SISTEMA PENALE

FASCICOLO

4/2026

ISSN 2704-8098

Sistema penale (*SP*) è una rivista *online*, aggiornata quotidianamente e fascicolata mensilmente, ad accesso libero, pubblicata dal 18 novembre 2019.
La *Rivista*, realizzata con la collaborazione scientifica dell'Università degli Studi di Milano e dell'Università Bocconi di Milano, è edita da Progetto giustizia penale, associazione senza fine di lucro con sede presso il Dipartimento di Scienze Giuridiche "C. Beccaria" dell'Università degli Studi di Milano, dove pure hanno sede la direzione e la redazione centrale.
Tutte le collaborazioni organizzative ed editoriali sono a titolo gratuito e agli autori non sono imposti costi di elaborazione e pubblicazione.
La *Rivista* si uniforma agli standard internazionali definiti dal *Committe on Publication Ethics* (COPE) e fa proprie le relative linee guida.

Peer review I contributi che la direzione ritiene di destinare alla sezione "*Articoli*" del fascicolo mensile sono inviati a un revisore, individuato secondo criteri di rotazione tra i membri del Comitato scientifico, composto da esperti esterni alla direzione e al comitato editoriale. La scelta del revisore è effettuata garantendo l'assenza di conflitti di interesse. I contributi sono inviati ai revisori in forma anonima. La direzione, tramite la redazione, comunica all'autore l'esito della valutazione, garantendo l'anonimato dei revisori. Se la valutazione è positiva, il contributo è pubblicato. Se il revisore raccomanda modifiche, il contributo è pubblicato previa revisione dell'autore, in base ai commenti ricevuti, e verifica del loro accoglimento da parte della direzione. Il contributo non è pubblicato se il revisore esprime parere negativo alla pubblicazione. La direzione si riserva la facoltà di pubblicare nella sezione "*Altri contributi*" una selezione di contributi diversi dagli articoli, non previamente sottoposti alla procedura di *peer review*. Di ciò è data notizia nella prima pagina della relativa sezione.
Di tutte le operazioni compiute nella procedura di *peer review* è conservata idonea documentazione presso la redazione.

Modalità di citazione Per la citazione dei contributi presenti nei fascicoli di *Sistema penale*, si consiglia di utilizzare la forma di seguito esemplificata: N. COGNOME, *Titolo del contributo*, in *Sist. pen.* (o *SP*), 1/2023, p. 5 ss.

SISTEMA PENALE

4/2026

SISTEMI DI MONITORAGGIO ALGORITMICO DELLE COMUNICAZIONI E TUTELA DEI DIRITTI FONDAMENTALI: *"CYBER SOCIAL SECURITY" RESEARCH PROJECT* (*)

di Alessandro Valenti, Maria Cecilia Cardarelli, Guido Colaiacovo, Valentina Vincenza Cuocci, Giorgio Fontana, Wanda Nocerino, Angela Procaccino, Francesco Giacomo Viterbo, Donatella Curtotti

Il contributo presenta alcuni tra i più rilevanti risultati del progetto di ricerca "Cyber Social Security", finalizzato allo sviluppo di sistemi di intelligenza artificiale per la prevenzione del cyberbullismo che siano al contempo efficaci e conformi al quadro normativo. L'indagine si focalizza sulla compatibilità tra l'addestramento e l'impiego di sistemi di monitoraggio algoritmico delle comunicazioni e la tutela dei diritti fondamentali, e inoltre si interroga sull'applicabilità dell'EU AI Act e sui vincoli imposti dal GDPR, tenendo conto anche della bozza di regolamento Digital Omnibus. Dopo aver esaminato tali profili, vengono delineate le possibili prospettive di innesto dei sistemi, in un contesto che vedrà crescere sempre più l'impiego dell'intelligenza artificiale negli scenari investigativi, così come la funzione preventiva delle piattaforme digitali, ma che resta ancora privo di solidi riferimenti normativi.

1. Il progetto "*Cyber Social Security*" per la lotta al cyberbullismo.

La diffusione degli ambienti comunicativi *online* ha reso le interazioni digitali una componente centrale della vita quotidiana, soprattutto per le giovani generazioni. Accanto ai benefici derivanti dalle nuove modalità di comunicazione, sono emersi rischi significativi. La comunicazione tecnologicamente mediata, affrancata dai tradizionali

(*) Il contributo si inserisce nell'ambito del Progetto di ricerca "*Cyber Social Security*" (CUP D73C23002120005), partenariato esteso SERICS (PE00000014 - Cup n. F53C22000740007), bando a cascata Spoke 3 " Attacks and Defences", ammesso a finanziamento PNRR, Missione 4 – Componente 2 - Investimento 1.3, Next GenerationEU.

limiti spazio-temporali e contraddistinta da distanza relazionale, nonché dalla possibile condizione di anonimato, favorisce l'emergere di condotte aggressive e devianti, e amplifica la loro offensività. Tali comportamenti, infatti, possono diffondersi con estrema rapidità e raggiungere un pubblico vasto e indeterminato. In questo scenario, desta particolare allarme il cyberbullismo, un fenomeno che è in costante crescita in tutto il mondo[1].

Alla luce delle evidenze emerse, quattro università pugliesi (Università del Salento - capofila, Università degli Studi di Foggia, Politecnico di Bari, Università LUM "Giuseppe Degennaro") hanno lavorato sinergicamente dal 1° luglio 2024 al 31 dicembre 2025 per elaborare strumenti efficaci di prevenzione del fenomeno, nell'ambito del più ampio progetto "*Cyber Social Security*" (C.S.S.) coordinato dall'Università degli studi di Bari. L'intento della ricerca è stato quello di progettare sistemi di intelligenza artificiale in grado di identificare e segnalare tempestivamente comportamenti-indicatori di cyberbullismo realizzati sui *social network* e sui canali pubblici di messaggistica istantanea, assicurando, al contempo, anche in fase di addestramento degli algoritmi, il pieno rispetto dei diritti fondamentali, con particolare riguardo ai diritti delle persone minorenni.

In tal modo, il progetto ha inteso perseguire con rigore l'obiettivo di soddisfare l'esigenza di tutela dei diritti dei minori nello spazio cibernetico, sempre più avvertita sia in Italia[2] sia nell'Unione europea. In particolare, la ricerca si colloca potenzialmente nel solco tracciato dall'art. 28 del Regolamento (UE) 2022/2065 ("*Digital Services Act*"), a mente del quale «i fornitori di piattaforme *online* accessibili ai minori adottano misure adeguate e proporzionate per garantire un elevato livello di tutela della vita privata, di sicurezza e di protezione dei minori sul loro servizio»[3].

Questo obiettivo è stato perseguito mediante un approccio multidisciplinare. La componente ingegneristica e informatica del gruppo di ricerca è stata chiamata a sviluppare i sistemi di intelligenza artificiale, e quindi a progettare gli algoritmi e ad addestrarli, dopo aver raccolto i dati necessari da *social network*, sistemi di messaggistica e telecamere ambientali. La componente psicologica ha chiarito le dinamiche relazionali

---

[1] OMS, *A focus on adolescent peer violence and bullying in Europe, central Asia and Canada Health Behaviour in School-aged Children international report from the 2021/2022 survey*, Volume 2, www.who.int, 27 marzo 2024.

[2] In Italia, tra le iniziative dirette a questo scopo è paradigmatico il disegno di legge AS n. 1136 del 13 maggio 2024 ("Disposizioni per la tutela dei minori nella dimensione digitale"), che ha come diretti destinatari «i fornitori di servizi della società dell'informazione». Se il d.d.l. venisse approvato, questi ultimi assumerebbero il dovere di verificare che gli utenti abbiano almeno quindici anni. Inoltre, i contratti conclusi con soggetti di età inferiore sarebbero «nulli» e «non po[trebbero] rappresentare idonea base giuridica per il trattamento dei dati personali», ai sensi dell'articolo 6, par. 1, lett. *b*), GDPR, salvo che non siano stati stipulati per conto del minore da chi esercita la responsabilità genitoriale o dal tutore. Sul rispetto di tali disposizioni dovrebbero vigilare l'Autorità per le garanzie nelle comu-nicazioni e il Garante per la protezione dei dati personali, i quali, in caso di violazioni, dovrebbero irrogare le sanzioni di cui al d.lgs. 31 luglio 2005, n. 177 e al GDPR (artt. 2 e 3).

[3] Sul regolamento, S. Braschi, *Il nuovo Regolamento sui servizi digitali: quale futuro per la responsabilità degli Internet Service Provider?*, in *Dir. pen. proc.*, 2023, p. 367 ss.; O. Murro, *Il ruolo delle piattaforme online nella prevenzione della violenza di genere: cosa cambia con il Digital Service Act*, in *Pen. dir. proc.*, 19 dicembre 2025.

del fenomeno. La componente giuridica ha assunto il ruolo di garante della conformità del progetto al quadro normativo.

Un primo macro-obiettivo affidato ai giuristi ha riguardato la predisposizione di *guidelines*, dirette alla componente tecnica del progetto, volte ad assicurare la costante rispondenza della ricerca, oltre che dei suoi prodotti tecnologici, ai principi costituzionali e sovranazionali in tema di libertà fondamentali, nonché al Regolamento (UE) 2024/1689 sull'intelligenza artificiale (c.d. "EU AI Act") e al Regolamento (UE) 2016/679 sulla protezione dei dati personali (c.d. "GDPR").

Corre l'obbligo di precisare che le linee guida non sono state sviluppate nell'ottica di garantire la piena conformità dei sistemi di intelligenza artificiale oggetto di studio al quadro normativo che interessa i loro possibili impieghi futuri: il perseguimento di questo obiettivo avrebbe rischiato di sacrificare eccessivamente la ricerca. Come si illustrerà in seguito, la normativa euro-unitaria prevede un regime di obblighi attenuato qualora il trattamento dei dati personali e lo sviluppo dei sistemi di intelligenza artificiale siano giustificati da finalità di ricerca scientifica. Tale scelta comporta che gli strumenti di monitoraggio automatizzato sviluppati nel corso del progetto, eventualmente, dovranno essere sottoposti ad un'attenta revisione prima di essere immessi sul mercato.

Un secondo macro-obiettivo è stato quello di identificare le condotte di cyberbullismo, analizzando la legge 29 maggio 2017 n. 71, volta alla prevenzione» e al «contrasto» del fenomeno[4], con particolare riguardo all'art. 2, che lo definisce, delineando una fattispecie alla cui integrazione non consegue l'applicabilità di una pena o di una circostanza aggravante[5], bensì l'operatività di una serie di istituti che hanno lo scopo di evitare che vengano realizzati nuovi episodi offensivi e, prima ancora, di neutralizzare gli ulteriori potenziali effetti dannosi di comportamenti già posti in essere.

Tale analisi è stata finalizzata ad escludere dal raggio di azione dei sistemi di intelligenza artificiale le condotte giuridicamente irrilevanti, e, conseguentemente, ad addestrare gli algoritmi sui c.d. "comportamenti spia", intesi come segnali di allarme di imminenti offese. In tal modo, si è perseguito l'obiettivo di garantire la progettazione di

---

[4] Inizialmente, si trattava di una legge esclusivamente sul cyberbullismo. La Legge 17 maggio 2024, n. 70 ha ampliato l'oggetto della disciplina al fenomeno del bullismo. L'incipit della nozione di cyberbullismo, prevista dall'art. 1, comma 2, legge n. 71/2017, è evocativo: «ai fini della presente legge, per "cyberbullismo" si intende». In effetti, la definizione normativa potrebbe non coincidere con quelle comuni, tratte dall'esperienza. Quelle nozioni, chiarisce il legislatore, non condizionano l'operatività degli istituti previsti dalla legge in esame. A tal fine, la nozione rilevante è questa: «per "cyberbullismo" si intende qualunque forma di pressione, aggressione, molestia, ricatto, ingiuria, denigrazione, diffamazione, furto d'identità, alterazione, acquisizione illecita, manipolazione, trattamento illecito di dati personali in danno di minorenni, realizzata per via telematica, nonché la diffusione di contenuti on line aventi ad oggetto anche uno o più componenti della famiglia del minore il cui scopo intenzionale e predominante sia quello di isolare un minore o un gruppo di minori ponendo in atto un serio abuso, un attacco dannoso, o la loro messa in ridicolo».

[5] Il cyberbullismo non integra un'autonoma fattispecie di reato, ma è pur vero che molte delle condotte descritte dall'art. 1, comma 2, legge n. 71/2017 hanno rilevanza penale e che altre possono acquisirla nell'ipotesi in cui ricorrano ulteriori elementi. Sul tema, M. C. Parmiggiani, *Il cyberbullismo*, in *Cybercrime*, a cura di A. Cadoppi - S. Canestrari - A. Manna - M. Papa, Utet, 2019, p. 673 e ss.

un sistema coerente con la normativa di settore, e quindi idoneo a perseguire finalità preventive. In questa prospettiva, all'interno dell'area di ricerca, è stata anche svolta un'analisi relativa agli strumenti apprestati dalla stessa legge, nell'ottica di verificare se quelli in progettazione siano funzionali all'adempimento di specifici obblighi gravanti sulle *Big Tech*.

Nel presente contributo si intende offrire una sintesi dei risultati dello studio orientato alla predisposizione delle *guidelines*, condotto in sinergia dall'Università degli studi di Foggia e dall'Università del Salento, includendo anche spunti riconducibili alla ricerca incentrata direttamente sulla legge n. 71/2017. Preliminarmente, appare necessario soffermarsi sulle ragioni che rendono imprescindibile lo sviluppo di sistemi di intelligenza artificiale nella lotta al cyberbullismo e poi, al fine di comprendere l'effettiva portata delle problematiche che sollevano, delinearne sinteticamente il funzionamento.

2. Limiti delle metodiche tradizionali e necessità di investire sui sistemi di monitoraggio algoritmico.

Al fine di comprendere la necessità di investire nello sviluppo di adeguati sistemi di monitoraggio algoritmico, occorre interrogarsi sulla sussistenza di metodiche alternative disponibili che siano in grado di realizzare la finalità del progetto *Cyber Social Security*. La finalità, si ripete, è quella di individuare tempestivamente, nello spazio cibernetico, segnali di allarme relativi al compimento di atti di cyberbullismo, così da consentire un intervento anticipato rispetto alla possibile evoluzione del comportamento lesivo.

Attualmente, non sono ipotizzabili strumenti alternativi all'intelligenza artificiale in grado di soddisfare tale esigenza. La ragione è piuttosto intuitiva. Qualsiasi forma di monitoraggio affidata esclusivamente all'operatore umano, anche quella che presenta il minore impatto sui diritti fondamentali, in quanto limitata all'osservazione dei comportamenti manifestati nello spazio pubblico digitale, si scontra frontalmente con l'estensione sterminata e incontrollabile delle interazioni sociali che si svolgono nell'universo cibernetico. In tale scenario, appare oggettivamente impossibile garantire l'effettività della funzione preventiva in assenza di sistemi automatizzati di rilevamento, tanto per gli attori pubblici quanto per quelli privati.

Sennonché, lo sviluppo di sistemi di intelligenza artificiale in grado di servire la finalità preventiva senza costituire seri pericoli per i diritti fondamentali è un processo ancora in atto e che non consiglia (imprudenti) accelerazioni da parte del legislatore. Ne consegue che, attualmente, nel quadro normativo che rileva in questa analisi, non si rinviene un obbligo di monitoraggio attivo e costante sulla natura dei contenuti pubblicati.

Allo stato attuale, il legislatore nazionale e quello unitario configurano l'azione preventiva del *cybercrime* come un adempimento esigibile nei limiti in cui il meccanismo è messo in moto dagli utenti dei servizi e, per quanto si dirà, sul presupposto di una condotta illecita.

Così, la l. 71/2017 riconosce alla vittima di cyberbullismo il diritto di segnalare condotte offensive e il conseguente obbligo per il gestore del *social* di attivarsi a sua tutela, mediante il potere di rimozione o oscuramento dei contenuti (art. 2, comma 1). Analogamente, il *Digital Services Act* obbliga i c.d. "prestatori di servizi della società dell'informazione", categoria nella quale rientrano i c.d. "*social network*" (Considerando n. 13) a predisporre meccanismi di segnalazione dei «contenuti illegali» e impone loro di provvedere tempestivamente (art. 16, par. 1 e 5), assumendo decisioni che possono consistere, ad esempio, nella rimozione dei contenuti o nella sospensione o chiusura del profilo utente (art. 17, par. 1).

Tuttavia, tale contegno diligente, da parte del soggetto interessato, e a maggior ragione dei terzi, non è imposto a livello generale da alcuna disciplina, per l'effetto che l'agire delle *Big Tech* contro il cyberbullismo rischia di assumere tratti ipotetici, casuali ed episodici. Peraltro, laddove a segnalare l'evento sia la vittima, la segnalazione non risulta il presupposto di un'azione di carattere effettivamente preventivo, essendosi l'offesa già in parte consumata.

Allora, per prevenire davvero il cyberbullismo, i gestori dei *social network* dovrebbero avvalersi del supporto di sistemi automatizzati per il monitoraggio sistematico dell'ambiente comunicativo digitale, finalizzato al rilevamento di comportamenti suscettibili di evolvere nell'evento lesivo.

Il regolamento euro-unitario, in realtà, tiene conto della possibilità di impiego dei sistemi di intelligenza artificiale per la c.d. "moderazione dei contenuti", tanto è vero che tale formula descrive l'insieme di «*attività, automatizzate o meno*, svolte dai prestatori di servizi intermediari con il fine, in particolare, di *individuare, identificare e contrastare contenuti illegali* e informazioni incompatibili con le condizioni generali, forniti dai destinatari del servizio» (art. 3, par. 1, lett. t).

Sennonché, il *Digital Services Act*, come si è anticipato, non introduce un obbligo di monitoraggio algoritmico dei contenuti offensivi, ma tenta di rendere il più possibile trasparente lo svolgimento di tale attività[6]. L'art. 8 non lascia adito a dubbi: «ai prestatori di servizi intermediari *non è imposto alcun obbligo generale di sorveglianza* sulle informazioni che tali prestatori trasmettono o memorizzano, né di accertare attivamente fatti o circostanze che indichino la presenza di attività illegali».

Per altro verso, i contenuti che costituiscono oggetto delle attività di segnalazione e di rimozione sono soltanto quelli «illegali»[7], non anche le condotte apparentemente

---

[6] Il *Digital Services Act* prevede che i prestatori di servizi intermediari debbano mettere a disposizione del pubblico, su base almeno annuale, uno o più *report* sulle attività di moderazione svolte nei quali devono risultare le informazioni su quelle «avviate *di propria iniziativa* dai prestatori, compres[o] l'utilizzo di strumenti automatizzati» (art. 15, par. 1, lett. c). Rileva in tal senso anche la previsione che impone di motivare le decisioni di moderazione (art. 17): «ove opportuno», va chiarito al soggetto interessato «se la decisione sia stata adottata in merito a contenuti *individuati o identificati per mezzo di strumenti automatizzati*».

[7] Il *Digital Services Act* definisce in questi termini il concetto di «contenuto illegale»: «qualsiasi informazione che, di per sé o in relazione a un'attività, tra cui la vendita di prodotti o la prestazione di servizi, non è conforme al diritto dell'Unione o di qualunque Stato membro conforme con il diritto dell'Unione, indipendentemente dalla natura o dall'oggetto specifico di tale diritto» (art. 3). Il *Considerando* n. 12 chiarisce che «tale concetto dovrebbe, in particolare, intendersi riferito alle informazioni, indipendentemente dalla

irrilevanti, ed eventualmente lecite, ma suscettibili di evolversi in comportamenti offensivi. Da questo punto di vista, i sistemi di intelligenza artificiale ai quali si riferisce il regolamento non sono sistemi di rilevamento dei c.d. "indicatori di rischio" o "comportamenti spia".

La vera sfida è questa: implementare sistemi di intelligenza artificiale capaci di *impedire* le offese; sistemi affidabili e rispettosi dei diritti fondamentali alla luce dei quali avviare una seria riflessione sull'imposizione di un obbligo di monitoraggio diretto agli intermediari della comunicazione telematica.

3. Natura e funzionamento dei sistemi di intelligenza artificiale in corso di sviluppo.

Tentano di rispondere a questa sfida i sistemi di intelligenza artificiale realizzati nell'ambito del progetto *Cyber Social Security*. Dal punto di vista funzionale, essi possono essere descritti come strumenti automatizzati potenzialmente in grado di analizzare grandi contenuti testuali pubblicamente accessibili sui *social network* e sui sistemi di messaggistica istantanea e di sottoporli ad un'analisi finalizzata ad individuare la ricorrenza di "*pattern*" comunicativi e comportamentali riconducibili al cyberbullismo.

Tali sistemi si basano su un procedimento articolato di trattamento dei contenuti testuali. In estrema sintesi, esso comprende le seguenti fasi:

I) raccolta dei contenuti in conformità al GDPR (tema, questo, che si approfondirà successivamente) e organizzazione degli stessi, la quale include procedure di tokenizzazione avanzata, normalizzazione lessicale, rimozione delle c.d. "*stop-word* contestuali", gestione di *slang*, abbreviazioni e forme ortografiche non *standard* tipiche della comunicazione *online*, nonché l'identificazione di elementi paralinguistici come *emoticon*, *emoji* e marcatori discorsivi, che, nelle dinamiche del fenomeno, assumono spesso un valore rilevante[8];

---

loro forma, che ai sensi del diritto applicabile sono di per sé illegali, quali l'illecito incitamento all'odio o i contenuti terroristici illegali e i contenuti discriminatori illegali, o che le norme applicabili rendono illegali in considerazione del fatto che riguardano attività illegali». Si fa riferimento, secondo un elenco non esaustivo, alle seguenti attività: «la condivisione di immagini che ritraggono abusi sessuali su minori, la condivisione non consensuale illegale di immagini private, il *cyberstalking* (pedinamento informatico), la vendita di prodotti non conformi o contraffatti, la vendita di prodotti o la prestazione di servizi in violazione della normativa sulla tutela dei consumatori, l'utilizzo non autorizzato di materiale protetto dal diritto d'autore, l'offerta illegale di servizi ricettivi o la vendita illegale di animali vivi».

[8] La fattispecie di cyberbullismo è spesso integrata da condotte plurime, estemporanee, di per sé inoffensive, ma che assumono rilevanza se lette alla luce del contesto di riferimento. Molto spesso l'offesa è determinata dalle c.d. "*reactions*", ossia la condivisione di stati d'animo o prese di posizione effettuate mediante le c.d. "*emojis*". Questo significa che un approccio appiattito sulla natura del contenuto pubblicato (un video, un'immagine o alcune parole) è insoddisfacente, essendo impensabile prescindere da elementi quali il contesto comunicativo, la frequenza, il tipo di reazioni e la loro provenienza. Per interessanti spunti in materia, E. Battelli, *Minori e nuove tecnologie*, in AA.VV., *Diritto privato delle persone minori di età*, a cura di E. Battelli, Torino, 2021, p. 125.

II) arricchimento semantico ed esame del linguaggio "naturale", ossia forme di analisi volte a cogliere il significato dei contenuti raccolti, avuto riguardo anche al contesto comunicativo[9];

III) applicazione di modelli di *machine learning* e *deep learning*[10] addestrati per l'identificazione di *pattern* comportamentali associabili a dinamiche aggressive, moleste, persecutorie e discriminatorie.

Sul punto, va sottolineato che il sistema è strutturato come un insieme di classificatori cooperanti, ciascuno specializzato nell'individuazione di specifici aspetti del fenomeno, quali l'aggressività verbale, l'asimmetria di potere comunicativo, la ripetitività temporale degli attacchi e la presenza di bersagli vulnerabili. Sono tutti elementi che, combinati, consentono di distinguere episodi isolati di conflittualità *online* da comportamenti di cyberbullismo[11].

Una volta terminato il procedimento, laddove il sistema abbia identificato un comportamento allarmante, genera un *alert* per segnalarne la presenza.

Alcuni modelli progettati prevedono anche la possibilità di oscurare quei contenuti. Si tratta, dunque, di una sorta di decisione automatizzata, ma dalla natura provvisoria.

È bene evidenziare, sul punto, che i sistemi relativi al progetto C.S.S. non sono stati progettati per sostituirsi all'uomo nell'assunzione di decisioni, bensì per costituire un valido supporto informativo destinato a orientare l'intervento umano. Per questo, essi sono stati integrati con meccanismi di *explainable AI*, finalizzati a rendere interpretabili le segnalazioni e a supportare gli operatori nella comprensione dei fattori che contribuiscono alla classificazione di un contenuto come potenzialmente riconducibile alle dinamiche del cyberbullismo.

Pertanto, operano meccanismi che consentono di individuare gli elementi che hanno avuto maggiormente peso nella segnalazione (es: il linguaggio impiegato), tecniche che consentono di visualizzare le parti del testo sul quale il sistema si è concentrato maggiormente durante l'analisi (c.d. "tecniche di *attention visualization* nei

---

[9] Sulle rappresentazioni "normalizzate" viene applicato un livello di *feature extraction* basato su tecniche di *Natural Language Processing* (NLP), che combina approcci tradizionali, quali n-grammi ponderati tramite TF-IDF, con rappresentazioni distribuzionali dense derivate da modelli di *word embedding* e *sentence embedding*, addestrati o adattati su *corpora* specifici del dominio sociale e linguistico di riferimento del progetto C.S.S.

[10] I modelli di apprendimento automatico impiegati includono algoritmi supervisionati come *Support Vector Machines*, *Random Forest* e *Gradient Boosting*, utilizzati come *baseline* per la classificazione binaria e multi-classe, e modelli di *deep learning* basati su architetture neurali ricorrenti e trasformative, in particolare reti LSTM, capaci di catturare dipendenze contestuali di lungo periodo e sfumature semantiche complesse tipiche del linguaggio offensivo e manipolativo.

[11] Dal punto di vista del flusso operativo, il sistema non si limita a una classificazione statica dei singoli messaggi, ma implementa un'analisi dinamica e temporale dei comportamenti, aggregando le predizioni su finestre temporali scorrevoli e su grafi di interazione tra utenti, al fine di rilevare *pattern* di accanimento, coalizione e isolamento sociale; elementi tipici delle dinamiche di cyberbullismo e difficilmente identificabili attraverso un'analisi puntuale dei contenuti. Questa componente di analisi comportamentale sfrutta modelli di *network analysis* e metriche di centralità, densità e reciprocità, integrandole con gli *output* dei classificatori linguistici per costruire indicatori compositi di rischio, utilizzabili in scenari di monitoraggio preventivo e di supporto alle politiche di intervento.

modelli neurali") e metodi che permettono di evidenziare le porzioni di testo maggiormente incidenti sulla segnalazione (c.d. "metodi *post-hoc*" come LIME e SHAP)[12].

4. Rischi per i diritti fondamentali e contromisure.

Lo sviluppo e l'utilizzo dei sistemi automatizzati per finalità preventive solleva questioni di particolare delicatezza sotto il profilo giuridico; questioni ineludibili, se l'obiettivo è quello di neutralizzare il serio rischio che un dispositivo di prevenzione del crimine finisca per rappresentare esso stesso una minaccia, potenzialmente ancor più incisiva, per i beni giuridici coinvolti, dando luogo ad una evidente eterogenesi dei fini.

Il perseguimento di tale finalità ha richiesto un approccio alla ricerca fondato sui principi in materia di libertà fondamentali e di protezione dei dati personali, senza dimenticare l'EU AI Act e, più in generale, la direttiva di azione secondo la quale il processo di trasformazione digitale in corso d'opera deve mettere al centro le persone, rispetto alle quali la tecnologia opera in chiave servente, nel rispetto dei loro diritti fondamentali e a garanzia della loro sicurezza[13].

La descrizione precedentemente svolta del meccanismo di funzionamento dei sistemi automatizzati chiarisce i rischi che la definizione delle *guidelines* ha mirato a contenere.

In primo luogo, il monitoraggio non deve sconfinare in forme di ingerenze indebite nella corrispondenza (art. 15 Cost.) e, più in generale, nella vita privata degli individui (art. 8 CEDU), né costituire uno strumento di sorveglianza di massa[14].

In secondo luogo, siffatti sistemi di monitoraggio algoritmico devono essere progettati in modo da non comprimere oltre ciò che è strettamente necessario la libertà comunicativa degli utenti, nel rispetto della loro libertà di espressione (artt. 21 Cost e 10 CEDU), e devono essere il più possibile esenti da errori sistemici, i quali possono produrre conseguenze rilevanti sugli individui, tra cui stigmatizzazioni o discriminazioni[15].

---

[12] Per approfondire questi profili, di carattere tecnico: S. Barocas - M. Hardt - A. Narayanan, *Fairness and Machine Learning. Limitations and Opportunities*, in www.fairmlbook.org, 2023; C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, in www.christophmolnar.com, 2022.

[13] Parlamento europeo – Consiglio – Commissione europea, *Dichiarazione europea sui diritti e i principi digitali per il decennio digitale*, 23 gennaio 2023, 2023/C 23/01.

[14] Con riferimento al problema del controllo e della sorveglianza, S. Zuboff, *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*, Roma, 2012, p. 529 e ss. Più di recente, C. F. Watson, *Protecting Children in the Frontier of Surveillance Capitalism,* in *Richmond Journal of Law & Technology*, vol. 27/2021.

[15] In argomento, F. Lagioia - G. Sartor, *Il sistema Compas: algoritmi, previsioni, iniquità*, in AA.VV., *XXVI Lezioni di diritto dell'intelligenza artificiale*, a cura di U. Ruffolo, Giappichelli, 2021, p. 226 ss.; S. Signorato, *Giustizia penale e intelligenza artificiale. Considerazioni in tema di algoritmo predittivo*, in *Riv. dir. proc.*, 2020, p. 614; G. Ubertis, *Intelligenza artificiale, giustizia penale, controllo umano significativo*, in AA.VV., *Giurisdizione penale, intelligenza artificiale ed etica del giudizio*, Giuffrè Francis Lefebvre, 2021, p. 12 ss.

Non da ultimo, deve essere garantita la conformità dei trattamenti di dati personali ai principi indicati nel GDPR, a partire dal principio di liceità, che impone la sussistenza di un'idonea base giuridica per ciascun trattamento di dati (art. 5, par. 1, lett. a) GDPR)[16].

Qualunque sforzo esegetico implica che sia dapprima risolta la questione di fondo: i principi fondamentali del nostro ordinamento giuridico consentono, ed eventualmente entro che limiti, la progettazione e l'impiego di sistemi di monitoraggio algoritmico con finalità preventiva del cyberbullismo?

4.1. Monitoraggio algoritmico delle comunicazioni e diritto alla riservatezza.

In questa indagine, la prima previsione normativa che assume rilevanza è l'art. 15 Cost., secondo il quale la libertà e la segretezza della corrispondenza, e di ogni altra forma di comunicazione, sono inviolabili. Si tratta di beni giuridici che possono subire limitazioni soltanto per atto motivato dell'autorità giudiziaria e nel rispetto delle ulteriori garanzie stabilite dalla legge. Non può disconoscersi sul punto anche la rilevanza dell'art. 8 CEDU, a mente del quale ogni individuo ha diritto al rispetto della propria vita privata e familiare, della propria abitazione e della propria corrispondenza. Le autorità pubbliche possono interferire con l'esercizio di tale diritto solo se tale interferenza è prevista dalla legge e costituisce una misura necessaria per la protezione di valori fondamentali[17].

Si impone l'esigenza di verificare se i sistemi di intelligenza artificiale in esame, già nella fase di addestramento, siano suscettibili di incidere sulla segretezza della corrispondenza, se non addirittura sulla sua libertà, dal momento che, in tali casi, l'impiego di tale metodica richiederebbe una previsione normativa relativa ai casi e alle garanzie di applicazione, prima tra le quali è senz'altro il provvedimento motivato dell'autorità giudiziaria[18].

---

[16] Sul tema, F.G. Viterbo, *The 'User-Centric' and 'Tailor-Made' Approach of the GDPR Through the Principles It Lays down*, in *The Italian Law Journal*, f. 2, pp. 631 ss.

[17] Avuto riguardo alla giurisprudenza della Corte europea dei diritti dell'uomo in tema di intercettazioni, l'assenza di una previa autorizzazione giudiziaria, la quale non è testualmente prevista dall'art. 8 CEDU, non dà necessariamente luogo ad una violazione convenzionale. La Corte, tuttavia, ha più volte evidenziato che, tra le garanzie che assumono primaria rilevanza nel giudizio sulla non arbitrarietà della «sorveglianza segreta» vi è la condizione secondo la quale l'autorità che autorizza la misura sia indipendente dal potere esecutivo (Corte Edu, G.c., 4 dicembre 2015, ric. n. 47143/06, *Roman Zakharov c. Russia*, in www.hudoc.echr.coe.int). Resta fermo che il controllo giurisdizionale preventivo «costituisce un'importante garanzia contro l'arbitrarietà» dell'ingerenza nella vita privata degli individui e «contribuisce a limitare il potere discrezionale delle autorità incaricate dell'applicazione di una legge formulata in maniera generale» (Corte Edu, Sez. I, 23 maggio 2024, ric. n. 2509/2019, *Contrada c. Italia n. 4*, in www.sistemapenale.it, 26 giugno 2024, con nota di L. Giordano, *Considerazioni sulla sentenza della Cedu Contrada c. Italia n. 4: per un'interpretazione convenzionalmente orientata delle norme di codice di rito*).

[18] Storicamente, il tema si è posto all'attenzione della dottrina ogni qualvolta nella prassi si è manifestato un mezzo tecnologico ad uso investigativo o preventivo che ha chiamato in causa i beni protetti dagli artt. 14 e 15 Cost. Rispetto al c.d. "captatore informatico", W. Nocerino, *Il captatore informatico nelle indagini interne e*

Il quesito impone di chiarire che la tutela costituzionale interessa la comunicazione *riservata*, protetta sia contro ostacoli al suo esercizio sia contro ingerenze nell'apprensione del contenuto[19]. Si tratta della trasmissione del pensiero rispetto alla quale, in ragione dell'«uso del mezzo espressivo» e delle «circostanze in cui avviene», è possibile escludere «una prevedibile possibilità di apprensione da parte dei terzi»[20]. La qualificazione dell'atto come riservato, dunque, interroga sulle sue «caratteristiche oggettive»[21]: non basta la volontà dell'autore del messaggio di delimitare la sfera dei riceventi, essendo necessario, dapprima, che il canale comunicativo sia «tecnicamente idoneo a garantire la segretezza»[22], e quindi lo scopo che il mittente si propone di raggiungere.

Per altro verso, è pacifico che la formula «corrispondenza», contenuta nell'art. 15 Cost., così come nell'art. 68 Cost., è «sufficientemente ampia da ricomprendere le forme di scambio di pensiero a distanza [via messaggi di posta elettronica e *WhatsApp*], costituenti altrettante "versioni contemporanee" della corrispondenza epistolare e telegrafica». Lo ha ribadito la Consulta nella celebre sentenza con la quale ha risolto il conflitto di attribuzione tra il Sen. Matteo Renzi e la Procura della Repubblica presso il Tribunale di Firenze[23].

La stessa pronuncia ha rafforzato il legame tra corrispondenza e riservatezza. Invero, secondo la Consulta, mantengono natura di corrispondenza i contenuti comunicativi «già ricevuti e letti dal destinatario, ma conservati nella memoria dei dispositivi elettronici del destinatario stesso o del mittente», «almeno fino a quando, per il decorso del tempo, essa non abbia perso ogni carattere di attualità, in rapporto all'interesse alla sua riservatezza, trasformandosi in un mero documento "storico"».

Le ricadute di queste coordinate d'insieme sono molteplici.

Laddove la comunicazione si svolga in un ambiente digitale aperto al pubblico (es: un "canale" pubblico di *Telegram*), non si verte in materia di "corrispondenza", dal momento che la comunicazione non presenta i connotati che consentono di qualificarla come "riservata". Per questo l'addestramento degli algoritmi e, in prospettiva, il loro impiego nelle piattaforme *social*, non trovano ostacolo nell'art. 15 Cost. Ciò vale non soltanto per la segretezza della comunicazione, ma anche per la sua libertà di esercizio. In una simile evenienza, si verte in materia di libertà di espressione. Il canone costituzionale di riferimento è l'art. 21 Cost., che la protegge, senza tuttavia apprestare, se non a tutela della stampa, una riserva di legge e di giurisdizione.

---

transnazionali, Wolters Kluwer, 2021, p. 210 ss.

[19] In merito alla sussistenza di due distinti, sebbene correlati, beni giuridici si veda V. Italia, *Libertà e segretezza della corrispondenza e delle comunicazioni*, Giappichelli, 1963, p. 91; C. Marinelli, I*ntercettazioni processuali e nuovi mezzi di ricerca della prova*, Giappichelli, 2017, p. 65.

[20] F. Cordero, *Il procedimento probatorio*, in *Tre studi sulle prove penali*, Giuffrè, 1963, p. 84, nota 234, rispetto alla comunicazione in forma verbale.

[21] F. Caprioli, *Colloqui riservati e prova penale*, Giappichelli, 2000, p. 46.

[22] Ibidem, p. 45.

[23] Corte cost. 27 luglio 2023, n. 170, in *Il diritto dell'informazione e dell'informatica*, 2023, f. n. 4-5, p. 708 ss., con nota di D. Curtotti, *La sentenza costituzionale n. 170 del 2023 e le comunicazioni "apparenti": quando un eccesso di garanzie non sempre è un moltiplicatore di garanzie*.

È problematico, invece, il monitoraggio automatico dei canali di messaggistica privata (es. una "chat *Whatsapp*"). In tal caso, le conversazioni si svolgono in un ambiente riservato, che resta tale anche una volta che il destinatario ha ricevuto e letto il messaggio. L'ingerenza sulla segretezza delle comunicazioni, pertanto, non è consentita in assenza di una base legale e di garanzie procedimentali, prima tra tutte l'atto motivato dell'autorità giudiziaria.

Alla luce di questa ricostruzione, nelle linee guida è stato previsto il divieto di addestrare gli algoritmi che impiegano tecniche idonee a visionare e carpire i contenuti comunicativi in ambienti digitali non connotati da pubblicità delle comunicazioni. Per analoghe ragioni, in prospettiva futura, per un eventuale utilizzo dei sistemi di intelligenza artificiale implementati nel corso del progetto, andrà ritenuta vietata l'attivazione di funzioni di monitoraggio[24].

Il gruppo di ricerca ha anche ipotizzato il caso in cui gli interlocutori prestino un consenso espresso, consapevole, specifico e attuale (relativo a ciascuna comunicazione in corso e per tutta la sua durata, specificamente individuata), rispetto all'addestramento degli algoritmi nelle *chat* private. In tale contesto, il consenso, se validamente prestato, potrebbe forse legittimare ingerenze nella comunicazione, laddove si ritenga mancante il requisito della sua riservatezza.

4.2. …e libertà di espressione.

Chiarite le implicazioni derivanti dal rapporto tra sistemi di intelligenza artificiale implementati nell'ambito del progetto e diritto alla riservatezza, occorre indagare l'impatto di tali meccanismi sulla libertà di espressione (artt. 21 Cost. e 10 CEDU).

Si tratta di un valore, questo, che potrebbe essere intaccato laddove il rilevamento di una comunicazione potenzialmente offensiva producesse come conseguenza "automatizzata" la rimozione o anche soltanto l'oscuramento del contenuto comunicativo. Come si è anticipato, taluni sistemi sperimentati nel progetto C.S.S. sono integrati da quest'ultima funzione.

Rispetto al problema della liceità dell'oscuramento o della rimozione automatizzati del contenuto offensivo o potenzialmente offensivo appare decisivo osservare che l'appartenenza ad una *community online* (es. *Twitter*) implica l'accettazione delle sue regole[25]. Tra queste regole può e, sempre più, tenuto conto del quadro regolatorio euro-unitario, *deve* essere prevista quella della inaccettabilità dei contenuti ritenuti violenti, e conseguentemente quella secondo la quale gli stessi verranno rimossi una volta individuati.

---

[24] Ciò vale quantomeno laddove tali metodiche venissero impiegate dai privati. Laddove, invece, esse servissero gli scopi investigativi, la questione risulterebbe più complessa. Si accennerà al tema al termine del contributo.

[25] Sul tema, cfr. C. Perlingieri, *Profili civilistici dei social network*, Napoli, 2014, p. 36 ss.

Così, l'art. 14, par. 1, *Digital Service Act* sancisce l'obbligo per i prestatori di servizi intermediari di includere nelle c.d. "*condizioni generali*", ossia le clausole che disciplinano il rapporto contrattuale con i destinatari del servizio (art. 3, par. 1, lett. *u*), «informazioni» riguardanti «le politiche, le procedure, le misure e gli strumenti utilizzati ai fini della moderazione dei contenuti, compresi il processo decisionale algoritmico e la verifica umana, nonché le regole procedurali del loro sistema interno di gestione dei reclami».

Tuttavia, nella misura in cui le piattaforme di *social network* operano come strumenti ordinari di manifestazione del pensiero, l'esercizio di tale "potere" risulta particolarmente delicato sul piano costituzionale, potendo risultare foriero di effetti ingiustamente pregiudizievoli per la libertà di espressione.

Ne consegue la necessità di adottare alcuni accorgimenti.

Il primo accorgimento, che opera a monte, è costituito dalla necessità di compiere uno sforzo, in fase di progettazione, per garantire un equilibrio tra la protezione dei minori e il rispetto della spontaneità delle interazioni *online*. Questo sforzo si traduce, a livello tecnico, nella necessità che gli algoritmi di rilevamento non siano configurati come meccanismi eccessivamente restrittivi. In altri termini, la finalità preventiva non dovrebbe giustificare la censura preventiva di qualunque comunicazione che possa ipoteticamente evolversi in uno scenario di cyberbullismo. È questa una direttiva di azione, modellata sul canone di necessarietà[26], il cui rispetto passa per la individuazione di segnali di rischio connotati da un adeguato grado di serietà, e che ha l'ulteriore effetto positivo di evitare un eccesso di segnalazioni che, altrimenti, rischierebbe di offuscare l'emersione di situazioni effettivamente problematiche.

Ancor prima, gli algoritmi non devono essere eticamente orientati e devono ridurre al minimo il rischio di falsi positivi o negativi, perché un sistema caratterizzato da elevato tasso di errore, oltre a risultare scarsamente utile, per inaffidabilità, si risolve nella ingiustificata compressione delle libertà fondamentali, dalla libertà di espressione sino alla libertà di iniziativa economica privata degli utenti. La mitigazione dei *bias* richiede procedure di valutazione continue e strategie di correzione che consentano di preservare l'affidabilità delle analisi.

In questo senso, e più in generale, appare irrinunciabile la c.d. "sorveglianza umana" sui risultati restituiti dal sistema algoritmico, rispetto alla quale si rivelano essenziali i suindicati di meccanismi di *explainable AI*. Invero, anche un sistema di intelligenza artificiale progettato per valorizzare i dati di contesto potrebbe non essere in grado di comprendere appieno la complessità della vicenda umana. Allora, in linea di principio, ogni segnalazione generata dal sistema deve essere sottoposta ad un controllo validante da parte di personale esperto, che sia in grado di contestualizzare le informazioni e determinare l'effettiva presenza di comportamenti a rischio.

Questo principio, tuttavia, deve essere adeguatamente tradotto in regole di azione. Il pericolo, infatti, è quello di posticipare eccessivamente l'intervento a tutela del minore e, in tal modo, vanificare lo scopo (preventivo) dello strumento impiegato. È

[26] In argomento, A. Sandulli, voce *Proporzionalità*, in *Dizionario di Diritto pubblico*, diretto da S. Cassese, vol. V, Giuffrè, 2006, p. 4643 ss.

sembrato dunque ragionevole, in un'ottica di bilanciamento, consentire che il contenuto che il sistema di intelligenza artificiale considera una minaccia concreta possa essere *oscurato* in via automatizzata (in un'ottica para-cautelare); tuttavia, l'oscuramento, per non assumere i connotati di una sostanziale cancellazione del contenuto, dovrebbe operare per il tempo strettamente necessario a consentire una rapida verifica da parte dei soggetti competenti.

Le linee guida, pertanto, hanno previsto un divieto di decisioni automatizzate, temperato dalla possibilità di un esercizio provvisorio del potere di oscuramento.

Sul punto, invece, il *Digital Service Act* appare maggiormente permissivo, nella misura in cui manifesta di tollerare decisioni definitive interamente automatizzate (art. 16, par. 6), pretendendo la "sorveglianza umana" soltanto a seguito della presentazione di un reclamo avverso tali decisioni e, dunque, nella fase "impugnatoria" in senso lato (art. 20, par. 6). Si tratta di una soluzione che è in grado di contemperare i diversi interessi in gioco, alla duplice condizione che il contenuto o il profilo censurato sia sempre ripristinabile e che la decisione sul reclamo venga adottata in tempi rapidi.

4.3. Applicabilità dell'EU AI Act, classificazione dei sistemi, implicazioni.

Il Regolamento (UE) 2024/1689 ha imposto ulteriori riflessioni.

Come è noto, tale normativa qualifica come «pratiche vietate» alcuni sistemi di intelligenza artificiale[27] mentre ad altri attribuisce un livello di rischio in base all'impatto prodotto sui diritti fondamentali[28]. Ci si è chiesti, allora, se i sistemi sviluppati nell'ambito del progetto *Cyber Social Security* potessero eventualmente rientrare tra le pratiche "vietate" o tra quelle "ad alto rischio". Nel primo caso, sarebbe venuta meno la stessa ragion d'essere dell'addestramento degli algoritmi. Nel secondo caso, invece, sarebbero conseguite una serie di conseguenze in punto di sviluppo, validazione e monitoraggio funzionale dei sistemi[29].

Occorre osservare che i sistemi di intelligenza artificiale progettati possono avvalersi della tecnica di estrazione massiva dei dati nota come "*web scraping*"[30] e che gli

---

[27] Tra i contributi di carattere generale relativi all'EU AI Act, si segnalano: R. Pardolesi, *Intelligenza artificiale (de)generativa? Svolazzi a cavallo tra ovvietà e disincanto*, in *Foro it.*, 2025, V, p. 5 e ss.; S. Orlando, *Sugli inviti rivolti dalla commissione europea all'Italia di modificare il d.d.l. sulla IA relativamente all'utilizzo dei sistemi di IA in ambito sanitario, nelle professioni intellettuali e nell'amministrazione della giustizia*, *ivi*, p. 49 e ss.; G. Vettori, *La "crisi del diritto" e il regolamento IA. Un esempio di positivismo inclusivo*, *ivi*, p. 12 e ss.; S. Pagliantini, *La base giuridica dell'AI Act ex art. 114 Tfue: l'intelligenza artificiale tra mercato e persona*, *ivi*, p. 19 e ss.

[28] In argomento, S. Orlando, *Regole di immissione sul mercato e pratiche di intelligenza artificiale vietate nella Proposta di Artificial Intelligence Act*, in *Persona e Mercato*, 2022, p. 346 e ss.

[29] Sul tema, E. Palmerini, *La governance dei sistemi ad alto rischio nell'Artificial Intelligence Act: uno sguardo panoramico*, in *Foro it.*, 2025, V, p. 35 e ss.

[30] Secondo il Garante per la protezione dei dati personali, *Delibera 20 maggio 2024, n. 329*, in Gazz. Uff. 7 giugno 2024, n. 132, «si parla di *web scraping* laddove l'attività di raccolta massiva ed indiscriminata di dati (anche personali) condotta attraverso tecniche di web crawling è combinata con un'attività consistente nella memorizzazione e conservazione dei dati raccolti dai bot per successive mirate analisi, elaborazioni ed utilizzi».

stessi sistemi hanno una dichiarata finalità di prevenzione del crimine, essendo finalizzati a monitorare, segnalare e oscurare i contenuti suscettibili di evolversi nell'offesa che la legge n. 71/2017 intende prevenire. Conseguentemente, trovano quali potenziali destinatari tanto i gestori dei *social network* quanto le forze dell'ordine.

Con riferimento alla metodica di addestramento e di funzionamento, va evidenziato che il *web scraping* costituisce una pratica vietata ai sensi dell'art. 5, lett. *e*), EU AI Act. E tuttavia il divieto interessa soltanto i sistemi che raccolgono massivamente immagini facciali dal *web* o filmati di telecamere a circuito chiuso per alimentare banche dati di riconoscimento facciale[31]. I sistemi C.S.S. non rientrano tra questi, non essendo in alcun modo finalizzati alla raccolta e alla identificazione dei dati biometrici.

Per escludere la sussistenza di una "pratica vietata" occorre confrontarsi anche con l'art. 5, lett. *d*), che proibisce alcuni sistemi di previsione del rischio di commissione di un reato da parte di una persona fisica (c.d. "*risk assessment tools*"[32]). Questo divieto concerne i sistemi che operano «unicamente sulla base della profilazione di una persona fisica o della valutazione dei tratti e delle caratteristiche della personalità», salvo che non siano utilizzati in funzione di supporto della valutazione umana - avente ad oggetto il coinvolgimento di una persona in un'attività criminosa - che si basa già su fatti oggettivi e verificabili direttamente connessi ad essa. Anche questo divieto non interessa i sistemi C.S.S., che non effettuano attività di profilazione finalizzata ad evitare la commissione di nuovi reati da parte di una specifica persona fisica. E ciò esclude pure che vadano classificati come "ad alto rischio" a norma dell'Allegato III, par. 6, lett. *d*) anch'esso concernente i sistemi di profilazione con finalità di giustizia predittiva[33].

Ciò non toglie che i sistemi implementati nel progetto C.S.S. hanno le caratteristiche per essere inquadrati tra quelli "ad alto rischio". Sono così classificati, a norma dell'Allegato III, par. 6, lett. *a*), quelli utilizzati dalle forze di polizia o per loro conto per determinare il rischio per una persona fisica di diventare vittima di reati. Di conseguenza, laddove i sistemi sviluppati fossero immessi sul mercato per essere in tal modo applicati dalle forze dell'ordine, dovrebbero rispettare tutta una serie di obblighi; obblighi che, invece, non sarebbero applicabili nel caso in cui fossero utilizzati dai gestori delle piattaforme *social*. Invero, la semplice moderazione dei contenuti comunicativi non attinge l'area applicativa della predetta normativa. Tuttavia, il progetto *Cyber Social Security* non è finalizzato ad un impiego da parte delle forze dell'ordine, ragione per la quale non si impone il problema del rispetto degli stringenti requisiti previsti dal regolamento.

---

[31] In argomento, E. Sacchetto, *Tecnologie di riconoscimento facciale e procedimento penale. Indagine sui fondamenti e sui limiti dell'impiego della biometria moderna*, Giappichelli, 2025.

[32] Sui "*risk assessment tools*", M. Gialuz, *Quando la giustizia penale incontra l'intelligenza artificiale: luci e ombre dei risk assessment tools tra Stati Uniti ed Europa*, in www.penalecontemporaneo.it, 29 maggio 2019, p. 3 ss.; L. Maldonato, *Risk and need assessment tools e riforma del sistema sanzionatorio: strategie collaborative e nuove prospettive*, in AA.VV., *Intelligenza artificiale e processo penale indagini, prove, giudizio*, a cura di G. Di Paolo - L. Pressacco, Napoli, 2022, p. 141 ss.; S. Quattrocolo, *Risk assessment: sentencing o non sentencing?*, in *Giurisdizione penale*, cit., p. 80.

[33] Con riferimento al problematico rapporto tra le due disposizioni del regolamento, S. Quattrocolo, *Intelligenza artificiale e processo penale: le novità dell'AI Act*, in www.dirittodidifesa.eu, 16.1.2025, p. 6.

Ad ogni modo, ai sensi dell'art. 2, par. 8, EU AI Act, il regolamento «non si applica alle attività di ricerca, prova o sviluppo relative a sistemi di IA o modelli di IA prima della loro immissione sul mercato o messa in servizio». Il rispetto di quegli obblighi, dunque, non si impone fintantoché il sistema di intelligenza artificiale non venga immesso sul mercato[34]. Ciò che conta è che le attività di ricerca, prova e sviluppo vengano «svolte in conformità del diritto dell'Unione applicabile», quindi, in particolare, al GDPR. Occorre comunque sottolineare che questa «esclusione» non si applica in caso di «prove in condizioni reali»[35].

Si è ritenuto comunque opportuno, tenuto conto degli interessi coinvolti, introdurre linee guida che fossero tendenzialmente orientate al rispetto di quelle previsioni. Peraltro, le linee guida hanno introdotto anche alcuni divieti, come quello di profilazione, finalizzati ad evitare di attingere le aree di rischio che richiedono obblighi più stringenti, oltre che l'onere di introdurre misure tecniche e misure di controllo in grado di impedire e di correggere eventuali *bias* di sistema, anche in coerenza con quanto precedentemente indicato.

4.4. Base giuridica del trattamento e principio di proporzionalità alla luce del GDPR.

Tema centrale è quello del rispetto della disciplina in materia di protezione dei dati personali, da analizzare in ossequio al principio di *accountability* di cui all'art. 5, par. 2, GDPR[36].

Al riguardo, occorre una breve premessa sulla titolarità dei trattamenti di dati personali eventualmente implicati dal progetto *Cyber Social Security*.

Nella fase di elaborazione del sistema di intelligenza artificiale volto al monitoraggio e all'individuazione di episodi di cyberbullismo, gli Atenei coinvolti nel progetto hanno agito come autonomi titolari del trattamento *ex* artt. 4, par. 1, n. 7) e 24 GDPR: ciascun Ateneo, nell'ambito dei processi di propria spettanza, ha dunque determinato le finalità e i mezzi del trattamento di dati personali, ove necessario. Di contro, nell'ambito della successiva fase di funzionamento (per quanto detto, ipotetica),

---

[34] Sul punto, G. Resta, *Commento all'art. 2 (commi 1-4)*, in *Intelligenza Artificiale. Commentario*, a cura di A. Mantelero - G. Resta - G.M. Riccio, Ipsoa, 2025, p. 16 ss.

[35] L'EU AI Act dedica un apposito articolo (art. 76) ai controlli dei *test* che devono essere effettuati in condizioni reali per i sistemi di intelligenza artificiale dall'autorità di vigilanza del mercato, la quale può richiedere al fornitore di condurre gli stessi e di fornire i risultati. I *test* in condizioni reali non sono altro che prove temporanee effettuate su sistemi di intelligenza artificiale al di fuori di uno spazio simulato, volti a valutare e verificare la conformità del prodotto ai requisiti del regolamento: così, S.A. Ibrahim El Sabi, *Commento all'art. 3 (numero 57)*, in *Intelligenza Artificiale. Commentario*, cit., p. 97 ss.

[36] Sul GDPR, tra i contributi di carattere generale più significativi: AA.VV., *I dati personali nel diritto europeo*, a cura di V. Cuffaro - R. D'Orazio - V. Ricciuto, Giappichelli, 2019; AA.VV., *Il nuovo regolamento europeo sulla privacy e sulla protezione dei dati personali*, opera diretta da G. Finocchiaro, Zanichelli, 2017; G. Finocchiaro, *Riflessioni sul poliedrico regolamento europeo sulla privacy*, *Quad. cost.*, 2018, p. 895; F.G. Viterbo, *Protezione dei dati personali e autonomia negoziale*, Edizioni Scientifiche Italiane, 2008. Alcuni riferimenti alla letteratura straniera: AA.VV., *Data Protection Around the World*, a cura di E. Kiesow Cortez, The Hague, 2021; M. Krzysztofek, *GDPR: General Data Protection Regulation (EU) 2016/679*, Wolters Kluwer, 2019.

è ipotizzabile che il titolare del trattamento di dati personali sia il soggetto, privato o pubblico, che utilizzerà concretamente il sistema, ad esempio, ai fini della gestione della piattaforma *online* o delle finalità di pubblico interesse connesse alla funzione istituzionale svolta.

Quanto osservato sottende che, nell'ambito del progetto, siano stati trattati dati personali. Il tema merita di essere approfondito, posto che, in tal caso, in ossequio al principio di liceità del trattamento, è necessario dapprima individuare un'adeguata base giuridica che giustifichi i suddetti trattamenti di dati e poi verificare la conformità dei medesimi ai principi di trasparenza, minimizzazione, esattezza, integrità e riservatezza. Al fine di rispondere all'interrogativo di fondo, bisogna chiedersi cosa siano i dati personali e perché i sistemi implementati nel progetto potrebbero comportare dei rischi per i diritti delle persone cui i dati medesimi si riferiscono (c.d. "interessati").

Ai sensi dell'art. 4, par. 1, GDPR, il dato personale è «qualsiasi informazione riguardante una persona fisica identificata o identificabile ("interessato"); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo *online* o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale». Il trattamento, invece, è «qualsiasi operazione o insieme di operazioni, compiute con o senza l'ausilio di processi automatizzati e applicate a dati personali o insiemi di dati personali, come la raccolta, la registrazione, l'organizzazione, la strutturazione, la conservazione, l'adattamento o la modifica, l'estrazione, la consultazione, l'uso, la comunicazione mediante trasmissione, diffusione o qualsiasi altra forma di messa a disposizione, il raffronto o l'interconnessione, la limitazione, la cancellazione o la distruzione»[37].

Ebbene, i diritti degli interessati potrebbero essere innanzitutto intaccati nelle modalità di addestramento degli algoritmi di intelligenza artificiale, dal momento che il progetto non è insensibile all'impiego della tecnica di raccolta massiva dei dati nota come "*web scraping*". Il Garante italiano per la protezione dei dati personali non ha escluso la liceità della pratica, e quindi sembra averla ritenuta ammissibile, pur nel rispetto dei principi e delle regole previste sulla protezione dei dati personali[38]. In secondo luogo, è chiaro che questi dati vengono trattati quando il sistema si trova a monitorare e a segnalare comportamenti sospetti.

---

[37] Sulla definizione di dato personale: P. Blume, *The Data Subject*, 1 *European Data Protection Law Review,* 2015, p. 258 ss.; R. Ducato, *La crisi della definizione di dato personale nell'era del web 3.0*, in *Le definizioni nel diritto*, Editoriale scientifica, 2016, p. 149 ss.; A. Nervi, *Il perimetro del Regolamento europeo: portata applicativa e definizioni*, in *I dati personali nel diritto europeo*, a cura di V. Cuffaro - R. D'Orazio - V. Ricciuto, Giappichelli, 2019, p. 161 ss. Sul piano comparativo e sulla definizione di *Personally Identifiable Information* nell'esperienza giuridica statunitense: P. M. Schwartz - D.J. Solove, *Reconciling Personal Information in the United States and European Union*, in California Law Review, vol. 102/2014, p. 877 ss.; Id, T*he PII Problem: Privacy and a New Concept of Personally Identifiable Information*, in *New York Law Review*, vol. 86/2011, p. 1814 ss; P. M. Schwartz - K. N. Peifer, *Transatlantic Privacy Law*, in *The Georgetown Law Journal,* vol 106/2017, p. 115 ss.

[38] Garante per la protezione dei dati personali, *Delibera 20 maggio 2024, n. 329*, cit.

La circostanza che possano essere trattati (massicciamente) dati personali (anche sensibili) richiede l'individuazione di una solida base giuridica, da rintracciare nell'art. 6 GDPR, che individua le «condizioni» al ricorrere delle quali il trattamento dei dati «è lecito». Fra le diverse basi giuridiche ivi elencate, si ritiene che quella più adeguata al caso di specie sia rintracciabile nella lett. *f*): è lecito il trattamento che «è necessario per il perseguimento del legittimo interesse del titolare del trattamento o di terzi, a condizione che non prevalgano gli interessi o i diritti e le libertà fondamentali dell'interessato che richiedono la protezione dei dati personali, in particolare se l'interessato è un minore». Pertanto, tale base giuridica è integrata al ricorrere di tre condizioni cumulative, da interpretare alla luce delle indicazioni fornite dall'*European Data Protection Board*[39]: 1. il perseguimento di un interesse legittimo da parte del responsabile del trattamento o di terzi, il quale è tale se è lecito, chiaro e attuale (di esso deve essere informato l'interessato); 2. la necessità di trattare i dati personali al fine di perseguire l'interesse legittimo, per l'effetto che occorre accertare se mezzi meno invasivi non potevano essere impiegati in modo altrettanto efficace; 3. nel bilanciamento tra gli interessi in gioco, le libertà e i diritti fondamentali degli utenti interessati non devono risultare prevalenti rispetto agli interessi legittimi del titolare del trattamento o di terzi[40].

È questa la pertinente base giuridica, atteso che l'impiego dei dati personali, nell'ambito del progetto C.S.S., si è reso necessario per scopi di *ricerca scientifica*, la quale costituisce un legittimo interesse, rispondente a quello perseguito dallo Stato italiano e dall'Unione europea, che ha finanziato la ricerca. Tale interesse si compenetra, dunque, con le finalità del trattamento, le quali, ai sensi dell'art. 5, par. 1, lett. *b*), devono essere «determinate, esplicite e legittime». Peraltro, sebbene i soggetti interessati al trattamento, in genere, siano minorenni, costituisce un fattore significativo nel giudizio di bilanciamento relativo alla liceità del trattamento, la circostanza che la finalità del progetto è costituita dalla tutela dei soggetti minorenni da condotte di cyberbullismo.

Rispetto a tale trattamento di dati personali, inoltre, il titolare sembra esonerato dagli obblighi di trasparenza previsti dall'art. 14, par. 1-4, GDPR (informazioni da fornire qualora i dati personali non siano ottenuti presso l'interessato), in quanto al caso di specie dovrebbe trovare applicazione il successivo par. 5, lett. *b*). In tale disposizione, infatti, si fa riferimento, tra l'altro, al trattamento a fini di ricerca scientifica rispetto al quale comunicare le informazioni prescritte risulterebbe impossibile o implicherebbe uno sforzo sproporzionato. Tuttavia, sono fatte salve le condizioni e le garanzie di cui all'art. 89, par. 1, GDPR. Resta fermo che «il titolare del trattamento adotta misure appropriate per tutelare i diritti, le libertà e i legittimi interessi dell'interessato, anche rendendo pubbliche le informazioni».

---

[39] European Data Protection Board, *Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR*, 8 ottobre 2024.

[40] Risalente ma utile il riferimento a: F. Ferretti, *Data Protection and the Legitimate Interest of Data Controllers: Much Ado About Nothing or the Winter of Rights?*, in *Common Market Law Review*, in vol. 51/2014, p. 843 ss. Sul problema relativo all'utilizzo del consenso quale base giuridica per la ricerca scientifica: D. Sborlini, *Il broad consent come mezzo per la valorizzazione dei dati personali nell'ambito della ricerca scientifica e il suo rilievo negli spazi di condivisione dei dati*, in *Contratto e Impresa*, 2024, p. 223 ss.

Il problema in esame è reso più complesso dalla circostanza che il trattamento svolto nell'ambito del progetto ha coinvolto anche "categorie particolari di dati personali". In tal caso, occorre che il trattamento sia altresì conforme all'art. 9 GDPR[41]. Tale previsione, al par. 1, vieta di «trattare dati personali che rivelino l'origine razziale o etnica, le opinioni politiche, le convinzioni religiose o filosofiche, o l'appartenenza sindacale» e vieta altresì di «trattare dati genetici, dati biometrici intesi a identificare in modo univoco una persona fisica, dati relativi alla salute o alla vita sessuale o all'orientamento sessuale della persona»[42].

Tuttavia, lo stesso art. 9, al par. 2, prevede tutta una serie di eccezioni alla regola. Una di queste è quella di cui alla lett. *j*): «il trattamento è necessario a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici in conformità dell'articolo 89, paragrafo 1, sulla base del diritto dell'Unione o nazionale, che è proporzionato alla finalità perseguita, rispetta l'essenza del diritto alla protezione dei dati e prevede misure appropriate e specifiche per tutelare i diritti fondamentali e gli interessi dell'interessato». Indubbiamente, in base a quanto già osservato sopra, il trattamento dei dati è qui funzionale tanto al pubblico interesse quanto alla ricerca scientifica. Pertanto esso è lecito, a patto, però, che rispetti l'art. 89 par. 1, oltre che al principio di minimizzazione dei dati e, più in generale, al canone di proporzionalità tra trattamento dei dati e scopo perseguito[43].

Ai sensi dell'art. 89, par. 1, «il trattamento a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici è soggetto a garanzie adeguate per i diritti e le libertà dell'interessato». Le «garanzie adeguate» consistono nella predisposizione di «misure tecniche e organizzative, in particolare al fine di garantire il rispetto del principio della minimizzazione dei dati». Le misure «possono includere la pseudonimizzazione, purché le finalità in questione possano essere conseguite in tal modo». Si aggiunge che «qualora possano essere conseguite attraverso il trattamento ulteriore che non consenta o non consenta più di identificare l'interessato, tali finalità devono essere conseguite in tal modo»[44].

Sulla scorta di questa analisi, si è ritenuto che il trattamento dei dati personali, anche di quelli rientranti nelle "categorie particolari", posto in essere nella fase di sviluppo del sistema cui è diretto il progetto *Cyber Social Security*, perseguisse una finalità legittima in quanto svolto per scopi di ricerca scientifica, ma dovesse rispettare il principio di proporzionalità, *sub specie* di necessarietà e di minimizzazione dei dati oggetto di trattamento, anche attraverso la tecnica della pseudonimizzazione.

---

[41] *Cfr*. L. GEROGIEVA - C. KUNER, sub *art. 9*, in *The EU General Data Protection Regulation (GDPR). A Commentary*, a cura di C. KUNER - L.A. BYGRAVE - C. DOCKSEY, Oxford University Press, 2020, p. 369 ss.

[42] Tra i numerosi contributi si segnala: M. Granieri, *Il trattamento di categorie particolari di dati personali nel Reg. UE 2016/679*, in *Nuove leggi civ. comm.*, 2017, p. 165 ss.

[43] Sul regime giuridico dei dati per ricerca scientifica: P. Guarda, *Il regime giuridico dei dati della ricerca scientifica*, Editoriale scientifica, 2021; I. Rapisarda, I. *Ricerca scientifica e circolazione dei dati personali*, in *Eur. dir. priv.*, 2021, p. 326 ss.; F.G. Viterbo, *Governance and processing of personal data in the Italian healthcare system in the light of EU principles*, in *Actualidad Jurídica Iberoamericana*, 2024, p. 1052 ss.

[44] *Cfr*. C. WIESE SVANBERG, *sub* art. *89*, in *The EU General Data Protection Regulation (GDPR). A Commentary*, cit., p. 1242 ss.

Con riferimento al principio di minimizzazione si deve aver riguardo in primo luogo all'art. 5, par. 1, lett. *c*), a mente del quale i dati personali devono essere «adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per le quali sono trattati». Il titolare deve osservare tale principio in ogni fase, sia in quella di raccolta, sia nel successivo trattamento e nella fase di conservazione dei dati personali e procedere alla anonimizzazione dei dati tutte le volte che ciò non sia di ostacolo alla realizzazione dello scopo perseguito.

Al riguardo, è bene evidenziare che l'art. 5, lett. *e*), legittima la conservazione dei dati per un tempo superiore al conseguimento della finalità del trattamento a condizione che siano trattati esclusivamente a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici, purché sempre nel rispetto di quanto previsto dall'art. 89, par. 1, e fatta salva l'attuazione di misure tecniche e organizzative adeguate. Anche in tal caso, cioè là dove risulti possibile l'ulteriore trattamento dei dati dopo la conclusione del progetto di ricerca, il titolare del trattamento, ai sensi dell'art. 5, lett. *f*), mantiene l'obbligo di garantire un'adeguata sicurezza dei dati conservati, soprattutto al fine di evitare trattamenti «non autorizzati o illeciti» e ridurre al minimo i rischi di perdita, distruzione o danno accidentali.

Riguardo, infine, al trattamento di dati personali, compresi quelli rientranti nelle "categorie particolari", ipotizzabile nella fase di sperimentazione successiva e di concreto utilizzo del sistema di controllo e individuazione di episodi di cyberbullismo, si è considerata l'ipotesi in cui a ricorrere al sistema sia un soggetto privato che gestisce una piattaforma di *social network*. In tal caso, sarà necessario fornire agli utenti/interessati le informazioni relative all'impiego del sistema di monitoraggio, assicurarsi che siano adempiuti gli obblighi di trasparenza di cui all'art. 13 GDPR, nonché acquisire il consenso esplicito di ciascun utente/interessato al trattamento dei dati personali anche sensibili per le finalità relative al funzionamento del sistema, ai sensi degli artt. 6, par. 1, lett. *a*) e 9, par. 2, lett. *a*) GDPR. In generale, il titolare del trattamento dovrà garantire, a tal fine, il rispetto della normativa in materia di protezione dei dati personali, se necessario anche in coordinamento con quanto previsto dal *Digital Services Act*.

4.5. Verso il "Digital Omnibus".

Sul tema, è stato necessario confrontarsi anche con le proposte di modifica del GDPR formulate nella bozza del c.d. "*Digital Omnibus*[45].

Si tratta del tentativo della Commissione europea di modificare una serie di normative per garantire, tra l'altro, una *governance* dei dati personali ispirata a semplificazione, a vantaggio della competitività delle imprese dell'Unione europea[46].

---

[45] Commissione europea, *Proposta di regolamento UE "Digital Omnibus"*, COM(2025) 837 final, 19 novembre 2025.

[46] Clément-Fontaine, *La rationalisation du droit numérique européen: le rapport parlementaire sur l'AI Act, prélude aux règlements Digital Omnibus et Omnibus IA*, in *Dalloz Actualité*, 14 novembre 2025.

Si è ritenuto che l'approvazione delle proposte non minerebbe la validità delle linee guida, atteso che il regolamento avrebbe l'effetto di attenuare gli obblighi euro-unitari sui quali si è basata l'analisi esegetica elaborata nel corso del progetto, tanto in tema di tutela dei dati personali quanto di tutela dei diritti fondamentali dai sistemi di intelligenza artificiale.

Occorre segnalare che il *Digital Omnibus* andrebbe a mutare finanche la nozione di "dato personale", modificando l'art. 4 GDPR sul concetto di persona fisica identificabile. Invero, la persona dovrebbe essere ritenuta identificabile, e quindi il dato "personale", soltanto nel caso in cui una data entità, ragionevolmente, abbia gli strumenti per procedere all'identificazione, e non già se ciò sia soltanto in astratto possibile. In tal modo, si va verso una valorizzazione delle operazioni di pseudonimizzazione, in coerenza con la più recente giurisprudenza della Corte di giustizia[47].

Inoltre, verrebbe introdotta nell'art. 4 GDPR la nozione di «ricerca scientifica», chiarendo che essa include le attività che mirano allo sviluppo tecnologico, e che non è incompatibile con un'ulteriore finalità di natura commerciale.

La proposta mira anche a modificare l'art. 5, lett. *b*), GDPR, rafforzando la base giuridica che giustifica la necessarietà del trattamento dei dati per la finalità di ricerca scientifica. Viene promosso il c.d. "uso secondario" dei dati, il cui trattamento viene sganciato dalle condizioni previste dall'art. 6 par. 4 GDPR[48].

Soprattutto, il *Digital Omnibus* tiene conto dell'ipotesi in cui un sistema di intelligenza artificiale vada ad impattare sulle categorie particolari di dati personali. Se la proposta venisse approvata, il nuovo art. 9, par. 2, lett. *k*), andrebbe a prevedere in maniera cristallina quanto il nostro studio ha ricomposto in via esegetica: il divieto di cui all'art. 9, par. 1, non troverebbe applicazione al trattamento finalizzato allo sviluppo di un sistema di intelligenza artificiale.

Non solo. Il nuovo par. 5 andrebbe a cristallizzare nel dato normativo la ricostruzione interpretativa effettuata dal gruppo di lavoro con riguardo alla necessità di rispettare, durante lo sviluppo dei sistemi di monitoraggio algoritmico, il principio di minimizzazione delle categorie particolari di dati, nel rispetto di un canone di esigibilità. Se possibile, occorre *evitare* il trattamento dei dati già nel corso dell'addestramento dell'algoritmo; laddove questo risultato non sia del tutto praticabile, occorre procedere alla *rimozione* rispetto al *set* di dati archiviati: laddove, tuttavia, lo sforzo sia sproporzionato, occorre adottare qualunque misura tempestiva volta ad evitare ogni impiego successivo o divulgazione a terzi.

---

[47] Corte giust., 4 settembre 2025, C-413/23, "Deloitte", § 86.

[48] Sul problema dell'uso secondario dei dati nell'ambito sanitario: F. Cascini - M. A. Arcuri, *Uso secondario dei dati personali relativi alla salute: panoramica della normativa europea e nazionale*, in *Dir. inf.*, 2024, p. 837 ss.; AA.VV., *Ricerca in sanità e protezione dei dati personali. Scenari applicativi e prospettive future*, a cura di P. Guarda - E. Chizzola - V. Maroni - L. Rufo, Editoriale scientifica, 2024, p. 1 ss.

5. Prospettive.

I risultati del progetto *Cyber Social Security* consentono di delineare alcune prospettive riguardanti l'eventuale futura messa in esercizio dei sistemi di monitoraggio implementati.

Una prospettiva in chiaro-scuro è rappresentata dalla messa a disposizione dei sistemi ai gestori delle piattaforme di *social network*. Se, da un lato, tali soggetti già sono tenuti a rimuovere i contenuti offensivi su segnalazione della vittima di cyberbullismo o di terzi, dall'altro lato non sono gravati da un obbligo di monitoraggio preventivo, il quale, per quanto detto, risulterebbe attualmente inesigibile e scarsamente opportuno. L'introduzione di accurati strumenti di intelligenza artificiale potrebbe favorire questo passaggio.

Il rischio è quello di una deriva privatistica nella prevenzione dei *cybercrime*, e quindi nel governo della libertà di espressione; un rischio che il rispetto delle *guidelines* dovrebbe essere in grado di neutralizzare. Del resto, è ormai imprescindibile la collaborazione delle piattaforme nella prevenzione e nel contrasto al fenomeno criminale.

Da questo punto di vista, i sistemi di intelligenza artificiale sviluppati nel progetto potrebbero essere ulteriormente perfezionati, mediante la predisposizione di meccanismi in grado di segnalare automaticamente all'autorità pubblica eventuali *notitiae criminis*.

È chiaro che i dati che finiscono sotto la lente degli algoritmi possono assumere valore di prova nei procedimenti aventi finalità preventive (es: il c.d. "ammonimento"), investigative e processuali. Per questo, per essere eventualmente trasmessi alle autorità competenti, tali dati, anche laddove vengano espunti dal *web* per finalità di tutela del soggetto minorenne, devono essere conservati, a norma della legge n. 71/2017[49]. Ben diverso, però, è immaginare meccanismi di partecipazione attiva delle piattaforme nell'accertamento del *cybercrime.*

Allo stato attuale, la trasmissione dei dati è spesso ostacolata dalla ubicazione dei servizi in Paesi diversi da quello procedente. Di regola, ciò richiede l'attivazione dei meccanismi di cooperazione giudiziaria[50]. Ma nell'immediato futuro sempre più si farà ricorso a canali di cooperazione non mediata tra autorità pubbliche e piattaforme[51]. In

---

[49] In particolare, l'art. 2, comma 1, legge n. 71/2017 dispone che, nel procedere all'oscuramento, alla rimozione o al blocco di contenuti offensivi, in accoglimento dell'istanza della vittima di cyberbullismo, il titolare del trattamento o il gestore del sito internet o del social media è tenuto alla «conservazione dei dati originali».

[50] In argomento, tra la vasta letteratura in materia, si segnala S. Signorato, *Le indagini digitali. Profili strutturali di una metamorfosi investigativa*, Giappichelli, p. 201; F. Siracusano, *La prova informatica transnazionale: un difficile "connubio" fra innovazione e tradizione*, in *Proc. pen. giust.*, 2017. p. 195.

[51] Un significativo passo in avanti verso questa direzione è rappresentato dal regolamento UE/2023/1543, il cui processo di attuazione interna è stato avviato con la Legge 13 giugno 2025, n. 91 (c.d. "Legge di delegazione europea 2024"). Esso consente ad uno Stato membro, nell'ambito di un procedimento penale, di ingiungere a un prestatore di servizi «della società dell'informazione», che opera nell'Unione europea e che è stabilito in altro Stato membro o che si avvale di un rappresentante legale in altro Stato membro, la

tale contesto, è ipotizzabile che maturi l'esigenza di un ruolo proattivo da parte degli attori privati nella giustizia penale[52]. Questa esigenza, per l'appunto, potrebbe essere soddisfatta da sistemi di rilevamento e segnalazione automatizzata dei reati commessi sul *web*. Fin troppo evidente che, in tal caso, si farebbe ancora più pressante l'esigenza di evitare pregiudizi sistemici, visto il potenziale impatto sulle libertà costituzionalmente protette.

Estremamente delicato è lo scenario in cui i sistemi di monitoraggio algoritmico vengano impiegati da parte delle autorità di *law enforcement*. L'ipotesi che appare astrattamente formulabile è quella di un loro utilizzo per l'analisi automatizzata *real time* degli scambi comunicativi o dei documenti multimediali nel corso dello svolgimento delle intercettazioni telematiche e dell'analisi del materiale giacente sui *server* dei fornitori di servizi digitali. Tali sistemi consentirebbero di semplificare l'attività investigativa, dal momento che all'analisi del contenuto monitorato o ispezionato seguirebbero segnalazioni automatizzate dei contenuti potenzialmente allarmanti, questo tanto con una finalità di accertamento, quanto con una finalità di prevenzione del *cybercrime*.

Questi sistemi di monitoraggio, evidentemente, esulano dalla presente ricerca. Il tema potrà essere alla base di un futuro studio, nell'ambito del quale ai giuristi spetterebbe il compito di valutare non soltanto i requisiti atti a garantire la rispondenza con gli obblighi previsti dall'EU AI Act per i sistemi ad alto rischio e dal GDPR per il trattamento dei dati personali, ma anche e soprattutto la loro strutturale compatibilità con la doppia riserva sancita dall'art. 15 Cost. e con le garanzie di cui all'art. 8 CEDU. Si tratterebbe di verificare se, in assenza di una specifica disciplina, la legittimazione degli strumenti di monitoraggio automatizzato possa essere rinvenuta nell'ambito di mezzi investigativi già regolamentati. Il quesito interroga sulla natura dei sistemi oggetto di studio: *nuovi strumenti* di controllo o *modalità tecniche ancillari* alle attività investigative già disciplinate?

Il tema rientra nella più generale questione posta dalla necessità di regolamentare gli usi consentiti dell'intelligenza artificiale a fini investigativi e predittivi. Sul punto, si attende l'intervento del Governo, delegato dall'art. 24, comma 5, lett. *e*), legge 23 settembre 2025, n. 132, a disciplinare la «regolazione dell'utilizzo dei sistemi di

---

produzione o la conservazione di «prove elettroniche», «indipendentemente dall'ubicazione dei dati» (art. 1). Le prove elettroniche sono «i dati relativi agli abbonati, i dati sul traffico o i dati relativi al contenuto conservati in formato elettronico da o per conto di un prestatore di servizi» (art. 3). Con riguardo alla genesi della disciplina e agli interessi in gioco, D. Curtotti, I*ndagini hi-tech, spazio cyber, scambi probatori tra Stati e Internet provider service e "Vecchia Europa": una normativa che non c'è (ancora)*, in *Dir. pen. proc.*, 2021, p. 745 ss. Sul regolamento, F. Sanvitale, *La tutela dei diritti nel c.d. e-evidence package: il lato oscuro della cooperazione diretta*, in *Cass. pen.*, 2025, p. 1475 ss.

[52] Non mancano casi in cui i gestori delle piattaforme già sono tenuti a comunicare notizie di reato all'autorità pubblica. L'art. 18 *Digital Services Act*, infatti, prevede che il prestatore di servizi che sia venuto a conoscenza di «informazioni che fanno sospettare che sia stato commesso, si stia commettendo o probabilmente sarà commesso un reato che comporta una minaccia per la vita o la sicurezza di una o più persone», deve informare «senza indugio le autorità giudiziarie o di contrasto dello Stato membro o degli Stati membri interessati»,«fornendo tutte le informazioni pertinenti disponibili».

intelligenza artificiale nelle indagini preliminari, nel rispetto delle garanzie inerenti al diritto di difesa e ai dati personali dei terzi, nonché dei principi di proporzionalità, non discriminazione e trasparenza». Il riferimento alle indagini preliminari lascia comunque intendere che la materia preventiva resterà problematicamente esclusa dall'intervento normativo.

In conclusione, gli esiti del progetto *Cyber Social Security* aprono scenari al contempo promettenti e critici, suscettibili di rilevanti implicazioni sul piano teorico e applicativo. Tali prospettive appaiono promettenti nella misura in cui dimostrano la possibilità di progettare sistemi di intelligenza artificiale orientati alla prevenzione del cyberbullismo e compatibili con la tutela dei diritti fondamentali; esse risultano, al contempo, problematiche poiché mettono in luce l'esistenza di vuoti di normazione destinati ad assumere un rilievo crescente al progredire dell'innovazione tecnologica. La questione centrale per il futuro consisterà nell'evitare che strumenti concepiti come ausilio alla protezione dei minori e alla sicurezza degli ambienti digitali possano, in assenza di una disciplina specifica, degenerare in forme di sorveglianza di massa. È lungo questo delicato crinale che si misurerà la sostenibilità giuridica dei sistemi di monitoraggio algoritmico nei prossimi anni.

Editore

ASSOCIAZIONE
"PROGETTO GIUSTIZIA PENALE"