stigmatizzati gestiscono la propria identità nell'interazione sociale. Goffman definisce lo stigma come *«un attributo profondamente screditante»*[17] e distingue tra individui discreditati, il cui attributo stigmatizzante è immediatamente visibile, e individui discreditabili che devono gestire le informazioni per evitarne la rivelazione. La sua analisi mostra il costo psicologico dello stigma, le strategie adottate per attenuarlo e i modi in cui un'identità stigmatizzata può finire per dominare l'autoconcezione individuale. Goffman individua diverse strategie di gestione dello stigma, tra cui il *passing* (occultare l'attributo stigmatizzante), il *covering* (renderlo meno evidente) e la *normification* (ridefinirlo come normale). Tuttavia, queste strategie richiedono spesso una vigilanza costante e un notevole lavoro psicologico; inoltre, il divario tra la identità sociale virtuale (ciò che gli altri si aspettano) e l'identità sociale reale (le caratteristiche effettive dell'individuo) può generare profonda ansia e alienazione (Goffman, 1963).

I contributi classici dell'etichettamento richiamano anche la dimensione politica dei processi di definizione della devianza. L'analisi di Cohen (1972) sulle *moral panics* mostra come media, retoriche politiche e forze dell'ordine possano amplificare timori pubblici e produrre risposte di controllo eccessive, come nel caso dei "mods" e dei "rockers", trasformati in folk devils che incarnavano ansie sociali più ampie. La teoria ha

[16] «Social groups create deviance by formulating the rules whose violation is defined as deviant and by applying those rules to specific individuals, labeling them as outsiders. In this perspective, deviance is not an inherent quality of the act committed, but the outcome of others applying rules and sanctions to an "offender." The deviant is the person to whom such a label has been successfully applied; deviant behavior is whatever people so label» (Becker, 1963, p. 9).

[17] «The term stigma, then, will be used to refer to an attribute that is deeply discrediting, but it should be seen that a language of relationships, not attributes, is really needed. An attribute that stigmatizes one type of possessor can confirm the usualness of another, and therefore is neither creditable nor discreditable as a thing in itself». (Goffman, 1963, p. 3).

© The Author(s) 2026

attirato critiche per aver trascurato i danni dei comportamenti criminali, per aver enfatizzato eccessivamente il potere delle etichette e per aver attribuito poca attenzione alle strutture economico-politiche (Taylor et al., 1973). Nonostante ciò, il suo contributo nel mettere in luce la costruzione sociale della devianza e la distribuzione asimmetrica del potere resta fondamentale. Le rielaborazioni contemporanee integrano prospettive come intersezionalità, *critical race theory* ed economia politica, mostrando come i processi di etichettamento si radichino in sistemi più ampi di disuguaglianza (Link & Phelan, 2001).

Le strategie di gestione dello stigma elaborate da Goffman trovano nell'ambiente digitale un ambito di applicazione particolarmente rilevante, pur assumendo configurazioni in parte nuove. Il *passing*, ossia l'occultamento dell'attributo stigmatizzante, si osserva nelle pratiche di alcuni creator LGBTQ+ che sostituiscono termini identitari nei titoli, nelle descrizioni o negli hashtag per ridurre il rischio di attivare filtri automatici. In questo contesto, parole come "gay" o "transgender" possono essere omesse o riformulate attraverso perifrasi considerate meno esposte ai sistemi di moderazione (Rodriguez, 2022). Il *covering*, che implica la riduzione della salienza dello stigma senza una sua completa rimozione, appare nelle forme di autocensura preventiva documentate tra creator appartenenti a gruppi marginalizzati. In tali casi, vengono modificati registro, stile espressivo o selezione tematica in funzione di penalizzazioni algoritmiche percepite, anche in assenza di notifiche esplicite (Kingsley et al., 2022). La *normification*, intesa come tentativo di presentare l'attributo stigmatizzante entro codici culturalmente dominanti, può manifestarsi nell'adozione di format e linguaggi aderenti alla cultura mainstream della piattaforma, con l'obiettivo di evitare la classificazione come contenuto borderline.

Rispetto al quadro goffmaniano, il contesto algoritmico introduce tuttavia un elemento ulteriore. Le strategie di gestione dello stigma non sembrano dipendere esclusivamente dall'iniziativa del soggetto, ma risultano fortemente condizionate dalle architetture tecniche che regolano visibilità e monetizzazione. Zhao (2024) utilizza l'espressione *algorithmic camouflage* per descrivere pratiche di adattamento semantico e stilistico indotte dalla struttura dei sistemi di classificazione. In questa prospettiva, lo stigma non è soltanto gestito dall'attore sociale, ma contribuisce a essere strutturato e amministrato dall'infrastruttura piattaformica, con effetti che incidono sulle condizioni di sostenibilità economica e comunicativa. Tali dinamiche risultano rilevanti per comprendere le trasformazioni contemporanee del controllo sociale, in cui la definizione di ciò che è accettabile o deviante è sempre più mediata da sistemi automatizzati. L'analisi di queste nuove forme di potere normativo costituisce il passaggio logico del paragrafo successivo.

3. Dal controllo sociale tradizionale al controllo sociale algoritmico

Il passaggio dalle forme tradizionali di controllo sociale alla governance algoritmica può essere letto come una trasformazione significativa nelle modalità con cui le società definiscono e riproducono i propri ordini normativi. L'opacità dei sistemi algoritmici contribuisce a ciò che Pasquale (2015) ha definito una *black box society*, nella quale decisioni rilevanti per le traiettorie individuali vengono affidate a meccanismi difficilmente accessibili al controllo pubblico. Questa opacità non riguarda soltanto la dimensione tecnica, ma si intreccia con profili giuridici ed economici, legati a segreti industriali, tutele proprietarie e alla complessità dei sistemi di machine learning, talvolta poco trasparenti anche per i loro sviluppatori. In modo complementare, O'Neil (2016) ha evidenziato come determinati modelli algoritmici possano attivare circuiti di retroazione che rafforzano disuguaglianze già esistenti in ambiti quali giustizia penale, credito ed educazione. Le cosiddette "armi di distruzione matematica" vengono descritte come sistemi opachi, scalabili e difficilmente contestabili, capaci di produrre effetti strutturali mantenendo un'apparenza di neutralità tecnica. Trasposta nei contesti digitali, la teoria dell'etichettamento offre una chiave per interrogare continuità e discontinuità rispetto al controllo sociale tradizionale. Come gli imprenditori morali individuati dalla sociologia classica, anche gli algoritmi contribuiscono a definire ciò che è accettabile o deviante, classificando soggetti e contenuti. Tuttavia, tali processi si caratterizzano per elementi inediti: operano su scala elevata, si fondano su opacità sistemiche e tendono ad aggregare preoccupazioni morali eterogenee in categorie semplificate che risultano difficili da contestare (O'Neil, 2016). Nelle piattaforme digitali queste dinamiche assumono

© The Author(s) 2026

una rilevanza quotidiana, poiché la definizione algoritmica di ciò che è appropriato o problematico si traduce direttamente in pratiche ordinarie di governo del discorso pubblico.

Dimensione	Controllo sociale tradizionale	Controllo sociale algoritmico	Implicazioni teoriche
Agenti del controllo	Imprenditori morali umani (Becker)	Sistemi algoritmici automatizzati	Trasferimento dell'agency morale da attori identificabili a black boxes tecnologiche
Scala operativa	Locale, circoscritta, contestuale	Globale, massiva, scalabile	Amplificazione esponenziale del potere normativo senza corrispondente aumento di accountability
Trasparenza dei criteri	Norme esplicite, negoziabili, contestabili	Opaca (segreti industriali, complessità ML)	Erosione della possibilità di contestazione democratica delle regole
Natura della classificazione	Giudizi situati, modificabili nel tempo	Categorizzazioni automatizzate, persistenti, aggregate	Reificazione delle etichette devianti attraverso l'automazione
Retroazione	Circuiti di feedback limitati e visibili	Circuiti autoamplificanti invisibili (O'Neil)	Le "armi di distruzione matematica" perpetuano e intensificano disuguaglianze strutturali
Legittimazione	Autorità morale, consenso comunitario	Presunta oggettività scientifica	Naturalizzazione delle scelte normative come necessità tecniche

Tabella 1: Evoluzione del controllo sociale: dal tradizionale all'algoritmico

Nota: La tabella mette in luce la frattura tra controllo sociale storico e governance algoritmica, mostrando che quest'ultima costituisce un regime normativo autonomo. Mentre le pratiche tradizionali coinvolgevano attori riconoscibili e norme contestabili, i sistemi algoritmici operano attraverso un'opacità strutturale che Pasquale definisce "black box society". Cambia anche la retroazione: processi un tempo aperti al confronto si trasformano in circuiti auto amplificanti, quelle "armi di distruzione matematica" che, secondo O'Neil, producono danni proprio perché percepite come neutrali. Rimane centrale l'interrogativo politico: chi definisce le categorie, con quali valori e quali margini di contestazione? *Fonte*: Ns elaborazione.

Applicare le categorie di Becker al controllo algoritmico richiede tuttavia un'operazione più precisa di una semplice analogia: occorre identificare con chiarezza chi esercita le funzioni di rule creator e di *rule enforcer* nell'ecosistema delle piattaforme, e in che misura tali funzioni si distinguano, per continuità e per rottura, dai loro equivalenti storici. Sul versante dei *rule creators*, la funzione creatrice di regole non risiede nell'algoritmo in sé, che è uno strumento, ma negli attori che ne definiscono i parametri: le imprese che gestiscono le piattaforme, i team di *policy* che traducono orientamenti valoriali e interessi commerciali in criteri operativi di moderazione, e gli inserzionisti che esercitano pressioni normative determinanti attraverso la logica della *brand safety*. Quando un'azienda pubblicitaria ritira i propri investimenti da contenuti associati a determinate identità, non compie un atto neutro: produce una regola de facto che l'algoritmo di demonetizzazione applicherà su scala globale. Gli imprenditori morali contemporanei non sono quindi le macchine, ma i soggetti economici e istituzionali che le programmano, le finanziano e ne orientano le priorità, con la differenza cruciale, rispetto ai rule creators beckeriani, che la loro funzione normativa è sistematicamente oscurata da retoriche di neutralità tecnica (Pasquale, 2015; Caplan & Gillespie, 2020).
Sul versante dei rule enforcers, gli algoritmi di moderazione e raccomandazione applicano le regole con caratteristiche che nessun enforcer umano ha mai posseduto: scala massiva, velocità di classificazione istantanea, assenza di discrezionalità situata e impossibilità di negoziazione da parte del soggetto classificato. L'agente di polizia di Becker poteva essere interpellato, contrattato, corrotto o convinto; l'algoritmo non è accessibile, non risponde e non motiva le proprie decisioni. Questa incontestabilità strutturale non è un

© The Author(s) 2026

difetto tecnico ma una caratteristica funzionale: produce un regime di enforcement che si autolegittima attraverso l'apparenza di oggettività scientifica, trasformando scelte morali e politiche in presunte necessità computazionali (O'Neil, 2016).
Gli elementi di continuità con il quadro beckeriano sono reali e significativi: come i moral entrepreneurs storici, gli algoritmi applicano le regole in modo asimmetrico, colpendo in misura sproporzionata popolazioni già marginalizzate; come gli enforcers tradizionali, operano con criteri che riflettono i valori e gli interessi dei gruppi dominanti; come i processi di etichettamento classici, producono effetti identitari che trascendono il singolo atto sanzionato. Gli elementi di discontinuità sono tuttavia altrettanto fondamentali e definiscono la specificità del controllo algoritmico come regime normativo autonomo: la scala operativa globale senza corrispondente accountability; l'opacità sistemica che impedisce la contestazione democratica delle classificazioni; la capacità di mascherare la propria natura politica non attraverso il silenzio, come accadeva ai rule enforcers tradizionali, ma attraverso la produzione attiva di un discorso di neutralità tecnica che nega strutturalmente l'esistenza stessa di una scelta normativa.

4. Governance delle piattaforme e moderazione dei contenuti

La moderazione dei contenuti, intesa come insieme di pratiche volte a filtrare, valutare e, se necessario, rimuovere o limitare i contenuti prodotti dagli utenti, rappresenta una forma di governance privata del discorso pubblico. Come osserva Klonick (2017), essa incide direttamente su questioni quali libertà di espressione, partecipazione democratica e ruolo delle imprese nella regolazione degli spazi comunicativi.
Nel suo studio sulla moderazione, Gillespie (2018) mostra come il controllo dei contenuti non costituisca un'attività accessoria, ma un elemento strutturale dell'identità e del modello economico delle piattaforme. Moderare significa, in questa prospettiva, tutelare gli inserzionisti, rispettare obblighi legali, contenere molestie e contenuti dannosi, nonché preservare norme comunitarie coerenti con l'immagine di marca. Tali pratiche implicano inevitabilmente valutazioni su ciò che può circolare online. Sebbene le piattaforme tendano a presentarle come decisioni tecniche, Gillespie le interpreta come espressione di scelte morali, politiche ed economiche. L'autore individua inoltre tre tensioni ricorrenti: l'ambivalenza tra neutralità dichiarata e intervento attivo sul discorso; il bilanciamento tra vincoli normativi, interessi economici e impegni valoriali; e la difficoltà di estendere il giudizio umano alla scala globale delle piattaforme, con il conseguente ricorso a sistemi automatizzati caratterizzati da opacità e potenziale bias.
In lavori successivi, Gillespie (2022) approfondisce la nozione di "riduzione" come modalità intermedia di moderazione. Attraverso strumenti quali demonetizzazione, limitazioni algoritmiche, restrizioni d'età o esclusione dai sistemi di raccomandazione, le piattaforme possono contenere la diffusione di contenuti ritenuti problematici senza procedere alla loro rimozione formale. Questa forma di intervento, meno visibile e talvolta meno contestabile rispetto alla cancellazione esplicita, tende a non essere pienamente tracciata nei report di trasparenza, pur producendo effetti significativi sulla circolazione del discorso. L'analisi del framework delle "Four Rs" adottato da YouTube, *remove*, *raise*, *reward*, reduce, consente a Gillespie di articolare queste dinamiche. La rimozione è la misura più evidente e, almeno in parte, soggetta a procedure di appello, ma non rappresenta necessariamente l'intervento più frequente. Le strategie di promozione e ricompensa, come l'amplificazione algoritmica, le raccomandazioni o la monetizzazione, operano secondo criteri discrezionali che producono effetti tangibili in termini di visibilità e vantaggio competitivo, pur restando poco trasparenti nei loro meccanismi decisionali. La riduzione occupa una posizione particolarmente delicata. Attraverso demonetizzazione, limitazioni d'età o esclusione dalle raccomandazioni, essa comprime la diffusione dei contenuti in una zona di bassa visibilità e scarsa contestabilità, anche perché l'applicazione di tali misure non è sempre oggetto di notifica esplicita. Secondo Gillespie, riconoscere la riduzione come forma piena di moderazione è essenziale per comprendere la portata del potere esercitato dalle piattaforme. L'asimmetria tra le quattro strategie, con livelli differenti di trasparenza e riconoscibilità, suggerisce che il controllo possa risultare tanto più incisivo quanto meno appare come atto esplicito di censura.

© The Author(s) 2026

Il concetto di *platformization* proposto da van Dijck et al. (2018) descrive l'integrazione profonda delle piattaforme nelle istituzioni sociali, con effetti significativi su mercati, pratiche sociali e processi democratici. Secondo gli autori, la logica combinata di datificazione, mercificazione e selezione algoritmica ridefinisce le forme di partecipazione economica, politica e culturale, configurando la piattaformizzazione come un processo al tempo stesso tecnologico e istituzionale. In questa prospettiva, la governance delle piattaforme è attraversata da tre sistemi di valori in competizione: quelli di mercato, centrati su efficienza e profitto; quelli pubblici, orientati a privacy, trasparenza e uguaglianza; e quelli civici, legati a inclusione e *accountability*. Gli autori argomentano che la predominanza dei valori di mercato, radicata nelle grandi imprese tecnologiche statunitensi, tenda ad alimentare tensioni con gli obiettivi europei di tutela dei valori pubblici e civici, generando conflitti sulla distribuzione delle responsabilità regolative tra aziende, governi e società civile. In un lavoro successivo, van Dijck (2020) articola la piattaformizzazione su tre livelli – infrastrutturale, intermedio e settoriale – mostrando come le Big Five (Google, Meta, Amazon, Apple, Microsoft) esercitino controllo trasversale su tutti, concentrando potere e dati in ecosistemi proprietari. Questa prospettiva multilivello evidenzia che l'influenza delle piattaforme deriva meno dai singoli servizi che dall'integrazione sistemica che vincola utenti, istituzioni e interi settori.

Sistema di valori	Principi fondanti	Attori di riferimento	Logiche operative	Obiettivi prioritari	Tensioni generate
Valori di mercato	Efficienza, profitto, crescita, scalabilità	Big Tech (GAFAM), investitori, azionisti	Datificazione, mercificazione, ottimizzazione algoritmica	Massimizzazione del valore azionario, espansione della base utenti	Subordinazione di privacy, equità e democrazia agli imperativi commerciali
Valori pubblici	Privacy, trasparenza, uguaglianza, diritti fondamentali	Regolatori (UE), istituzioni pubbliche, garanti	Normative (GDPR, DSA, DMA), supervisione, sanzioni	Protezione dei cittadini, controbilanciamento del potere privato	Conflitto con la logica proprietaria e l'opacità algoritmica delle piattaforme
Valori civici	Inclusione, accountability, partecipazione democratica, empowerment	Società civile, comunità marginali, attivisti	Advocacy dal basso, audit indipendenti, mobilitazione	Equità distributiva, giustizia informativa, controllo democratico	Asimmetria di potere rispetto alle corporation; dipendenza da piattaforme che controllano

Tabella 2: Sistemi di valori in conflitto nella *platformization*

Nota: La tabella sintetizza il framework della platformization (van Dijck et al., 2018; van Dijck, 2020) mettendo in evidenza i tre sistemi valoriali che esso individua e la loro tensione strutturale. L'interpretazione proposta qui introduce un elemento ulteriore: questi sistemi non sono semplicemente "in competizione", ma gerarchicamente ordinati, poiché i valori di mercato risultano costitutivi dell'architettura stessa delle piattaforme. I valori pubblici e civici operano invece come forze esterne che faticano a incidere su ecosistemi proprietari orientati all'estrazione di valore. La concentrazione verticale e orizzontale delle Big Five lungo i livelli infrastrutturale, intermedio e settoriale produce dipendenze che rendono regolazione e resistenza civica intrinsecamente asimmetriche. Ne discende che il conflitto non oppone valori equivalenti, ma una logica endogena di mercato a due logiche esogene che tentano di civilizzare un sistema che non le contempla. Ciò apre la questione se una reale democratizzazione delle piattaforme sia possibile senza trasformarne proprietà e assetti di governance. *Fonte*: Ns elaborazione.

Bucher (2018) mostra come gli algoritmi non si limitino a organizzare relazioni già esistenti, ma partecipino attivamente alla loro produzione. Riprendendo la nozione di *technicity*, evidenzia che l'accesso alle reti sociali tramite piattaforme digitali ridisegna le modalità di connessione, comunicazione e percezione delle relazioni. Funzionalità come *"People You May Know"* non si limitano a suggerire contatti, ma contribuiscono a

© The Author(s) 2026

naturalizzare un'idea di amicizia calcolabile e misurabile, creando nuove normatività intorno alla socialità e oscurando le logiche commerciali che orientano tali suggerimenti. Parallelamente, il concetto di *affective encounters* descrive il modo in cui gli utenti vivono gli algoritmi come entità dotate di intenzionalità, sviluppando risposte emotive che influenzano la gestione delle interazioni digitali e la modellazione delle soggettività. L'automazione della moderazione dei contenuti ha attirato crescente attenzione (Gorwa et al., 2020; Katzenbach et al., 2020). Pur rendendo possibile la gestione su larga scala dei contenuti generati dagli utenti, i sistemi algoritmici introducono *bias*, errori e forte opacità. I modelli di machine learning, addestrati su dati distorti, tendono a riprodurre e amplificare tali distorsioni, con effetti sproporzionati sulle comunità marginalizzate. L'opacità che caratterizza molte di queste tecnologie ostacola la contestazione delle decisioni e la comprensione dei criteri che guidano gli interventi di moderazione. Gorwa e colleghi (2020) evidenziano le principali difficoltà tecniche e politiche: da un lato, l'incapacità degli algoritmi di cogliere contesti, sfumature culturali e ironia genera sia falsi positivi sia falsi negativi; dall'altro, emergono interrogativi sui valori incorporati nei sistemi di moderazione, sulla distribuzione del potere decisionale e sulla necessità di maggiore trasparenza, accountability e partecipazione delle comunità coinvolte.

5. *Bias* algoritmico e marginalizzazione digitale

Gli studi critici sugli algoritmi hanno mostrato come i sistemi di classificazione automatizzata tendano a produrre svantaggi sistematici per gruppi già collocati in posizioni marginali. In questa prospettiva, il bias non viene interpretato come semplice errore tecnico, ma come espressione di strutture sociali di disuguaglianza che trovano nelle architetture computazionali una forma di sedimentazione. L'analisi di Safiya Noble (2018) sui motori di ricerca evidenzia come la visibilità algoritmica privilegi prospettive bianche, maschili ed eteronormative, mentre contribuisce a sessualizzare o stigmatizzare le donne nere. Il caso delle query come "Black girls", associate a risultati pornografici, è letto dall'autrice non come un malfunzionamento isolato, ma come effetto di logiche economiche e culturali incorporate nei sistemi commerciali. In tale quadro, l'opacità algoritmica e la concentrazione di potere nelle grandi piattaforme favoriscono la riproduzione di tali distorsioni, rendendo plausibile l'ipotesi di considerare i motori di ricerca come dispositivi normativi che richiedono forme di regolazione orientate alla giustizia informativa. Zuboff (2019) offre una cornice di economia politica che amplia questa lettura. Il cosiddetto capitalismo della sorveglianza descrive un modello in cui i dati comportamentali vengono trasformati in previsioni commercializzabili e in strumenti di intervento sulle condotte. Ciò che l'autrice definisce potere strumentario si fonda su monitoraggio continuo, automazione e modulazione algoritmica dei comportamenti, sostenuti da marcate asimmetrie informative. In questa prospettiva, la questione non riguarda solo l'efficienza dei sistemi, ma le implicazioni per l'autonomia individuale e per le condizioni dell'autogoverno democratico.

Le analisi di Noble e Zuboff suggeriscono che la classificazione algoritmica non produca effetti uniformi, ma operi su più livelli, incidendo sia sulle traiettorie individuali sia sulle configurazioni collettive. Per articolare tali livelli può risultare utile ricorrere alla distinzione proposta da Lemert tra devianza primaria e devianza secondaria, reinterpretata nel contesto digitale. Si può parlare di devianza primaria algoritmica quando contenuti associati a identità marginalizzate vengono classificati come problematici indipendentemente dal loro contenuto effettivo. In linea con lo stigma by association descritto da Goffman (1963), non è l'atto in sé a determinare la sanzione simbolica, ma l'identità attribuita al soggetto. Determinati marcatori linguistici o identitari possono attivare filtri orientati alla brand safety senza una valutazione sostanziale del contesto. In termini beckeriani, la devianza non deriva dall'azione, ma dall'etichetta applicata; nei sistemi algoritmici tale etichetta può precedere l'analisi concreta del contenuto.

La devianza secondaria algoritmica riguarda invece gli effetti nel tempo della classificazione sistematica. La demonetizzazione o la riduzione di visibilità possono generare pressioni verso l'autocensura, modificando progressivamente l'espressione creativa. Se tali adattamenti si stabilizzano, le restrizioni vengono interiorizzate come condizioni normali di produzione, con una ridefinizione dell'identità professionale entro i confini percepiti del sistema (Kingsley et al., 2022). La carriera deviante, nel senso lemertiano, assume qui una configurazione peculiare: non si costruisce sull'assunzione dell'etichetta, ma sul suo costante

© The Author(s) 2026

evitamento. Tuttavia gli effetti sull'autonomia espressiva possono risultare comparabili. Per evitare sovrapposizioni concettuali è utile distinguere tra tre fenomeni spesso confusi nel dibattito pubblico. La devianza algoritmica riguarda la classificazione automatica di contenuti o identità come problematici. La discriminazione algoritmica implica un trattamento differenziale sistematico tra gruppi comparabili, con una chiara dimensione asimmetrica. L'esclusione algoritmica concerne infine la riduzione o soppressione della visibilità attraverso pratiche quali limitazione della reach o esclusione dai sistemi di raccomandazione. Tali livelli sono interconnessi ma non equivalenti. Un contenuto può essere classificato come problematico senza generare discriminazione strutturale; una discriminazione può non tradursi in esclusione totale; e forme di esclusione possono verificarsi anche in assenza di una classificazione esplicita di devianza, come suggeriscono gli studi sui limiti invisibili alla viralità su Douyin (Zhao, 2024). Mantenere distinte queste dimensioni consente una valutazione più rigorosa delle evidenze empiriche e riduce il rischio di inferenze teoriche non adeguatamente supportate.

Le ricerche sullo *shadow banning* descrivono una forma particolarmente complessa di marginalizzazione algoritmica, caratterizzata dalla soppressione della visibilità senza notifica esplicita. I contenuti restano formalmente accessibili, ma vengono esclusi da ricerca, raccomandazioni e sistemi di scoperta, risultando di fatto difficilmente rintracciabili dai pubblici potenziali. Savolainen (2022) interpreta questo fenomeno come una forma di "folklore algoritmico": di fronte a una governance opaca, gli utenti elaborano spiegazioni collettive basate su esperienze condivise, riconoscimento di pattern e processi di sense-making. Tali narrazioni non hanno soltanto una funzione interpretativa, ma contribuiscono all'orientamento pratico, al sostegno emotivo e alla costruzione di forme di resistenza simbolica. L'opacità che caratterizza queste pratiche è stata letta da Cotter (2021) in termini critici rispetto ai principi di buona governance. Attraverso il concetto di "black box gaslighting", l'autrice descrive situazioni in cui le piattaforme tenderebbero a negare o ridimensionare fenomeni percepiti dagli utenti, con effetti potenzialmente erosivi sulla loro capacità di critica. Il framework proposto da Delmonaco et al. (2024) contribuisce a sistematizzare il fenomeno, documentandone diffusione e impatti. In un sondaggio su oltre mille utenti, circa il 10% riferisce di aver sperimentato shadow banning, con incidenze più elevate tra persone non cisgender, utenti ispanici e individui che si identificano come Repubblicani. Gli autori suggeriscono che tale pratica risponda a una pluralità di interessi organizzativi, tra cui la gestione discreta di contenuti problematici e la prevenzione di comportamenti opportunistici da parte di attori malevoli. Al tempo stesso, vengono rilevati effetti negativi significativi: difficoltà di contestazione, stress emotivo, perdite economiche e incremento della sfiducia nei confronti delle piattaforme. Per gruppi già marginalizzati, la riduzione invisibile della visibilità può tradursi in una forma di esclusione dal discorso pubblico, resa più incisiva dalla mancanza di trasparenza.

Evidenze analoghe emergono dalle ricerche sui creator LGBTQ+. Rodriguez (2022) descrive il caso di YouTube come una strategia di incorporazione selettiva, in cui la valorizzazione pubblica della diversità coesiste con pratiche di demonetizzazione e limitazione della visibilità di contenuti queer. Video dedicati a educazione sessuale, identità transgender o sessualità lesbica risultano segnalati come inadatti agli inserzionisti con frequenza superiore rispetto a contenuti eterosessuali comparabili. L'autore collega tali disparità a presupposti eteronormativi incorporati nelle policy e nei sistemi automatizzati di classificazione, osservando che determinati termini tendono ad attivare flag algoritmici anche in contesti educativi o di empowerment. Ne deriverebbe una configurazione "bifronte" della cultura di piattaforma, in cui riconoscimento simbolico e penalizzazione economica possono coesistere.

Le interviste raccolte da Kingsley et al. (2022) rafforzano questa lettura. I creator colpiti da demonetizzazione automatica appartengono frequentemente a gruppi svantaggiati e percepiscono le classificazioni algoritmiche come veicolo di stereotipi. Contenuti LGBTQ+ vengono interpretati come eccessivamente sessualizzati, quelli legati alla cultura nera come violenti, quelli sulla salute mentale come oscuri. Gli autori documentano danni economici, riduzione della visibilità e conseguenze psicologiche quali stress, perdita di fiducia e senso di impotenza. Emergono inoltre pratiche diffuse di autocensura e adattamento linguistico che, nel lungo periodo, possono restringere l'offerta culturale disponibile per comunità marginalizzate.

© The Author(s) 2026

Reynolds e Hallinan (2024) analizzano infine forme di "accountability generata dagli utenti". Alcuni creator hanno condotto veri e propri audit indipendenti, testando sistematicamente termini e categorie per individuare quelli associati a maggiore rischio di demonetizzazione. Queste pratiche producono conoscenze collettive sui meccanismi di moderazione e suppliscono parzialmente alla scarsa trasparenza delle piattaforme. Tuttavia, gli autori ne evidenziano anche i limiti: tali iniziative richiedono risorse significative, non consentono di accedere alle infrastrutture profonde dei sistemi e possono finire per favorire strategie di adattamento più che trasformazioni strutturali. Nonostante ciò, rappresentano una forma rilevante di partecipazione critica che amplia la comprensione pubblica dei processi di classificazione algoritmica.

Forma di Marginalizzazione	**Meccanismo Tecnico**	**Gruppi Target**	**Effetti Documentati**
Bias nei motori di ricerca (Noble, 2018)	Logiche commerciali integrate nell'architettura di ranking	Donne nere, minoranze razziali	Sessualizzazione, stigmatizzazione, discriminazione informativa
Shadow Banning (Savolainen, 2022; Delmonaco et al., 2024)	Soppressione invisibile da ricerca/raccomandazioni	LGBTQ+, ispanici, gruppi politicizzati	10% utenti colpiti; stress, perdite economiche, "black box gaslighting"
Demonetizzazione sistematica (Rodriguez, 2022-23; Kingsley et al., 2022)	Flag automatici su termini identitari	Creator LGBTQ+, salute mentale, giustizia sociale	Penalizzazione economica, autocensura, riduzione visibilità
Incorporazione selettiva (Rodriguez, 2022-23)	Celebrazione retorica + discriminazione operativa	Comunità queer	Cultura "bifronte": valorizzazione simbolica, soppressione materiale
Camuffamento algoritmico (Zhao, 2024)	Manomissione keywords, hashtag-trappola, tetti invisibili viralità	Uomini gay (Douyin)	Discorso "de-gay": disumanizzazione, deemotionalizzazione
Duplice emarginazione (Lyu et al., 2024)	Falsi positivi da disabilità + moderazione indotta da troll	BlindTokers, persone con disabilità	Erosione fiducia, spirale stigmatizzante
Censura identitaria percepita (Ungless et al., 2024)	Gap tra esperienza diretta e credenze collettive	Non binari, disabili, minoranze etniche	Autocensura preventiva, sfiducia sistemica
Modifica comportamentale (Zuboff, 2019)	Capitalismo della sorveglianza, "potere strumentario"	Intera popolazione	Erosione autonomia, manipolazione, deficit democratico

Tabella 3: Marginalizzazione algoritmica - meccanismi, manifestazioni e resistenze

Nota: Le forme di resistenza - documentate (audit amatoriali (Reynolds & Hallinan, 2024), folklore algoritmico (Savolainen, 2022), creatività linguistica e migrazione verso piattaforme alternative (Zhao, 2024) - rappresentano risposte adattive strutturalmente asimmetriche rispetto al potere centralizzato delle piattaforme. La tabella evidenzia che la marginalizzazione algoritmica non deriva da errori contingenti ma da meccanismi tecnici strutturati che producono schemi ricorrenti di esclusione, operando come dispositivo di stratificazione digitale che riproduce disuguaglianze preesistenti. *Fonte:* Ns elaborazione.

© The Author(s) 2026

L'opacità algoritmica può essere letta non come un fenomeno unitario, ma come un insieme stratificato di non trasparenze che operano su piani differenti e tendono a rafforzarsi reciprocamente. Una prima dimensione è quella tecnica. La complessità dei sistemi di machine learning, talvolta difficilmente interpretabile anche per i loro stessi sviluppatori, viene spesso richiamata per spiegare la limitata spiegabilità delle decisioni automatizzate (Pasquale, 2015). A questa si affianca un'opacità di natura giuridica: il ricorso a segreti industriali e tutele proprietarie restringe l'accesso pubblico alle architetture algoritmiche e limita le possibilità di scrutinio e contestazione informata (van Dijck, 2020). Una terza dimensione riguarda il piano economico. Modelli di business fondati su asimmetrie informative contribuiscono a mantenere inaccessibili le logiche operative dei sistemi, con effetti che parte della letteratura collega all'estrazione di valore dai comportamenti degli utenti e all'orientamento delle loro preferenze (Zuboff, 2019). Sul piano operativo, pratiche come la riduzione della visibilità o lo shadowbanning non notificato introducono forme di intervento difficilmente verificabili dall'esterno. In tali situazioni, l'esperienza soggettiva degli utenti può non trovare corrispondenza nelle comunicazioni ufficiali delle piattaforme, come osservato da Cotter (2021) e Gillespie (2022). Si può inoltre individuare un'opacità di carattere valoriale. Presupposti normativi legati a modelli culturali dominanti possono essere incorporati nei sistemi di classificazione in modo poco visibile, presentandosi come esiti tecnici anziché come scelte discutibili sul piano pubblico (Noble, 2018; Bucher, 2018). Zhao (2024) propone infine di distinguere tra limiti strutturali della trasparenza e discrepanze tra principi dichiarati e pratiche effettive, suggerendo una forma di opacità riconducibile alla gestione selettiva dell'informazione.

Queste dimensioni indicano che la trasparenza algoritmica non può essere considerata una questione meramente tecnica. Essa coinvolge condizioni di accesso alla conoscenza, possibilità di verifica e margini di contestazione, con implicazioni dirette per il funzionamento degli spazi pubblici digitali.

6. Estensioni comparative: TikTok, disabilità e contesti culturali

La marginalizzazione dei creator LGBTQ+ osservata su YouTube, attraverso pratiche di demonetizzazione e restrizioni d'età, trova riscontri anche su TikTok, dove tuttavia il ventaglio dei gruppi coinvolti appare più ampio. Le evidenze raccolte in questi contesti hanno indotto parte della letteratura a interpretare gli algoritmi come attori regolativi che incorporano nelle proprie architetture specifiche configurazioni culturali, contribuendo a definire ciò che risulta visibile o marginale. Zhao (2024), analizzando Douyin, versione cinese di TikTok, introduce il concetto di *camuffamento algoritmico* per descrivere una forma di governance che combina tolleranza dichiarata e restrizioni selettive. L'etnografia condotta con utenti gay cinesi individua tre meccanismi ricorrenti. Il primo riguarda la manipolazione delle parole chiave, con sostituzioni semantiche che alterano espressioni legate all'identità o alla sessualità. Il secondo concerne hashtag apparentemente attivi che, secondo l'autore, segnalerebbero i contenuti per una revisione più intensa. Il terzo è rappresentato da limiti non dichiarati alla diffusione virale, per cui i contenuti circolano all'interno di comunità ristrette senza accedere ai sistemi di raccomandazione di ampia portata. Zhao distingue questa forma di opacità da quella tecnica descritta da Pasquale (2015), sottolineando la discrepanza tra principi dichiarati e pratiche effettive. In tale contesto, il discorso pubblico sull'identità gay risulterebbe modulato attraverso processi di problematizzazione, neutralizzazione linguistica e selezione di rappresentazioni stereotipiche. Riletto in chiave beckeriana, il caso può essere interpretato come una forma di imprenditoria morale esercitata attraverso infrastrutture tecniche, pur presentandosi come gestione neutrale dei contenuti.

Quando la moderazione algoritmica interagisce con la disabilità, le dinamiche si articolano ulteriormente. Lyu et al. (2024) mostrano come utenti non vedenti su TikTok sperimentino una duplice forma di marginalizzazione. Da un lato, si registrano errori di classificazione legati a sistemi addestrati su modelli visivi standardizzati. In un caso documentato, un'inquadratura atipica è stata interpretata come contenuto inappropriato per minori. Dall'altro lato, si osservano interazioni antagonistiche in cui commenti abilisti, formulati entro i limiti formali delle policy, provocano reazioni emotive che finiscono per attivare la moderazione contro le vittime stesse. Gli autori suggeriscono che tali meccanismi possano amplificare vulnerabilità preesistenti, richiamando in parte la nozione lemertiana di devianza amplificata.

© The Author(s) 2026

L'assenza di procedure di appello accessibili contribuisce a erodere la fiducia e favorisce la circolazione di spiegazioni condivise all'interno delle comunità interessate. In questo senso, il cosiddetto *folklore algoritmico* descritto da Savolainen (2022) assume un rilievo particolare nei contesti di disabilità, fungendo da risorsa interpretativa e da rete di sostegno. L'indagine quantitativa di Ungless *et al.* (2024), condotta su 627 utenti britannici di TikTok, offre un quadro sistematico di tali dinamiche. Gli utenti non binari riportano livelli più elevati di rimozione o soppressione dei contenuti; le persone con disabilità dichiarano esperienze di censura superiori anche controllando per la controversialità dei contenuti. I dati suggeriscono che l'etichettamento algoritmico possa operare non solo sul contenuto, ma anche sull'identità del creator, richiamando lo stigma by association di cui parlava Goffman (1963). Al contempo, emerge uno scarto tra esperienza diretta e percezione della censura. Alcuni utenti convinti dell'esistenza di una repressione sistematica non avevano necessariamente sperimentato rimozioni personali. Le percezioni sembrano formarsi anche attraverso osservazioni indirette e narrazioni condivise, confermando il ruolo del folklore algoritmico nella costruzione collettiva di senso. Questa produzione di conoscenza può offrire strumenti cognitivi e supporto emotivo, ma può anche favorire autocensura preventiva e diffusa sfiducia nei confronti delle piattaforme.

Dimensione	**YouTube**	**TikTok (Globale)**	**Douyin (Cina)**	**Pattern Convergenti**
Logica dominante	Monetizzazione selettiva	Visibilità algoritmica (For You Page)	Conformità culturale-politica	Esclusione economica/sociale/culturale
Meccanismo primario	Demonetizzazione, limited advertising (Rodriguez, 2022-23)	Shadow banning, soppressione FYP (Lyu et al., 2024; Ungless et al., 2024)	Camuffamento algoritmico: manomissione keywords, hashtag-trappola (Zhao, 2024)	Soppressione invisibile > rimozione esplicita
Trasparenza dichiarata	Policy pubbliche dettagliate ma applicazione opaca (Gillespie, 2022)	Community Guidelines generiche, algoritmo FYP black box (Ungless et al., 2024)	Norme vaghe su "contenuti appropriati" (Zhao, 2024)	Retorica inclusività vs pratica escludente
Ipocrisia operativa	Celebrazione Pride + demonetizzazione sistematica contenuti queer (Rodriguez, 2022)	Supporto pubblico diversità + shadow banning identitario (Ungless et al., 2024)	Tolleranza apparente + "discorso de-gay" sistematico (Zhao, 2024)	Camuffamento universale dell'esclusione
Gruppi target primari	LGBTQ+, salute mentale, attivismo (Rodriguez, 2022-23; Kingsley et al., 2022)	LGBTQ+, disabilità, minoranze etniche, utenti non binari (Lyu et al., 2024; Ungless et al., 2024)	Uomini gay, dissidenza politica (Zhao, 2024)	Minoranze sessuali across platforms
Bias tecnici incorporati	Flag automatici su termini identitari: "gay",	Falsi positivi da disabilità; presupposti	Sostituzione semantica forzata: "coming out" →	Training data normativi riflettono maggioranze

© The Author(s) 2026

	"transgender", "queer" (Rodriguez, 2023)	abilisti nel training (Lyu et al., 2024)	"armadio" (Zhao, 2024)	
Amplificazione sociale	Inserzionisti pressano per "brand safety" (Caplan & Gillespie, 2020)	Troll abilisti strumentalizzano moderazione contro vittime (Lyu et al., 2024)	Segnalazioni massive attivano censura accurata (Zhao, 2024)	Interazione algoritmi-attori sociali ostili
Temporalità controllo	Retroattiva: contenuto pubblicato → demonetizzato (Rodriguez, 2022)	Mista: pre-moderazione + shadow banning post-pubblicazione (Ungless et al., 2024)	Predittiva: tetti invisibili prevengono viralità (Zhao, 2024)	Shift verso anticipazione della devianza
Forme resistenza	Audit terminologici user-generated, linguistic codes (Reynolds & Hallinan, 2024)	Teorie popolari, documentazione collettiva esperienze (Ungless et al., 2024)	Creatività linguistica, platform migration (Zhao, 2024)	Sense-making dal basso + adattamento tattico
Effetto identitario	Autocensura, perdite economiche, "morte economica" (Rodriguez, 2023; Kingsley et al., 2022)	Erosione fiducia, invisibilità sociale, "morte sociale" (Lyu et al., 2024)	Disumanizzazione, deemotionalizzazione, drammatizzazione (Zhao, 2024)	**Devianza secondaria algoritmica**
Possibilità appello	Limitata ma formalmente esistente (Gillespie, 2022)	Praticamente assente per shadow banning (Delmonaco et al., 2024)	Inaccessibile; linee guida oscure (Zhao, 2024)	Incontestabilità crescente
Contestualizzazione culturale	Pressioni inserzionisti + brand safety occidentale (Caplan & Gillespie, 2020)	Logiche algoritmiche globalizzate + moderazione culturalmente variabile (Ungless et al., 2024)	Governance autoritaria + norme confuciane su "armonia sociale" (Zhao, 2024)	Architetture globali, discriminazioni locali

Tabella 4 - Comparativa Cross-Platform

Nota: Nonostante le differenze culturali e politiche, YouTube, TikTok e Douyin condividono una medesima logica di governance che maschera l'esclusione sistematica. Le piattaforme promuovono retoriche di inclusività mentre i loro algoritmi marginalizzano gruppi formalmente protetti (Rodriguez, 2022; Zhao, 2024; Ungless et al., 2024). Il tratto comune è la soppressione invisibile: YouTube demonetizza senza rimuovere, TikTok limita la visibilità escludendo dal For You Page, Douyin manipola le parole chiave. Tale convergenza indica che l'occultamento algoritmico non deriva da contesti autoritari specifici, ma costituisce una caratteristica strutturale della governance delle piattaforme video. Ne risulta un'ipocrisia operativa in cui inclusione dichiarata ed esclusione praticata coesistono. A differenza degli

© The Author(s) 2026

imprenditori morali classici che agivano pubblicamente, gli algoritmi esercitano potere normativo negando la propria natura politica attraverso retoriche tecniche. Si osserva inoltre un passaggio dalla moderazione retroattiva alla classificazione predittiva della devianza. Le forme di resistenza basate su folklore algoritmico e audit informali restano deboli rispetto all'asimmetria del potere delle piattaforme (Reynolds & Hallinan, 2024; Savolainen, 2022). *Fonte*:: Ns elaborazione.

7. Conclusioni

Questo studio ha riconfigurato la teoria dell'etichettamento di Howard Becker per analizzare come gli algoritmi di moderazione e raccomandazione delle piattaforme video operino come imprenditori morali contemporanei, trasformando content creator e utenti in outsiders digitali attraverso meccanismi di classificazione opachi e incontestabili. L'analisi ha documentato continuità e rotture rispetto al controllo sociale tradizionale: come i moral entrepreneurs descritti da Becker (1963), gli algoritmi definiscono norme e classificano soggetti e contenuti; tuttavia operano con caratteristiche inedite, scala massiva, opacità sistemica e capacità di mascherare la propria natura politica attraverso retoriche di neutralità tecnica. Le evidenze empiriche attraverso YouTube (Rodriguez, 2022; 2023; Kingsley et al., 2022; Reynolds & Hallinan, 2024), TikTok globale (Lyu et al., 2024; Ungless et al., 2024) e Douyin (Zhao, 2024) suggeriscono pattern convergenti di marginalizzazione che investono minoranze sessuali, persone con disabilità, minoranze etniche e attivisti. Il pattern più significativo emerso dall'analisi comparata è una preferenza sistematica, documentata su tutte e tre le piattaforme esaminate, per forme di soppressione invisibile piuttosto che rimozione esplicita: demonetizzazione senza eliminazione su YouTube, esclusione dal For You Page su TikTok, pratiche di camuffamento su Douyin. Il fatto che tali dinamiche si riscontrino in contesti culturali e politici differenti suggerisce, pur con le cautele derivanti dall'assenza di ricerca primaria originale, che esse possano rappresentare una caratteristica strutturale della governance delle piattaforme video in presenza di contenuti considerati borderline.

Il concetto di algorithmic camouflage elaborato da Zhao (2024) consente di distinguere tra opacità tecnica e discrepanze tra dichiarazioni pubbliche e pratiche operative. Studi su YouTube (Rodriguez, 2022) e TikTok (Lyu et al., 2024) evidenziano dinamiche in cui forme di moderazione automatizzata interagiscono con pressioni sociali e culturali, producendo effetti differenziati su gruppi marginalizzati. Rispetto agli imprenditori morali tradizionali, gli algoritmi si caratterizzano per una minore visibilità del processo decisionale, il che rende più complessa la contestazione delle classificazioni.

La distinzione di Lemert (1967) tra devianza primaria e secondaria assume, in questo contesto, una configurazione specifica. La devianza primaria algoritmica si manifesta quando contenuti associati a identità marginalizzate vengono classificati automaticamente come problematici indipendentemente dal loro contenuto effettivo, configurando stigma by association (Goffman, 1963) in cui l'affiliazione identitaria stessa diventa marker di devianza. La transizione alla devianza secondaria algoritmica avviene attraverso meccanismi di amplificazione invisibile: shadow banning senza notifica (Delmonaco et al., 2024), demonetizzazione automatizzata senza spiegazioni granulari (Rodriguez, 2022-23), manomissione di keywords che nega la propria esistenza (Zhao, 2024). Questa opacità strutturale genera autocensura preventiva (Kingsley *et al.*, 2022), erosione della fiducia istituzionale (Lyu *et al.*, 2024) e interiorizzazione delle restrizioni come naturali (Zhao, 2024), configurando profezia che si autoavvera in cui il controllo algoritmico produce marginalizzazione identitaria ed esclusione dal discorso pubblico.

Le implicazioni per la democrazia digitale eccedono la questione tradizionale della censura. Quando piattaforme come YouTube e TikTok assumono un ruolo infrastrutturale per l'espressione culturale e il sostentamento economico, le loro scelte di moderazione acquisiscono rilevanza pubblica, pur restando collocate in ambito privato (van Dijck *et al.*, 2018). In tale quadro, l'equilibrio tra tutela degli inserzionisti, gestione dei rischi e pluralità espressiva diventa una questione centrale. Riconoscere la natura normativa e politica degli algoritmi non implica tuttavia ignorare i vincoli operativi entro cui le piattaforme gestiscono contenuti su scala globale. La moderazione algoritmica risponde anche a esigenze reali di contrasto all'hate

© The Author(s) 2026

speech, alla disinformazione e agli abusi, esigenze che a quella scala rendono l'automazione non solo una scelta politica ma anche un vincolo tecnico difficilmente eludibile con strumenti esclusivamente umani.
Il punto critico, nell'ottica qui adottata, non è l'automazione in sé, ma l'assenza di accountability democratica sui criteri che ne governano l'implementazione: è possibile progettare sistemi automatizzati di moderazione che siano al tempo stesso efficaci nel contrasto ai contenuti dannosi e trasparenti nei meccanismi di classificazione delle identità. La tensione non è tra sicurezza e libertà di espressione, ma tra governance privata opaca e governance pubblica accountable, una distinzione che il dibattito regolatorio europeo sta cominciando ad articolare.
Le forme di resistenza documentate, audit amatoriali (Reynolds & Hallinan, 2024), folklore algoritmico (Savolainen, 2022), teorie popolari (Lyu *et al.*, 2024), rivelano capacità innovative delle comunità marginalizzate di produrre conoscenza collettiva, ma presentano limiti strutturali: non possono penetrare infrastrutture opache e rischiano di contribuire più all'adattamento che alla resistenza trasformativa.
Questo studio presenta limiti che indicano direzioni per ricerca futura. L'analisi si è concentrata su letteratura esistente senza condurre ricerca etnografica originale; studi futuri potrebbero approfondire meccanismi di marginalizzazione attraverso osservazione partecipante e interviste con creator colpiti da moderazione algoritmica.
Un limite connesso riguarda la rappresentazione delle forme di resistenza: la letteratura esaminata documenta prevalentemente strategie adattive – folklore algoritmico, audit amatoriali, creatività linguistica – che operano entro i confini del sistema piuttosto che trasformarlo. Esistono tuttavia casi in cui la pressione dal basso ha prodotto cambiamenti nelle policy delle piattaforme: le campagne di documentazione condotte da creator LGBTQ+ su YouTube tra il 2019 e il 2021 hanno contribuito a rendere pubblico il problema della demonetizzazione sistematica, portando la piattaforma ad annunciare revisioni dei propri sistemi di classificazione automatica, sebbene l'effettiva implementazione di tali revisioni rimanga oggetto di dibattito (Reynolds & Hallinan, 2024). Analogamente, le indagini giornalistiche e accademiche sul shadowbanning di TikTok hanno preceduto e in parte alimentato l'apertura di procedimenti regolatori in sede europea. Questi casi suggeriscono che la resistenza dal basso possa produrre effetti istituzionali, specialmente quando si connette con advocacy organizzata e pressione regolatoria esterna. Studi futuri potrebbero approfondire sistematicamente le condizioni in cui il folklore algoritmico e gli audit amatoriali riescono a tradursi in accountability reale, distinguendo le condizioni di resistenza adattiva da quelle di resistenza trasformativa.
La dimensione comparativa cross-platform richiederebbe approfondimento attraverso analisi sistematica delle policy e audit tecnici degli algoritmi di raccomandazione. L'intersezionalità delle identità marginalizzate merita attenzione specifica: come interagiscono razza, genere, disabilità e orientamento sessuale nell'amplificare vulnerabilità alla classificazione algoritmica? La dimensione longitudinale resta infine sottosviluppata: come evolvono nel tempo le strategie di camuffamento delle piattaforme e le tattiche di resistenza delle comunità marginalizzate?
Sul piano delle direzioni future, un'area di ricerca particolarmente rilevante riguarda l'interazione tra i meccanismi di marginalizzazione algoritmica qui descritti e il quadro regolatorio europeo in evoluzione. Il Digital Services Act (DSA), entrato in vigore nel 2024 (European Parliament & Council of the European Union, 2022), introduce per le piattaforme di grandi dimensioni obblighi di trasparenza algoritmica, valutazione dei rischi sistemici e audit indipendenti che, almeno sul piano normativo, si rivolgono direttamente alle dinamiche di opacità analizzate in questo studio. L'AI Act (European Parliament & Council of the European Union, 2024), nelle sue disposizioni sui sistemi ad alto rischio, apre ulteriori questioni sulla classificabilità dei sistemi di moderazione automatizzata come tecnologie soggette a requisiti di spiegabilità e non discriminazione. È ancora prematuro valutare se questi strumenti regolatori siano sufficienti a incidere sui meccanismi strutturali di camuffamento algoritmico documentati dalla letteratura, la distanza tra obblighi normativi formali e pratiche operative effettive delle piattaforme è storicamente significativa, ma rappresentano un campo di osservazione privilegiato per verificare empiricamente la tenuta del framework qui proposto in un contesto di governance in trasformazione.

© The Author(s) 2026

L'applicazione della *labelling theory* ai contesti algoritmici rivela che le piattaforme video hanno ridefinito i confini tra normalità e devianza nella società contemporanea, configurando forme di controllo sociale che combinano caratteristiche inedite, automazione, scala, opacità, con logiche ricorrenti di stigmatizzazione ed esclusione. La sfida è trasformare piattaforme che operano come imprenditori morali ipocriti in infrastrutture che servano valori democratici di pluralità, equità e accountability. Finché il potere di classificazione resta concentrato in oligopoli privati opachi, il framework qui proposto suggerisce che la marginalizzazione algoritmica tenda a funzionare come meccanismo di stratificazione sociale digitale capace di riprodurre e amplificare disuguaglianze preesistenti, producendo outsiders digitali attraverso processi strutturalmente resistenti alla contestazione democratica.

Riferimenti bibliografici

Becker, H. S. (1963). *Outsiders: Studies in the sociology of deviance*. Free Press.

Bucher, T. (2018). *If...then: Algorithmic power and politics*. Oxford University Press. https://doi.org/10.1093/oso/9780190493028.001.0001

Cohen, S. (1972). *Folk devils and moral panics: The creation of the mods and rockers*. MacGibbon & Kee.

Cotter, K. (2021). "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *25*(11), 1685-1703. https://doi.org/10.1080/1369118X.2021.1994624

Cotter, K., Jonas, R., & De, A. (2025). "Why can't I just be fat and exist on the internet?": The embeddedness of antifatness on TikTok. *Platforms & Society*, *2*, 1–14. https://doi.org/10.1177/29768624251377163

Delmonaco, D., Haimson, O. L., Nie, P., & Wegner, A. (2024). Shadowbanning. *Business & Information Systems Engineering*, 66, 733-750. https://doi.org/10.1007/s12599-024-00905-3

European Parliament & Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). Official Journal of the European Union, L 277, 1–102. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689, 1–144. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221117552

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Prentice-Hall.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Katzenbach, C., Binns, R., Orwat, C., & Noordegraaf-Eelens, L. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Kingsley, S., Sannon, S., Schneider, G., Watson, N. F., Yue, S., & Haimson, O. L. (2022). "Give everybody [...] a little bit more equity": Content creator perspectives and responses to the algorithmic demonetization of content associated with disadvantaged groups. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), Article 412. https://doi.org/10.1145/3555149

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*(6), 1598-1670.

Lemert, E. M. (1951). *Social pathology: A systematic approach to the theory of sociopathic behavior*. McGraw-Hill.

© The Author(s) 2026

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology*, *27*, 363-385. https://doi.org/10.1146/annurev.soc.27.1.363
Lyu, Y., Cai, J., Callis, A., Cotter, K., & Carroll, J. M. (2024). "I got flagged for supposed bullying, even though it was in response to someone harassing me about my disability": A study of Blind TikTokers' content moderation experiences. *CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–24. https://doi.org/10.1145/3613904.3642148
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
Reynolds, C. J., & Hallinan, B. (2024). User-generated accountability: Public participation in algorithmic governance on YouTube. *New Media & Society*. https://doi.org/10.1177/14614448241251791
Rodriguez, N. S. (2022). LGBTQ incorporated: YouTube and the management of diversity. *Journal of Homosexuality*, *70*(9), 2030-2056. https://doi.org/10.1080/00918369.2022.2042664
Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *New Media & Society*, *24*(6), 1291-1309. https://doi.org/10.1177/01634437221077174
Taylor, I., Walton, P., & Young, J. (1973). *The new criminology: For a social theory of deviance*. Routledge & Kegan Paul.
Ungless, E. L., Markl, N., & Ross, B. (2024). Experiences of censorship on TikTok across marginalised identities. *arXiv*. https://doi.org/10.48550/arXiv.2407.14164
van Dijck, J. (2020). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, *22*(12), 2187-2206. https://doi.org/10.1177/1461444820940293
van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.
Zhao, L. (2024). Algorithmic camouflage: Exploring the shadowbans imposed by algorithms to moderate the content of Chinese gay men. *Big Data & Society*, *11*(October–December), 1–13. https://doi.org/10.1177/20539517241296037
Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

© The Author(s) 2026