Gli algoritmi come decisori nel contrasto al terrorismo online. Alcune riflessioni a partire dal Regolamento (UE) 2021/784

di Micol Ferrario

Abstract: *Algorithms as decision-makers in the fight against online terrorism. Some considerations stemming from Regulation (EU) 2021/784* – Regulation (EU) 2021/784 mandates hosting service providers to promptly remove terrorist contents upon request by competent national authorities. While the Regulation addresses the crucial need to fight online terrorism, it raises significant concerns regarding the balance between security and human rights. A key issue lies in the extensive reliance on automated tools for content removal. This article critically examines the Regulation's provisions, highlighting the challenges posed by automated tools. To address this issue, the article proposes key solutions such as incorporating human review for complex cases, ensuring transparent content moderation processes, and establishing clear and timely appeal procedures.

Keywords: Regulation (EU) 2021/784; Online terrorism; Automated tools; Freedom of expression; Freedom of information

1. Introduzione

A partire dalla fine degli anni '80, internet si è rivelato essere uno dei più importanti mezzi di comunicazione esistenti e, ad oggi, viene infatti utilizzato da circa il 67.9% della popolazione globale[1]. Questo suo crescente uso è stato riscontrato anche nelle organizzazioni terroristiche che, in particolare dopo l'11 settembre, hanno accresciuto e affinato la loro presenza sul web, avvantaggiandosi di alcune delle sue caratteristiche innate come, per esempio, il fatto di essere scarsamente regolato o, ancora, di garantire in linea di massima l'anonimità degli *users*[2]. Nel corso degli anni, i terroristi si sono avvalsi di Internet per perseguire molteplici fini[3]. Tra gli altri, figura

[1] Secondo i dati forniti da Statista e aggiornati a febbraio 2025: https://www.statista.com/statistics/617136/digital-population-worldwide/.

[2] B. Todorovic, D. Trifunovic, *Prevention of (Ab-)Use of the Internet for Terrorist Plotting and Related Purposes*, in A.P. Schmid (Ed), *Handbook of Terrorism Prevention and Preparedness*, Aja, 2021, 594, 596.

[3] Per una ricostruzione puntuale si vedano A. Vedaschi, C. Graziani *Artificial Intelligence, Counter-Terrorism, and the Rule of Law. At the Heart of National Security*, Cheltenham, 2025, 7 ss. Sul punto si veda altresì G. Weimann, *Terror on the Internet: The New Arena, the New Challenges*, Washington, 2006. Un'analisi altrettanto interessante è offerta da

DPCE online
ISSN: 2037-6677

innanzitutto quello propagandistico, nel senso che i terroristi se ne sono crescentemente serviti per fare circolare le loro ideologie e i loro obiettivi. In secondo luogo, Internet ha rappresentato l'arena ideale per reclutare nuove leve, anche tramite la profilazione. Il mondo digitale è stato altresì sfruttato per condividere informazioni cruciali – come quelle relative alla costruzione di esplosivi – nonché per pianificare e coordinare plurimi attacchi terroristici. A tal fine, oltre ai propri siti web[4], i terroristi si sono largamente serviti dei social media e, in particolare, di Facebook, Telegram, X e YouTube. Se si considera che almeno il 59% dei cittadini dell'Unione europea[5] fa costantemente uso di questi canali, si comprende immediatamente l'ammontare del rischio di vulnerabilità alla propaganda terroristica. In considerazione di ciò, nel corso degli ultimi anni si sono infatti moltiplicate le iniziative per contrastare la minaccia terroristica online a livello internazionale, regionale e nazionale.

A livello regionale di Unione europea, tra gli sforzi più importanti[6] si registra l'adozione del Regolamento (UE) 2021/784[7], il cui principale obiettivo è quello di combattere la diffusione di contenuti terroristici online. Questo atto, che è divenuto applicabile a partire dal 7 giugno 2022, impone agli *hosting service providers* di rimuovere i contenuti terroristici dalle loro piattaforme entro un'ora dalla ricezione del relativo ordine da parte dell'autorità nazionale competente. Per individuarli e rimuoverli nel rispetto di questo rigoroso lasso di tempo, si è rivelato tuttavia necessario per gli *hosting service providers* prevalersi di strumenti automatizzati. In pratica, questi strumenti vengono utilizzati per monitorare i contenuti caricati sulle piattaforme, identificando quelli che potrebbero essere classificati come terroristici in base a criteri prestabiliti. Queste tecnologie sono fondamentali per la gestione di un flusso così elevato di contenuti e per garantire una rimozione rapida visto che la vastità dei dati, da un lato, e la necessità di una risposta immediata, dall'altro, sono difficili (se non impossibili) da gestire manualmente.

Seppure rappresenti un passo significativo nella lotta contro la propaganda terroristica online, il Regolamento (UE) 2021/784 presenta dei rischi significativi per l'uso di strumenti automatizzati con riferimento, in

M. Conway, *Terrorism and the Internet: New Media – New Threat?*, in 59(2) *Parl. Affs.* 283 (2006).

[4] G. Weimann, *www.terror.net*. *How Modern Terrorism Uses the Internet*, United States Institute of Peace Special Report, 2004, 116, 1.

[5] Secondo i dati forniti da Eurostat e aggiornati a marzo 2024: https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20240319-1.

[6] Tra i più recenti, si registra anche il Regolamento sui servizi digitali (Reg. UE n. 2065/2022 del P.E. e del Cons. del 19-10-2022, relativo a un mercato unico dei servizi digitali) che, pur non essendo specificamente focalizzato sulla lotta al terrorismo, vi contribuisce indirettamente introducendo un obbligo per le piattaforme digitali di valutare i rischi dei contenuti pubblicati e di motivare la loro eventuale rimozione. Altrettanto rilevante è l'ancora più attuale Regolamento sull'intelligenza artificiale (Reg. UE n. 1689/2024 del P.E. e del Cons. del 13-6-2024, che stabilisce regole armonizzate sull'intelligenza artificiale) che, dal momento in cui entrerà in vigore, avrà importanti ripercussioni sul possibile uso di strumenti biometrici. Per un'ampia trattazione, si vedano A. Vedaschi, C. Graziani, *op. cit.*, 63 ss.

[7] Reg. UE n. 784/2021 del P.E. e del Cons. del 29-4-2021, relativo al contrasto della diffusione di contenuti terroristici online.

particolare, alla precisione dei sistemi impiegati, alla protezione di taluni diritti individuali ed al rispetto della trasparenza nel processo decisionale.

Prendendo come caso studio il Regolamento (UE) 2021/784, questo contributo si propone di mettere in luce i rischi sottesi all'uso di strumenti automatizzati nel contesto della lotta al terrorismo online. A tal fine, dopo avere analizzato i caratteri salienti e le criticità del Regolamento (UE) 2021/784 (§ 2), l'articolo si focalizza sull'obbligo della rimozione entro un'ora, sulle tecnologie impiegate per assolvervi e sui rischi tecnici, giuridici ed etici a ciò connessi (§ 3). Seguiranno in chiusura alcune riflessioni conclusive (§ 4).

2. Il Regolamento (UE) 2021/784 sulla rimozione dei contenuti terroristici online: genesi, contenuto e criticità

Il processo di adozione del Regolamento (UE) 2021/784 è stato lungo e controverso e ha visto il coinvolgimento di diversi attori della governance digitale e di associazioni a tutela dei diritti fondamentali[8]. Le sue origini risalgono al 2017, segnatamente a quando la Commissione europea ha pubblicato la comunicazione sulla lotta ai contenuti illeciti online[9] con cui ha invitato per la prima volta le piattaforme ad essere più proattive nella messa a punto di strategie e di strumenti per individuare e rimuovere prontamente i contenuti illegali, anche di matrice terroristica[10]. Poco dopo, nel marzo 2018, la Commissione è nuovamente intervenuta adottando una raccomandazione[11] per incoraggiare ulteriormente gli Stati membri ad implementare delle misure effettive per contrastare i contenuti illeciti online, fornendo altresì raccomandazioni specifiche con riferimento a quelli terroristici. Dopo avere constatato che gli strumenti non vincolanti erano insufficienti, anche su invito del Consiglio europeo[12], nel settembre 2018 la Commissione presenta una prima bozza di regolamento specificamente incentrata sul contrasto alla diffusione dei contenuti terroristici online. Questo primo testo prevedeva che i fornitori di servizi di *hosting* fossero tenuti a rimuovere entro il termine di un'ora (decorrente dall'emissione dell'ordine di rimozione da parte dell'autorità nazionale competente) i contenuti terroristici pubblicati sulle loro piattaforme, nonché l'obbligo di adottare delle misure proattive per contrastarne la diffusione. Poco dopo la sua pubblicazione, diversi attivisti per i diritti digitali[13] hanno sottoscritto

[8] C. Graziani, *Intelligenza artificiale e fonti del diritto: verso un nuovo concetto di* soft law*? La rimozione dei contenuti terroristici online come* case-study, in *DPCE Online*, 2022, n. spec., 1473, 1479.
[9] Com. UE n. 555/2017 della C.E. del 28-9-2017 sulla lotta ai contenuti illeciti online.
[10] Questa Comunicazione è stata il risultato del vertice del Consiglio europeo di giugno 2017, in occasione del quale è stato chiesto alle aziende tecnologiche di creare nuovi strumenti per favorire la lotta alla propaganda terroristica online e si è, per la prima volta, profilata la possibilità di adottare una legislazione europea in tal senso. Per maggiori informazioni sul punto si consultino: concl. del Cons. del 22/23-6-2017.
[11] Racc. UE n. 334/2018 della C.E. dell'1-3-2018 sulle misure per contrastare efficacemente i contenuti illegali online.
[12] Concl. del Cons. del 28-6-2018.
[13] Si pensi, per esempio, alla lettera sottoscritta da *Access Now* e altre organizzazioni e singole persone con sui è stato richiesto di emendare il Regolamento sotto plurimi

DPCE online
ISSN: 2037-6677

una lettera aperta con cui hanno esortato i Ministri dell'UE a modificare il contenuto della suddetta proposta, sollevando preoccupazioni con riferimento, tra le altre cose, al fatto che il concetto di "contenuto terroristico" fosse troppo vago o, ancora, al termine perentorio di un'ora. In questa occasione, anche l'Agenzia europea per i diritti fondamentali (FRA) e i relatori speciali delle Nazioni Unite (ONU) sulla libertà di espressione, sul diritto alla privacy e sulla lotta al terrorismo hanno espresso le loro perplessità in merito al contenuto della proposta[14]. Viste le criticità sollevate, la Commissione ha quindi esortato il Consiglio dell'Unione europea (d'ora innanzi, il Consiglio) e il Parlamento a concordare un testo finale prima delle elezioni di quest'ultimo (previste per maggio 2019). Mentre il Consiglio ha prontamente adottato la sua posizione accogliendo quasi interamente il testo della Commissione[15], il Parlamento ha pubblicato la sua prima lettura solo nell'aprile 2019, suggerendo tra l'altro svariati emendamenti[16]. È stata proprio la divergenza tra la posizione di quest'ultimo, da un lato, e quella della Commissione e del Consiglio, dall'altro, ad avere notevolmente prolungato la durata dei negoziati. Questo processo si è infatti concluso soltanto nel dicembre 2020, quando finalmente il Parlamento, la Commissione e il Consiglio hanno trovato un accordo sul testo finale del Regolamento, che è stato firmato il 21 aprile 2021 ed è diventato applicabile a partire dal 7 giugno 2022.

Nella sua formulazione attuale, il Regolamento (UE) 784/2021 stabilisce delle norme uniformi per garantire la rimozione tempestiva dei contenuti terroristici dal web. In particolare, esso impone agli *hosting service providers* di eliminare tali contenuti o di disabilitarne l'accesso entro massimo un'ora dalla loro identificazione e segnalazione ad opera dell'autorità nazionale competente[17]. In aggiunta, gli *hosting service providers* che sono

profili: *Access Now* et al., *Joint Letter on Terrorist Content Regulation*, Brussels, 4-12-2018, https://www.accessnow.org/press-release/joint-letter-opposing-the-proposed-terrorist-content-regulation/

[14] Circa le perplessità dell'Agenzia europea per i diritti fondamentali (FRA) si veda: Fundamental Rights Agency, *Proposal for a Regulation on Preventing the Dissemination of Terrorist Content Online and its Fundamental Rights Implications*, Opinion n. 2/2019, 12-2-2019, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-opinion-online-terrorism-regulation-02-2019_en.pdf. Con riferimento a quelle dei Relatori speciali delle Nazioni Unite si consulti invece: United Nations, *Mandates of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression; The Special Rapporteur on the Right to Privacy and the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism*, OL OTH 71/2018, 7-12-2018, https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=24234.

[15] Cons. UE, *Proposta di regolamento del Parlamento europeo e del Consiglio relativo alla prevenzione della diffusione di contenuti terroristici online - Orientamento generale*, 15336/18, 7-12-2018.

[16] P.E., *Risoluzione legislativa del Parlamento europeo del 17 aprile 2019 sulla proposta di regolamento del Parlamento europeo e del Consiglio relativo alla prevenzione della diffusione di contenuti terroristici online* (COM(2018)0640 - C8-0405/2018 - 2018/0331(COD)), P8_TA(2019)0421, 17-4-2019.

[17] Reg. UE n. 784/2021, art. 3.3: «I prestatori di servizi di *hosting* rimuovono i contenuti terroristici o disabilitano l'accesso ai contenuti terroristici in tutti gli Stati membri il prima possibile e in ogni caso entro un'ora dal ricevimento dell'ordine di rimozione».

particolarmente a rischio di ospitare questi contenuti, sono anche tenuti ad adottare in maniera autonoma delle misure proattive[18]. Per garantire l'ottemperanza di questi doveri, il Regolamento stabilisce che dal mancato rispetto di queste disposizioni derivi la comminazione di importanti sanzioni economiche, che possono ammontare fino al 4% del fatturato globale del precedente esercizio finanziario dell'*hosting service provider* interessato[19].

Come anticipato in apertura, il Regolamento in oggetto presenta almeno quattro criticità importanti[20].

La prima riguarda il concetto di "contenuto terroristico" su cui esso si fonda. In effetti, l'art. 2.7 ne fornisce una definizione molto ampia, che riprende sostanzialmente quella contenuta nella Direttiva (UE) 541/2017 sulla lotta al terrorismo[21]. In dettaglio, il Regolamento identifica cinque categorie di materiali che possono essere qualificati come "contenuti terroristici": i primi sono quelli che incitano, direttamente o indirettamente, alla commissione di atti terroristici; i secondi sono quelli che sollecitano la realizzazione di tali atti; i terzi sono quelli che invitano a partecipare ad attività di gruppi terroristici; i quarti sono quelli che forniscono istruzioni sull'uso di esplosivi, armi e sostanze pericolose; i quinti ed ultimi sono invece quelli che rappresentano una minaccia di realizzazione di reati di stampo terroristico. Si consideri che all'art. 1.3 viene stabilito che non possono essere considerati terroristici – e quindi rimossi – i materiali diffusi a scopi educativi, giornalistici, artistici, di ricerca, di prevenzione o contrasto al terrorismo, nonché le opinioni espresse nei dibattiti politici. Cionondimeno, la dottrina ha sottolineato che avere accolto una definizione così ampia di "contenuto terroristico" lascia ampio margine all'interpretazione soggettiva e, in ultima istanza, potrebbe anche portare a rimuovere materiali che, nella sostanza, non sono realmente tali[22]. La realizzabilità di questo rischio è tra l'altro consolidata dalla previsione di importanti sanzioni economiche: invero, per la paura di essere sanzionati, gli *hosting service providers*

[18] Reg. UE n. 784/20211, art. 5.2: «Un prestatore di servizi di *hosting* esposto a contenuti terroristici di cui al paragrafo 4, adotta misure specifiche per proteggere i propri servizi dalla diffusione al pubblico di contenuti terroristici [...]».

[19] Reg. UE n. 784/2021, art. 18.3: «Gli Stati membri provvedono a che la sistematica o persistente inosservanza degli obblighi ai sensi dell'articolo 3, paragrafo 3, sia passibile di sanzioni pecuniarie fino al 4 % del fatturato mondiale del prestatore di servizi di *hosting* del precedente esercizio finanziario».

[20] A. Vedaschi, *La lucha contra la difusión de contenidos terroristas en línea: entre poder público y acción de los privados*, in M.J. Ridaura Martínez (cur.), *Retos para la seguridad*, Valencia, 2023, 223, 231.

[21] Dir. UE n. 541/2017 del P.E. e del Cons. del 15-3-2017 sulla lotta contro il terrorismo e che sostituisce la decisione quadro 2002/475/GAI del Cons. e che modifica la decisione 2005/671/GAI del Cons. Tra l'altro, la trasposizione della Direttiva in questione senza alcun adattamento è stata oggetto di severe critiche in dottrina, poiché essa assume come riferimento la realtà materiale, non quella digitale: T. Gherbaoui, M. Scheinin, *A Dual Challenge to Human Rights Law: Online Terrorist Content and Governmental Orders to Remove It*, in 1 *Eur. J. Hum. Rts.* 1, 13 (2023).

[22] M. Rojszczak, *Gone in 60 Minutes: Distribution of Terrorist Content and Free Speech in the European Union*, in 20(2) *Democracy & Sec.* 179, 192 (2024).

DPCE online
ISSN: 2037-6677

potrebbero non fare gli approfondimenti necessari e prediligere una rimozione *tout court*[23].

Un secondo aspetto problematico attiene alla mancata indicazione nel Regolamento della natura dell'autorità competente ad ordinare la rimozione del supposto contenuto terroristico. Invero, in sua assenza, gli Stati membri hanno assegnato questo ruolo ad autorità profondamente diverse tra di loro: alcuni hanno affidato tale compito alle forze di polizia (si pensi, ad esempio, alla Germania, all'Irlanda o alla Svezia), altri al Ministero dell'Interno (come in Spagna), altri ancora alla procura competente (è il caso dell'Italia) e taluni all'autorità giurisdizionale (tra gli altri, la Danimarca). Il fatto che gli Stati membri possano assegnare questo ruolo all'autorità che preferiscono potrebbe, secondo quanto sottolineato dalla dottrina, incidere profondamente sulla separazione dei poteri poiché, in alcuni casi, verrebbe a mancare un controllo giurisdizionale circa l'ammissibilità dell'ordine di rimozione[24]. In aggiunta, tale disomogeneità potrebbe anche compromettere l'applicazione uniforme del Regolamento[25].

Una terza criticità è stata sollevata in riferimento alla portata transnazionale del Regolamento[26]. Invero, constatando che i contenuti terroristici vengono spesso disseminati da *hosting service providers* aventi la propria sede legale all'infuori dell'Unione europea, l'art. 4 del Regolamento autorizza le autorità nazionali competenti a richiederne la rimozione se accessibili (e limitatamente al) nel territorio degli Stati membri. In altre parole, anche i fornitori di servizi di *hosting* con sede legale in Paesi terzi possono essere obbligati a rimuovere o a bloccare l'accesso a contenuti terroristici, e ciò esclusivamente nel territorio degli Stati membri. A tal fine, l'art. 17 stabilisce che i fornitori di servizi di *hosting* senza una sede principale nell'Unione europea siano tenuti a designare un rappresentante legale che si occupi di ricevere e dare seguito agli ordini di rimozione emessi dalle autorità competenti degli Stati membri. Sebbene questa disposizione rinforzi il Regolamento, essa presenta almeno due problematiche principali. In primo luogo, da una prospettiva interna, è stato osservato che questa norma potrebbe essere soggetta ad abusi da parte di alcuni Stati membri, nello specifico quelli con tendenze autoritarie[27]. In pratica, approfittandosi della vaga definizione di "contenuto terroristico" accolta nel Regolamento, essi potrebbero facilmente richiedere la rimozione di diversi materiali e, in questo modo, reprimere il dissenso della politica governativa. In secondo luogo, guardando a una prospettiva esterna, l'applicazione transfrontaliera[28] del Regolamento potrebbe entrare in conflitto con le normative di Stati terzi

[23] W. Bellaert, V. Selimi, R. Gouwy, *The End of Terrorist Content Online?*, in 92(2) *Rev. int. droit pénal* 163, 178 (2021).
[24] M. Rojszczak, *Gone in 60 Minutes*, *op. cit.*, 194-196.
[25] T. Gherbaoui, M. Scheinin, *op. cit.*, 13.
[26] M. Ferrario, *La portata transnazionale del regolamento (UE) 2021/784 e i possibili profili di incompatibilità con le normative di Stati terzi: un'analisi comparata*, in E.A. Imparato, G. Giorgini Pignatiello (cur.), *La libertà di espressione nel diritto comparato tra stato di diritto e stati di emergenza*, Torino, 2024, 339, 344 ss.
[27] J. Burchett, *Countering Extremist Ideologies: What are the Synergies Between the EU's Internal and External Action?*, in 14(2) *New J. Eur. Crim. L.* 138, 155 (2023).
[28] Si parla, in questo senso, anche di "effetto Bruxelles": A. Bradford, *The Brussels Effect: How the European Union Rules the World*, New York, 2020.

che, pur non essendo membri dell'Unione europea, potrebbero nondimeno trovarsi obbligati a rispettarne ed applicarne il contenuto[29].

3. La regola del *one-hour takedown* e i relativi rischi

Sin da quando ne erano state presentate le prime bozze, uno degli aspetti più controversi del Regolamento era rappresentato dalla regola che prevedeva la rimozione entro un'ora (massimo) del contenuto qualificato come terroristico[30]. Concretamente, dal momento in cui ricevono l'ordine di rimozione da parte dell'autorità nazionale competente, gli *hosting service providers* dispongono di massimo sessanta minuti per ottemperarvi e quindi rimuovere o non rendere più accessibile il contenuto.

La crescita esponenziale dei materiali condivisi online, da un lato, e le stringenti tempistiche imposte, dall'altro, hanno reso praticamente impossibile affidare all'operato umano tale attività[31]. Così, per ottemperare alle misure imposte dal Regolamento[32], gli *hosting service providers* hanno iniziato a servirsi di strumenti automatizzati[33] e, in particolare, delle tecnologie di *hashing* e di algoritmi basati sul *machine learning* e sul *natural language processing*[34].

Con il termine *hashing* si fa riferimento a quel processo di *matching* possibile grazie alla creazione di un *hash*. Praticamente, nel momento in cui viene identificato un contenuto potenzialmente terroristico (come, per esempio, un video o un'immagine) esso viene codificato tramite un *hash* (una sorta di impronta digitale), che viene poi conservato in un database[35]. In

[29] A. Vedaschi, *op. cit.*, 236.
[30] M. Rojszczak, *Online Content Filtering in EU Law – A Coherent Framework or Jigsaw Puzzle?*, in 47 *Comput. L. & Sec. Rev.* 1, 13 (2022).
[31] S. Macdonald, S. Giro Correia, A.L. Watkin, *Regulating Terrorist Content on Social Media: Automation and the Rule of Law*, in 15 *Int'l J.L. Context* 183, 184 (2019). Tradizionalmente, la moderazione dei contenuti online veniva affidata all'operato umano; poi, in ragione del numero sempre più crescente di contenuti condivisi online e dei relativi costi per monitorarli, sono stati sviluppati degli strumenti automatizzati per supportare questa attività: M.B. Martínez, *Platform Regulation, Content Moderation, and AI-Based Filtering Tools: Some Reflections from the European Union*, 14 *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 211, 212 (2023).
[32] È bene tuttavia ricordare che già dai primi anni 2000, su incoraggiamento dell'Unione europea, gli *hosting service providers* avevano iniziato ad utilizzare strumenti automatizzati per procedere alla rimozione di contenuti terroristici online: S. Tosza, *Internet Service Providers as Law Enforcers and Adjudicators. A Public Role of Private Actors*, in 43 *Comput. L. & Sec. Rev.* 1, 3-4 (2021).
[33] A norma dell'art. 5.8 Regolamento, gli *hosting service providers* non hanno l'obbligo di adottare degli strumenti automatizzati. Tuttavia, è bene prendere atto che, al considerando 25, viene stabilito che essi dovrebbero poterlo fare se lo ritengono opportuno e necessario per contrastare la diffusione di contenuti terroristici online.
[34] S. Macdonald, A. Mattheis, D. Wells, *Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online, Policy Briefing: Tech Against Terrorism Europe*, Brussels, 2024, 15 ss. Sul punto si consulti anche M. Thomas-Evans, *A Critical Analysis into the Beneficial and Malicious Utilisations of Artificial Intelligence*, in R. Montasari (Ed), *Artificial Intelligence and National Security*, Cham, 2022, 81, 82-83.
[35] R. Bellanova, M. De Goede, *Co-Producing Security: Platform Content Moderation and European Security Integration*, in 60(5) *J. Common Mkt. Stud.* 1316, 1328 (2022).

DPCE online
ISSN: 2037-6677

questo modo, se dopo la sua creazione viene caricato online un contenuto analogo, esso verrà facilmente individuato tramite questo processo di corrispondenza ed eliminato prima di diventare accessibile a troppe persone.

Il *machine learning* è invece un campo dell'intelligenza artificiale che si concentra sullo sviluppo di algoritmi di apprendimento automatico in grado di identificare, raccogliere ed elaborare un elevato e diversificato numero di dati in tempi brevissimi[36]. Tra gli strumenti di *machine learning* più utilizzati per la rimozione di contenuti terroristici vi è, ad esempio, quello di *flag and remove*[37], ossia quella procedura in cui i contenuti sospetti di terrorismo vengono segnalati (*flag*) da un algoritmo e successivamente rimossi (*remove*) dalla piattaforma online. Un'altra procedura ampiamente impiegata è quella di *prediction and prevention*[38], in cui gli algoritmi di *machine learning* vengono addestrati per identificare quei contenuti che potrebbero essere associati a comportamenti estremisti e cancellarli ancora prima di essere pubblicamente accessibili. Gli algoritmi di *machine learning* vengono usati anche per scopi di *takedown and staydown*[39], ossia per garantire che un contenuto già rimosso in passato perché considerato pericoloso non possa poi essere ripubblicato.

Il *natural language processing* (NLP) rappresenta un'altra branca dell'intelligenza artificiale e consiste in un insieme di metodi computazionali che mirano a rendere il linguaggio umano accessibile ai computer e, nello specifico, danno loro la possibilità di comprenderlo e generarlo a loro volta[40].

Nonostante l'uso di strumenti automatizzati permetta di gestire grandi quantità di dati e quindi di individuare e rimuovere prontamente i contenuti terroristici, esistono delle criticità importanti correlate al loro impiego e, segnatamente, quelle di natura tecnica, giuridica ed etica.

3.1 I rischi tecnici

Secondo alcuni studiosi, l'uso di strumenti automatizzati nella moderazione dei contenuti online presenta dei limiti tecnici evidenti che possono inevitabilmente influire sull'accuratezza di questa attività[41].

Innanzitutto, è stato rilevato che il loro impiego ha nel tempo determinato, da un lato, l'eliminazione di contenuti non terroristici (c.d. falsi positivi) e, dall'altro, anche la mancata rimozione di contenuti che in realtà erano tali (c.d. falsi negativi)[42]. Questo rischio è tanto più elevato quanto la definizione del contenuto che deve essere individuato e quindi rimosso dai

[36] R. Montasari, *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution. Threats, Assessment and Responses*, Cham, 2024, 137.
[37] N. Elkin-Koren, *Contesting Algorithms: Restoring the Public Interest in Content Filtering by Artificial Intelligence*, in *Big Data & Soc'y* 1, 3 (2020).
[38] *Ibidem.*
[39] M. Rojszczak, *Gone in 60 Minutes*, *op. cit.*, 184.
[40] J. Torregrosa, G. Bello-Orgaz, E. Martínez-Cámara, J. Del Ser, D. Camacho, *A Survey on Extremism Analysis Using Natural Language Processing: Definitions, Literature Review, Trends and Challenges*, in *J. Ambient Intel. & Humanized Comput.*, 1 (2022).
[41] E. Vargas Penagos, *ChatGPT, Can You Solve the Content Moderation Dilemma?*, in 32(1) *Int'l J.L. & Info. Tech.* 1, 2 (2024).
[42] K. Gunton, *The Use of Artificial Intelligence in Content Moderation in Countering Violent Extremism on Social Media Platforms*, in R. Montasari (Ed), *Artificial Intelligence*, *op. cit.*, 73.

sistemi automatizzati non è accurata[43], come è il caso del Regolamento in oggetto. Tra l'altro, tutto ciò è ulteriormente enfatizzato anche dal fatto che gli strumenti automatizzati non hanno la percezione del contesto in cui si trovano[44]. Il sarcasmo, la satira, i riferimenti culturali così come le opinioni personali possono essere facilmente fraintesi e portare quindi a censurare dei contenuti che, in realtà, non violano le linee guida.

In secondo luogo, è stato sottolineato che l'impiego settoriale di strumenti automatizzati possa portare ad eliminare un contenuto terroristico su una piattaforma ma non dalle altre, favorendo quindi la loro continua circolazione[45]. A tale proposito, è utile sottolineare che, nel 2017, le cosiddette *Big Four* (ossia Facebook, Google, Microsoft e Twitter) si sono impegnate in un progetto di contrasto congiunto al terrorismo attraverso la creazione del Global Internet Forum to Counter Terrorism[46]. Invero, uno dei suoi strumenti chiave è stata la creazione di un database condiviso di contenuti illegali che, grazie alla tecnica di *hashing*, permette di evitare che un contenuto terroristico già eliminato da una delle loro piattaforme possa essere ricaricato su un'altra.

In terzo luogo, è stato evidenziato che l'evoluzione delle tecniche impiegate dai terroristi può portare alla creazione di contenuti sempre più sofisticati in grado di aggirare le pratiche di moderazione. Basti pensare, ad esempio, a quanto possa essere facile eludere l'*hash* di un'immagine semplicemente modificandone il colore[47]: poiché queste tecnologie funzionano secondo un metodo di combinazione, anche soltanto una piccola modifica nella gradazione della colorazione potrebbe portare ad una mancata individuazione della stessa.

3.2 I rischi giuridici

Alla base del dibattito relativo al Regolamento in oggetto vi è sempre stata la dubbia capacità di riuscire a bilanciare la necessità di lottare contro il terrorismo online con quella di salvaguardare taluni diritti fondamentali e, segnatamente, la libertà di espressione e di informazione, le garanzie

[43] N. Duarte, E. Llanso, A.C. Loup, *Mixed Messages? The Limits of Automated Social Media Content Analysis*, in 81(1) *Proc. Mach. Learning Rsch.* 100, 106 ss. (2018).

[44] G. Gosztonyi, D. Gyetván, A. Kovács, *Theory and Practice of Social Media's Content Moderation by Artificial Intelligence in Light of European Union's AI Act and Digital Services Act*, in 4(1) *Eur. J. L. & Pol. Sci.* 33, 39 (2025).

[45] K. Gunton, *op. cit.*, 72.

[46] B. Clifford, *Moderating Extremism: The State of Online Terrorist Content Removal Policy in the United States*, *Program on Extremism*, The George Washington University, 2021, 11. Per un'analisi approfondita si veda anche B. Heller, *Combating Terrorist-Related Content Through AI and Information Sharing*, Working Paper: Transatlantic Working Group, 2019, 1 - 7.

[47] E. Llansó, J. van Hoboken, P. Leerssen, J. Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, Working Paper: Transatlantic Working Group, 2020, 6. Per evitare che i contenuti multimediali vengano aggirati troppo facilmente è stato di recente favorito l'approccio del *perceptual hashing*, che consente di individuare anche i contenuti simili (e non solo identici). A differenza di quello tradizionale, quindi, questo è meno sensibile a cambiamenti minori come, ad esempio, le dimensioni o il colore di un'immagine: H. Farid, *An Overview of Perceptual Hashing*, in 1(1) *J. Online Tr. & Safety* 1 (2021).

DPCE online
ISSN: 2037-6677

giurisdizionali e il diritto alla non discriminazione tra piattaforme. Questa perplessità è stata ulteriormente esacerbata dall'impiego di strumenti automatizzati.

Sin da quando la Commissione aveva pubblicato la prima bozza del Regolamento, diversi attivisti per i gruppi digitali avevano sottolineato i rischi che potevano derivare dalla sua applicazione con riferimento alla libertà di espressione e informazione[48] tutelate dall'art. 11 della Carta dei diritti fondamentali dell'Unione europea[49]. Innanzitutto, fin dall'inizio, era stato messo in luce che la vaga definizione di "contenuto terroristico" ivi accolta avrebbe potuto condurre anche alla rimozione di contenuti che non erano effettivamente tali, e ciò a detrimento dei fondamentali diritti di potere esprimere un'opinione e di potersi aggiornare. Con riferimento a questo aspetto, la dottrina ha evidenziato l'ulteriore pericolo di strumentalizzazione da parte di alcuni Stati, che potrebbero avvalersi di questa definizione così ampia per rimuovere contenuti di oppositori politici o anche di matrice accademica, giornalistica o religiosa[50]. Tra l'altro, tutto ciò potrebbe essere ulteriormente esacerbato proprio dall'uso di strumenti automatizzati che, se alimentati con talune informazioni, potrebbero eliminare questi contenuti in serie (c.d. rischio di *over-blocking*)[51]. In secondo luogo, la libertà di espressione e informazione è minacciata anche dal fatto che la decisione di rimuovere un certo contenuto pertiene solo ed esclusivamente agli *hosting service providers*[52]. Oltre ad affidarsi a strumenti automatizzati, essi sono dei soggetti privati mossi da fini squisitamente economici, che non hanno quindi interesse a valutare la reale compatibilità delle loro scelte con i diritti fondamentali coinvolti[53]. Come sottolineato dai relatori speciali delle Nazioni Unite (ONU) sulla libertà di espressione, sul diritto alla privacy e sulla lotta al terrorismo, nel contesto della lotta al terrorismo online gli *hosting service providers* hanno delle «quasi-regulative, quasi-enforcement, and quasi-adjudicative functions»[54], ciò che può profondamente influire sull'integrità della libertà di espressione e informazione. In terzo luogo, anche la portata transnazionale del Regolamento può compromettere questi diritti. Basti pensare al caso in cui un ordine di rimozione viene indirizzato verso uno Stato che accoglie uno standard di protezione della libertà di

[48] *Access now* et al., cit.
[49] Carta dei diritti fondamentali dell'Unione europea (2000/C 364/01) del 18-12-2000, art. 11: «1. Ogni individuo ha diritto alla libertà di espressione. Tale diritto include la libertà di opinione e la libertà di ricevere o di comunicare informazioni o idee senza che vi possa essere ingerenza da parte delle autorità pubbliche e senza limiti di frontiera. 2. La libertà dei media e il loro pluralismo sono rispettati».
[50] T. Gherbaoui, M. Scheinin, *op. cit.*, 3.
[51] R. Gorwa, R. Binns, C. Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, in 7(1) *Big Data & Soc'y* 1, 5 (2020).
[52] R. Ahmed, *Negotiating Fundamental Rights: Civil Society and the EU Regulation on Addressing the Dissemination of Terrorist Content Online*, in *Stud. Conflict & Terrorism* 1, 13 ss. (2023).
[53] T. Gherbaoui, M. Scheinin, *op. cit.*, 22.
[54] United Nations, cit., 7.

espressione più esteso[55]: in un caso come questo, il suddetto ordine potrebbe avere un effetto negativo sui diritti fondamentali in questione.

Il Regolamento comporta poi anche dei rischi con riferimento alle garanzie giurisdizionali. Invero, come visto sopra, l'ordine di rimozione è il frutto della decisione esclusiva dell'autorità nazionale competente (spesso incardinata nel potere esecutivo) che, quindi, non necessita di un'autorizzazione da parte di un giudice. Pertanto, qualora una persona dovesse essere vittima di una rimozione ingiustificata disporrebbe soltanto della possibilità di ricorrere contro questa decisione *ex post*. Tuttavia, è risaputo che in diversi Stati dell'Unione europea[56] la durata media di un processo può essere di anni, con il risultato che anche una decisione a favore del ricorrente perderebbe inevitabilmente di significato nella situazione in cui, per il decorrere del tempo, la notizia in questione diventerebbe irrilevante[57]. Tra l'altro, negli Stati con derive autoritarie, il ricorso di un individuo contro un ordine di rimozione potrebbe essere completamente compromesso se l'autorità giudiziaria in questione non gode dei requisiti di autonomia e indipendenza richiesti per svolgere opportunamente questa funzione[58].

Infine, proprio per il fatto di doversi necessariamente affidare a strumenti automatizzati, è emerso anche il rischio che il Regolamento possa ingenerare una disparità di trattamento tra gli *hosting service providers* e incidere quindi negativamente sulla libera concorrenza nel mercato[59]. Invero, il fatto di assoggettare egualmente alla regola di un'ora tanto le piattaforme più grandi quanto quelle più piccole, non considera che le due hanno una forza economica differente e, quindi, non tiene presente che le ultime non possono avvalersi degli stessi strumenti delle prime. Oltre a mettere gli *hosting service providers* più piccoli in una posizione precaria, questa regola potrebbe addirittura determinarne l'uscita dal mercato in assenza di una soluzione alternativa[60].

3.3 I rischi etici

[55] Si pensi, per esempio, al caso degli Stati Uniti ove esistono almeno due ostacoli significativi alla regolamentazione dei contenuti online, ossia il primo emendamento e la sez. 230 del *Communications Decency Act 1996*, Pub. L. 104-104: M. Ferrario, *op. cit.*, 347 ss. Sul punto si veda anche G. De Gregorio, *Il diritto delle piattaforme digitali: un'analisi comparata dell'approccio statunitense ed europeo al governo della libertà di espressione*, in *DPCE Online*, 2021, n. spec., 1455.

[56] Per una disamina puntuale della durata media dei processi civili e penali si consulti l'ultimo rapporto della Commissione europea per l'efficacia della giustizia pubblicato nel 2024: European Commission for the Efficiency of Justice (CEPEJ), *European Judicial Systems CEPEJ Evaluation Report. 2024 Evaluation Cycle (2022 Data)*, Strasbourg, 2024, https://www.coe.int/en/web/cepej/special-file.

[57] M. Rojszczak, *Gone in 60 Minutes*, *op. cit.*, 197.

[58] Ivi, 198.

[59] A. Vedaschi, *op. cit.*, 234.

[60] G. Kalpakis et al., *AI-Based Framework for Supporting Micro and Small Hosting Service Providers on the Report and Removal of Online Terrorist Content*, in I. Gkotsis et al. (Eds), *Paradigms on Technology Development for Security Pratictioners*, Cham, 2025, 249, 255 ss.

DPCE online
ISSN: 2037-6677

L'impiego di strumenti automatizzati può, infine, anche ingenerare dei problemi di natura etica relazionabili, segnatamente, all'assenza di trasparenza e all'impossibilità di capire a chi imputare la responsabilità di un certo fatto[61].

Invero, è largamente risaputo che la moderazione automatizzata dei contenuti è, in generale, un processo ampiamente opaco. Questo comporta che i criteri attraverso i quali un contenuto (nel nostro caso, terroristico) viene individuato e poi rimosso restino in gran parte sconosciuti. Oltre a non consentire agli utenti di potere predeterminare quali sono i contenuti proibiti, questo potrebbe addirittura portarli ad autocensurarsi per evitare di essere bannati[62].

Per quanto riguarda invece l'*accountability*, l'uso di strumenti automatizzati rende praticamente impossibile risalire al responsabile della rimozione in quanto non vi è chiarezza circa chi abbia fatto una determinata scelta[63]. Detto diversamente, non essendo chiaro se la decisione di eliminare un certo contenuto sia imputabile al sistema automatizzato oppure all'umano che non ha sorvegliato, è di fatto impossibile ascrivere a qualcuno la responsabilità di una certa decisione.

4. Riflessioni conclusive

Il Regolamento (UE) 784/2021 sulla rimozione dei contenuti terroristici online rappresenta un'iniziativa importante nella lotta al terrorismo digitale. Esso istituisce infatti un sistema di notifica e rimozione in base al quale gli *hosting service providers* (europei e non) sono obbligati a conformarsi, entro un massimo di 60 minuti, all'ordine di rimozione di un contenuto terroristico emesso dall'autorità nazionale competente.

Pur rispondendo all'esigenza urgente di contrastare la diffusione online di materiale incitante al terrorismo, il Regolamento (UE) 784/2021 presenta una serie di criticità di natura giuridica e pratica che meritano una riflessione approfondita, soprattutto con riferimento alla moderazione automatizzata dei contenuti.

In primo luogo, come visto, il Regolamento accoglie una definizione molto vaga di "contenuto terroristico", che potrebbe portare ad eliminare dei contenuti che, pur potenzialmente controversi, non sono tali. In secondo luogo, in mancanza di un'indicazione circa la natura dell'autorità nazionale competente ad emettere l'ordine di rimozione, gli Stati membri sono stati di fatto autorizzati ad affidare questo compito a qualsivoglia autorità, con possibili ripercussioni sul principio della separazione dei poteri e sull'applicazione uniforme del Regolamento. In terzo luogo, risulta potenzialmente problematico anche il fatto che il Regolamento abbia una portata transnazionale, in quanto un ordine di rimozione può essere rivolto

[61] R. Gorwa, R. Binns, C. Katzenbach, *op. cit.*, 10 ss.
[62] Si parla in questo senso anche di "*chilling effect*": A. Zornetta, I. Pohland, *Legal and Technical Trade-offs in the Content Moderation of Live Streaming*, in 30(3) *Int'l J. L. & Info. Tech.* 302, 313 (2022). Sul rischio di auto-censura si veda anche S. Schinello, *New (Digital) Media in Creative Society: Ethical Issues of Content Moderation*, in 35(1) *Phil. Socio.* 67, 70 (2024).
[63] R. Montasari, *Cyberspace, Cyberterrorism and the International Security*, *op. cit.*, 162 ss.

anche ad un'autorità di uno Stato non-UE avente una disciplina diversa e potenzialmente contrastante in materia. Il quarto e più critico aspetto attiene alle tempistiche previste e alla conseguente necessità di affidarsi a strumenti automatizzati. Invero, visto che i contenuti terroristici devono essere eliminati o resi indisponibili entro al massimo 60 minuti dalla ricezione dell'ordine di rimozione, l'uso di strumenti automatizzati (come gli *hash*, il *machine learning* o il *natural language processing*) si è reso praticamente necessario. Questo, come visto, implica che l'individuazione, la qualificazione e finanche la decisione di rimuovere un contenuto terroristico venga in ultima istanza ed unicamente presa da un algoritmo, con i rischi tecnici, giuridici ed etici che ne possono conseguire. Infatti, la scelta di affidare esclusivamente ad uno strumento automatizzato la decisione di rimuovere un certo contenuto può implicare, a causa di un falso positivo, l'eliminazione di un contenuto legittimo. Questo avrebbe necessariamente delle ripercussioni tanto sulla libertà di espressione, quanto su quella di informazione. In aggiunta a queste libertà, anche la garanzia della via giudiziaria potrebbe essere compromessa, soprattutto se si considera che, a causa dell'opacità degli strumenti impiegati, è quasi impossibile individuare il vero responsabile di una certa decisione. Infine, è bene considerare che, avendo il Regolamento imposto le stesse regole a tutti gli *hosting service providers*, ha di fatto sfavorito quelli piccoli e medi per cui è più difficile avvalersi di strumenti automatizzati efficienti.

Per confinare il rischio di censurare dei contenuti legittimi e compromettere la libertà di espressione e informazione sarebbe innanzitutto necessario che le piattaforme venissero obbligate ad implementare degli algoritmi in grado di ridurre al minimo il rischio di errore. Nei casi più complessi, sarebbe opportuno prevedere anche una procedura di revisione manuale[64], così da garantire una corretta valutazione contestuale.

Con riferimento invece al diritto di difesa degli utenti, sarebbe necessario prevedere delle procedure di ricorso più snelle ed efficaci, oltre che un obbligo di motivate giustificazioni in caso di rimozione. In questo contesto, l'integrazione di meccanismi di trasparenza relativi agli algoritmi di moderazione diventa una *conditio sine qua non*.

Infine, per quanto attiene alla protezione dei piccoli e medi *hosting service providers* sarebbe opportuno prevedere delle misure differenziate in modo, da un lato, di evitare la loro uscita dal mercato e, dall'altro, anche di ridurre il rischio di eliminazione di contenuti legittimi. Infatti, non potendo sempre disporre di strumenti automatizzati efficienti, essi potrebbero essere più inclini a rimuovere un certo contenuto per la paura di essere sanzionati.

In conclusione, per evitare che la lotta al terrorismo online si trasformi in un pretesto per limitare i diritti e le libertà fondamentali, è indispensabile che il Regolamento (UE) 2021/784 venga accompagnato da misure di controllo trasparenti, da un sistema di ricorsi effettivi e da un attento monitoraggio delle sue applicazioni. Solo in questo modo si può effettivamente garantire un giusto equilibrio tra la protezione della sicurezza pubblica, da un lato, e il rispetto dei diritti umani, dall'altro.

---

[64] Come è stato, per esempio, il caso di Facebook: K. Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, in 129(8) *Yale L.J.* 2418 (2020).

DPCE online
ISSN: 2037-6677

Micol Ferrario
Dip.to di Giurisprudenza
Università di Torino
micol.ferrario@unibocconi.it