Gestione automatizzata dei contenuti e discriminazione:

Come l'algoritmo può impattare la libertà di espressione dei gruppi marginalizzati.

di Pietro Dunn

1. Il ruolo della moderazione automatica nell'attuale contesto digitale

La moderazione dei contenuti rappresenta forse il ruolo più importante di una piattaforma online, di un social network o di un motore di ricerca. Invero, secondo Tarleton Gillespie[1], la moderazione è in sé stessa il cuore e l'essenza stessa del prodotto offerto da un intermediario digitale o da una azienda *Big Tech*.

La moderazione, da intendersi come l'insieme di quei meccanismi di *governance* che strutturano la partecipazione a una comunità online al fine di facilitarne la cooperazione interna e di prevenire abusi[2], è ciò che determina ciò che compare o non compare su un determinato sito internet, nonché in che modo i contenuti vengano organizzati ed effettivamente offerti al pubblico. È pertanto grazie alla moderazione che, per esempio, è possibile ridurre la diffusione di materiale illecito o comunque socialmente dannoso o pericoloso (ad esempio pedopornografia e pornografia non consenziente, discorsi d'odio, insulti etc.). Inoltre, è grazie alla moderazione (soprattutto algoritmica) che una piattaforma o un motore di ricerca sono capaci di offrirci quei materiali e contenuti che più ci interessano o più ci servono.

La moderazione dei contenuti, pertanto, è in sé e per sé uno strumento essenziale e ineludibile all'interno dell'attuale contesto digitale al fine di garantire la migliore esperienza possibile agli utenti e al pubblico in generale. Al contempo, a fronte dello straordinario incremento nell'utilizzo di internet che ha caratterizzato gli ultimi anni (e a fronte della crescente digitalizzazione che ha seguito la recente pandemia di COVID-19[3]), le piattaforme hanno fatto ricorso in modo sempre più massiccio a tecniche di moderazione automatizzata[4]. Tali tecniche vengono utilizzate non solo al fine di prendere decisioni relative alla conformità o meno di un contenuto alla legge o ai termini e condizioni della piattaforma stessa, e che conducano pertanto alla scelta di rimuovere o meno un contenuto o di sospendere o cancellare il profilo di un utente (*hard moderation* o moderazione in senso stretto), ma altresì al fine di creare una "gerarchia" tra i contenuti stessi per quanto concerne la visibilità degli stessi (*content curation* o *soft moderation*).

2. Benefici e rischi della moderazione automatica

L'utilizzo di tali sistemi di intelligenza artificiale (IA), da un lato, è essenziale a garantire una protezione quanto maggiore possibile dalla diffusione di materiali illeciti, dannosi o comunque pericolosi per gli individui e la società e, dall'altro lato, contribuisce a migliorare le condizioni di vita e di lavoro dei c.d. "moderatori umani". Il loro scopo è, in effetti, quello di filtrare e "scremare" i casi più manifesti e di portare all'attenzione del moderatore umano soltanto quei contenuti la cui illiceità o contrarietà ai termini e condizioni sia dubbia. In tal

modo, per esempio, i più lampanti episodi di pedopornografia o *hate speech* non dovranno essere visionati da un operatore umano, che potrebbe, a lungo andare, subire serie conseguenze anche sul piano psicologico[5].

Questi sistemi di IA, d'altro canto, si fondano su presupposti meramente probabilistici e sono, pertanto, esposti a un maggiore o minore rischio di errori. In particolare, laddove l'individuazione di un contenuto dannoso o illecito (quale per esempio è il caso dell'*hate speech* o dei cosiddetti "discorsi tossici") richieda un'accurata valutazione del contesto[6], l'utilizzo di macchine pone problemi significativi. Un algoritmo, per esempio, è difficilmente in grado di comprendere l'ironia di un messaggio, o di contestualizzare l'utilizzo di uno specifico termine.

Oltre a ciò, l'utilizzo di sistemi di *machine learning* e *deep learning*, quali le reti neurali, implica la necessità di ricorrere a complesse banche dati per insegnare agli stessi sistemi di IA a riconoscere i contenuti da moderare e da rimuovere. Se, però, il *dataset* utilizzato non è costruito in modo adeguato e rappresentativo delle modalità espressive di tutte le componenti sociali, si pone il concreto rischio di risultati di bassa qualità e, in taluni casi, discriminatori[7]. Numerosi studi, nel campo dell'intelligenza artificiale applicata alla moderazione dei contenuti, hanno in effetti confermato come, allo stato dell'arte, l'utilizzo di sistemi di rilevazione automatica di discorsi "tossici" o d'odio tendano ad avere ripercussioni particolarmente significative sulle categorie già tradizionalmente soggette a discriminazione e marginalizzazione[8].

Così, per esempio, Davidson *et al*.[9] hanno dimostrato come in numerosissimi casi gli algoritmi volti alla rilevazione di contenuti d'odio tendono a incorporare e riprodurre *bias* discriminatori nei confronti della comunità afro-americana. Invero, i *dataset* utilizzati da molteplici piattaforme sono costruiti utilizzando esempi e modelli tratti dal linguaggio comunemente utilizzato dalla comunità bianca. L'*African American English* (AAE), inteso come il tipico modo di parlare, comprensivo dello *slang*, tipico della comunità afro-americana, è generalmente sottorappresentato: quando si trovi dinnanzi a un esempio di contenuto in AAE, la probabilità che una macchina produca falsi positivi si alza notevolmente.

Un altro studio, pubblicato da Oliva *et al*.[10], ha allo stesso modo rilevato come un applicativo sviluppato da Google (Perspective) avesse la tendenza a sovrastimare la tossicità dei contenuti pubblicati dai membri della comunità LGBTQIA+ (in particolare da alcune celebri *drag queen* statunitensi) rispetto a quelli condivisi da noti esponenti di idee suprematiste. Secondo l'articolo, ampia letteratura in ambito sociolinguistico ha evidenziato il diffuso utilizzo interno alla comunità LGBTQIA+ di termini originariamente insultanti con il doppio fine di riappropriarsi di tali termini svuotandoli della loro carica denigratoria[11] e di aiutare gli stessi membri della comunità a sviluppare una "scorza dura". La finalità proattiva e pro-sociale dell'utilizzo di simili espressioni non è tuttavia facilmente individuabile da parte dell'algoritmo che, posto di fronte a termini quali "*bitch*", "*tranny*" o "*fag*" rischia di rilevare automaticamente la presenza di contenuti d'odio. Per di più, in taluni casi, l'utilizzo di *dataset* di qualità ridotta rischia di condurre al risultato per cui la semplice presenza di termini, anche neutri, facenti riferimento a una categoria discriminata tende a essere interpretata quale indizio della presenza di contenuti tossici o d'odio. Così, Perspective riteneva che il tweet

«and I'm……. GAY. #HairsprayLive»[12] presentasse un grado di tossicità del 92,31% (probabilmente per la sola presenza del termine "gay" scritto in maiuscolo) a fronte dell'11,71% rilevato per il tweet «Mixed-race children have higher rates of various dysfunctions – anyone can marry anyone, of course, but people should be aware of the risks»[13].

Ancora, nel marzo 2021 diversi social network (Instagram e Twitter *in primis*) sono stati aspramente criticati per avere filtrato automaticamente, e senza offrire alcuna spiegazione, qualsiasi contenuto pubblicato a sostegno della popolazione palestinese a seguito dell'occupazione da parte di Israele di alcune abitazioni nel quartiere di Sheikh Jarrah e a seguito delle conseguenti proteste. In un secondo momento, le piattaforme si sono scusate pubblicamente per la rimozione arbitraria di tali contenuti, adducendo come giustificazione la presenza di un *glitch* nei loro algoritmi[14].

Per di più, i problemi non si pongono solo con riferimento alla cosiddetta *hard moderation*, ma investono altresì la *content curation* (o *soft moderation*), ovverosia la modalità attraverso la quale i contenuti vengono organizzati e quindi presentati e diffusi tra lo stesso pubblico. Questo aspetto è particolarmente significativo in quanto, nei casi più seri, la *content curation* può addirittura portare allo *shadow banning* di un contenuto o di uno specifico account: in questi casi, nonostante non vi sia una sanzione esplicita nei confronti dell'utente (il suo contenuto non è rimosso, né il suo account è sospeso o cancellato), vi è la sostanziale impossibilità di raggiungere la propria *audience*. In caso di *shadow banning*, i contenuti pubblicati dal soggetto che ne è vittima rimarranno nascosti dal *feed* o dalla *homepage* dei destinatari, né il profilo apparirà tra i suggerimenti di ricerca.

Anche in questo caso, è stato rilevato in molti casi come l'utilizzo di sistemi di IA ai fini della *content curation* tenda in molti casi a replicare e amplificare forme di discriminazione e oppressione delle categorie tradizionalmente marginalizzate. Nel suo libro *Algorithms of Oppression: How Search Engines Reinforce Racism*[15], Safiya U. Noble mostra come, da un lato, motori di ricerca quali Google tendano a proporre immagini distorte e stereotipate della comunità afro-americana (soprattutto femminile) e come, dall'altro lato, siti dedicati alla pubblicizzazione delle attività commerciali sponsorizzino in modo ridotto gli esercizi gestiti da persone non bianche. Allo stesso modo, numerosi autori hanno sottolineato come l'utilizzo di sistemi di raccomandazione (*recommender systems*) da parte di piattaforme online e social network tendano a depotenziare notevolmente i contenuti proposti da attiviste femministe[16] o da comunità etnico-linguistiche minoritarie[17]. Questo a fronte, contestualmente, della tendenza dell'algoritmo a premiare proprio quei contenuti fortemente controversi, quali *hate speech* e/o *fake news*, che tendono in quanto tali a massimizzare l'*engagement* degli utenti (e quindi i profitti della piattaforma)[18].

3. Conclusioni

L'utilizzo di sistemi di intelligenza artificiale e di *machine learning* per finalità legate alla moderazione dei contenuti pubblicati dagli utenti in rete è quanto mai essenziale al giorno d'oggi, per far fronte allo straordinario flusso di informazioni che circolano quotidianamente in rete e, pertanto, per far fronte alla diffusione di materiali potenzialmente dannosi e

pericolosi per gli individui e la società. Al tempo stesso, l'utilizzo di questi sistemi non è esente dalla produzione di rischi significativi non solo per la libertà di espressione in generale, ma soprattutto per la libertà di espressione di quelle categorie della popolazione che, storicamente vittime di marginalizzazione, oppressione e discriminazione, richiederebbero invece una maggiore tutela a livello di diritti umani.

Se dunque, da un lato, non è certamente auspicabile una demonizzazione di tali sistemi, la cui utilità e necessità appare inevitabile, allo stesso tempo occorre una maggiore consapevolezza e attenzione relativa ai loro *side effects*. Invero, la Commissione Europea ha, in alcune recenti iniziative legislative, dimostrato di aver iniziato a fare sua tale consapevolezza. Così, il Regolamento (UE) 2021/784[19], relativo al contrasto della diffusione di contenuti terroristici online, prevede espressamente che un prestatore di servizi di *hosting* esposto a contenuti terroristici, nel porre in essere le specifiche misure volte alla minimizzazione del rischio di diffusione di tali contenuti, debba agire da un lato in una maniera che tenga pienamente conto dei diritti e degli interessi legittimi degli utilizzatori (in particolare, libertà di espressione e informazione, rispetto della vita privata e protezione dei dati personali) e debba dall'altro lato applicare tali misure in maniera non discriminatoria (Art. 5). Allo stesso tempo, la proposta di Regolamento volta alla promulgazione di un *Digital Services Act*[20] ha previsto l'introduzione di alcune garanzie a tutela dell'individuo i cui contenuti siano stati rimossi o il cui account sia stato sospeso, tra cui la previsione di un meccanismo attraverso il quale l'utente possa contestare le decisioni prese dal *provider* e ottenere una seconda decisione non ottenuta solamente tramite sistemi automatizzati (Art. 17).

Tale consapevolezza sembra, tuttavia, essere ancora ai suoi albori. Saranno invero necessari ulteriori e meno vaghi[21] interventi atti a incentivare la trasparenza nell'utilizzo dei sistemi automatizzati di moderazione e cura dei contenuti, nonché a tutelare, anche proceduralmente, gli individui da un'applicazione discriminatoria degli stessi.

---

[1] T. GILLESPIE, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018.

[2] J. GRIMMELMANN, *The Virtues of Moderation*, in *Yale Journal of Law and Technology*, n. 17, 2015, p. 42.

[3] G. DE GREGORIO *et al.*, *Big tech, così la pandemia ne ha accresciuto il potere: i rimedi che servono*, *Agenda Digitale*, 2021 https://www.agendadigitale.eu/cultura-digitale/big-tech-cosi-la-pandemia-ne-ha-accresciuto-il-potere-i-rimedi-che-servono/ (consultato 12/12/21).

[4] E. DOUEK, *Governing online speec: from"posts-as-trumps" to proportionality and probability*, in *Columbia Law Review*, n. 121, 3, 2021, p. 759–834; G. SARTOR-A. LOREGGIA, *The impact of algorithms for online content filtering or moderation. «Upload filters»* (Study Requested by the JURI Committee n. PE 657.101), Study Requested by the JURI Committee, European Parliament, 2021.

[5] S. T. ROBERTS, *Behind the screen: content moderation in the shadows of social media*, Yale University Press, New Haven; London, 2019; G. SARTOR-A. LOREGGIA, *The impact of algorithms for online content filtering or moderation. «Upload filters»* (Study Requested by the JURI Committee n. PE 657.101), cit.

[6] Ivi compresa la valutazione del tempo e del luogo in cui il contenuto è stato pubblicato, nonché dell'identità di chi lo abbia pubblicato e degli stessi destinatari.

[7] In tal senso, si parla sovente di "*garbage in, garbage out*", a indicare che l'utilizzo di dati scadenti in partenza determina risultati e decisioni automatizzate altrettanto scadenti.

[8] Suzor sottolinea per esempio come, durante il genocidio dei Rohingya in Myanmar (2017), l'algoritmo di Facebook abbia portato alla rimozione di numerosi contenuti pubblicati dagli stessi Rohingya per denunciare le violenze subite (mentre i materiali che incitavano all'odio e alla violenza nei loro confronti non venivano rimossi). N. P. SUZOR, *Lawless: The Secret Rules That Govern our Digital Lives*, Cambridge University Press, Cambridge, 2019.

[9] T. DAVIDSON *et al.*, *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, in *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, 2019, p. 25–35.
[10] T. DIAS OLIVA *et al.*, *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online*, in *Sexuality & Culture*, n. 25, 2, 2021, p. 700–732.
[11] Si pensi al termine "*queer*", inizialmente utilizzato avverso le persone non cis-gender e non eterosessuali e adesso facente parte integrante della sigla LGBTQIA+.
[12] Lett. "e sono……. GAY! #HairsprayLive".
[13] Lett. "I figli di razza mista hanno probabilità più alte di avere varie disfunzioni – certamente, ognuno si può sposare con chiunque, ma le persone dovrebbero essere consapevoli dei rischi".
[14] M. GEBEILY, *Instagram, Twitter blame glitches for deleting Palestinian posts*, *Reuters*, 2021 https://www.reuters.com/article/israel-palestinians-socialmedia-idUSL8N2MU624 (consultato 13/12/21).
[15] S. U. NOBLE, *Algorithms of oppression: how search engines reinforce racism*, New York University Press, New York, 2018.
[16] C. ARE, *The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram*, in *Feminist Media Studies*, n. 0, 0, 2021, p. 1–18.
[17] E. LLANSÓ *et al.*, *Artificial intelligence, Content Moderation, and Freedom of Expression*, TWG, 2020, 32 p.
[18] Ibidem.
[19] *Regolamento (UE) 2021/784 del Parlamento europeo e del Consiglio, del 29 aprile 2021, relativo al contrasto della diffusione di contenuti terroristici online (Testo rilevante ai fini del SEE)*, vol. LXIV OJ L 172 79–109, 2021 (2021).
[20] *Proposta di Regolamento del Parlamento Europeo e del Consiglio relativo a un mercato unico dei servizi digitali (legge sui servizi digitali) e che modifica la direttiva 2000/31/CE* (COM/2020/825 final).
[21] N. APPELMAN *et al.*, *Article 12 DSA: Will platforms be required to apply EU fundamental rights in content moderation decisions? – DSA Observatory*, *DSA Observatory*, s.d. https://dsa-observatory.eu/2021/05/31/article-12-dsa-will-platforms-be-required-to-apply-eu-fundamental-rights-in-content-moderation-decisions/ (consultato 03/12/21).