Ordine internazionale e diritti umani
International Legal Order and Human Rights
Ordenamiento Jurídico Internacional y Derechos Humanos
Ordre Juridique International et Droits de l'Homme
Diretta da Claudio Zanghì, Lina Panella, Carlo Curti Gialdino
EDITORIALE SCIENTIFICA

OSSERVATORIO NUOVE TECNOLOGIE E DIRITTI FONDAMENTALI

Coordinato da Francesco Battaglia

N.4
15 ottobre 2026

Jean Monnet Chair 2024-2027
DREAM-EU
Digital Rights Empowerment Across Market in the European Union

Co-funded by
the European Union

Ordine internazionale e diritti umani
International Legal Order and Human Rights
Ordenamiento Jurídico Internacional y Derechos Humanos
Ordre Juridique International et Droits de l'Homme
Diretta da Claudio Zanghì, Lina Panella, Carlo Curti Gialdino
EDITORIALE SCIENTIFICA

OSSERVATORIO NUOVE TECNOLOGIE E DIRITTI FONDAMENTALI N. 4/2026

1. L'INTELLIGENZA ARTIFICIALE COME STRUMENTO PER PREVENIRE E REPRIMERE LA VIOLENZA DI GENERE CONTRO LE DONNE

1. L'intelligenza artificiale (IA) può essere – ed è ampiamente – impiegata per manipolare immagini, generare messaggi d'odio, diffondere disinformazione, riprodurre stereotipi di genere (*gender-bias*), esacerbare disuguaglianze strutturali e generare violenza di genere che colpisce in modo sproporzionato le donne (tra i numerosi studi, v. AI FORENSICS, *Harassment As Infrastructure*, rapporto del 2026). L'IA viene, ad esempio, sfruttata per produrre *deepfake*, un immagine o video che presenta figure corporee e facciali catturate in Internet – o "rubate" tramite la condivisione non consensuale – rielaborate e adattate a un contesto diverso da quello originario tramite un sofisticato algoritmo (questa definizione è una rivisitazione più aggiornata, che contempla ad esempio il furto di immagini senza il consenso della persona ritratta, della definizione contenuta nella Treccani, che ha inserito *deepfake* tra i neologismi del 2018). Il termine descrive sia la tecnologia che il contenuto fittizio che ne deriva. Quando a sfondo sessuale e non consensuali, il *deepfake* costituisce una forma di abuso sessuale basato sulle immagini ed è una forma di violenza di genere nel mondo digitale generata dall'IA (sul concetto di *image-based sexual abuse*, v. C. MCGLYNN, E. RACKLEY, R. HOUGHTON, *Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse*, in *Feminist Legal Studies*, 25/2017, p. 25 ss.).

Esiste tuttavia un altro volto dell'IA: se da un lato essa può essere fonte di violenza e riprodurre dinamiche di subordinazione e oppressione, dall'altro può altresì costituire una risorsa potenziale all'interno di strumenti concepiti per proteggere le vittime e garantire il consenso (ad es. la moderazione dei contenuti guidata dall'IA, le applicazioni di supporto alle vittime e i sistemi di verifica del consenso informato), di strumenti volti a rafforzare il rispetto delle regole come i Termini di Servizio delle piattaforme online (ad es. mediante avvisi automatici generati dall'IA inviati agli utenti), nonché nella rilevazione di contenuti illeciti in rete. Nuove applicazioni e *chatbot* (si veda, ad esempio, FollowItApp, NonPossoParlare, Sophia Chat) sono state sviluppate per sostenere le vittime di violenza di genere e violenza domestica, fornendo un canale sicuro e anonimo per riconoscere le dinamiche relazionali, raccogliere prove utilizzabili in eventuali futuri procedimenti contro chi perpetra l'abuso e individuare i pregiudizi algoritmici (si veda a riguardo il focus O. GIOLO, M. BERNARDINI (a cura di), *L'algoritmo alla prova del caso concreto: stereotipi, serializzazione, discriminazione*, in *GenIUS*, 1/2022, p. 6 ss.). Come riconosciuto dal Consiglio dell'Unione europea, nelle conclusioni del 2025, «quando si applicano regole chiare e prive di pregiudizi e si utilizzano dati di alta qualità, l'IA può di fatto risultare meno incline a generare distorsioni rispetto al processo decisionale

umano. Essa può infatti agevolare l'individuazione dei pregiudizi grazie alla sua capacità di raccogliere e analizzare ingenti quantità di dati» (*Council Conclusions on Advancing Gender Equality in the AI-Driven Digital Age: 6th horizontal review of the implementation of the Beijing Platform for Action by the Member States and the EU institutions,* Bruxelles, 19 giugno 2025).

Questo contributo vuole approfondire questo secondo aspetto, meno conosciuto, dell'IA, a partire dall'adozione di due raccomandazioni del 2026 nel quadro del Consiglio d'Europa e di recenti sviluppi in seno all'Unione europea. Rifletterà sul come l'IA possa essere utilizzata nella rilevazione dei contenuti illeciti online e nella valutazione del rischio e affronterà alcune delle complesse questioni etico-giuridico che il suo uso solleva.

2. L'analisi condotta nel presente contributo assume quale presupposto normativo le due raccomandazioni del Comitato dei Ministri del Consiglio d'Europa del 2 marzo 2026. Tali atti, non vincolanti, indirizzano la complessa convergenza tra l'evoluzione delle tecnologie digitali e la tutela sovranazionale dei diritti fondamentali, con specifico riguardo alla parità di genere, al divieto di discriminazione e alla prevenzione della violenza di genere contro le donne e le ragazze. La prima, la Raccomandazione CM/Rec(2026)1 sull'uguaglianza e l'intelligenza artificiale (IA), traccia le linee guida destinate agli Stati parte e agli attori chiave per l'integrazione del principio di eguaglianza, anche di genere, e per la prevenzione delle condotte discriminatorie in ciascuna fase di sviluppo, impiego e dismissione dei sistemi algoritmici. Il testo rimarca inoltre la centralità delle attività informative, della formazione specialistica e della vigilanza indipendente nei comparti ad alto impatto sociale. La seconda, la Raccomandazione CM/Rec(2026)2, si concentra sulla responsabilità per la violenza contro le donne e le ragazze facilitata dalle tecnologie della comunicazione e dell'informazione. Si fonda su standard esistenti, inclusa la Convenzione di Istanbul del Consiglio d'Europa per la prevenzione e la lotta contro la violenza nei confronti delle donne e la violenza domestica (Convenzione di Istanbul, su cui, diffusamente, v. S. DE VIDO, M. FRULLI (eds), *Preventing and Combating Violence against Women and Domestic Violence. A Commentary on the Istanbul Convention*, Cheltenham, 2023), che viene citata espressamente nel preambolo. Questa seconda Raccomandazione è stata poi diffusa ufficialmente il 10 giugno 2026 a Strasburgo.

Sebbene entrambe le Raccomandazioni del Consiglio d'Europa siano fortemente focalizzate sui rischi e sulla mitigazione dei danni causati dalla tecnologia e dagli algoritmi, è possibile rintracciare alcuni passaggi dei testi in cui viene valorizzato il potenziale positivo dell'intelligenza artificiale e della tecnologia per prevenire e contrastare la violenza di genere. Così, ad esempio, nella Raccomandazione CM/Rec (2026)1, è previsto un quadro organico (*comprehensive*) di misure volto a promuovere l'eguaglianza, ivi inclusa la parità di genere, nonché a prevenire e combattere le discriminazioni nel ciclo di vita dei sistemi di IA (*ivi*, par. 3). Gli algoritmi non devono solo evitare la discriminazione, ma possono essere progettati ed addestrati per promuovere attivamente la parità di genere e rimuovere gli svantaggi storici sofferti dalle donne (*ivi*, par. 4). Inoltre, si sofferma sulla progettazione orientata alla mitigazione del rischio: l'IA può essere utilizzata come uno strumento di diagnosi e monitoraggio: integrando criteri di equità sin dalla fase di progettazione, i sistemi algoritmici possono essere impiegati da autorità pubbliche ed enti privati per rilevare in modo tempestivo schemi di discriminazione, squilibri di genere o segnali di rischio prima che sfocino in abusi sistemici (*ivi*, parr. 4 e 5; par. 16). L'adozione di sistemi di monitoraggio e di valutazione dell'impatto supportati da tecnologie di IA consente alle istituzioni di raccogliere ed elaborare dati disaggregati per genere, migliorando la precisione e l'efficacia delle politiche pubbliche di prevenzione contro la violenza (*ivi*, par. 14).

La Raccomandazione CM/Rec(2026)2 incentiva lo sviluppo e l'attuazione di soluzioni tecnologiche di prevenzione da parte delle piattaforme digitali. L'IA gioca un ruolo chiave nel rilevamento tempestivo e nella rimozione automatizzata/assistita di contenuti lesivi. La Raccomandazione promuove altresì strumenti tecnologici a supporto delle vittime per facilitare la segnalazione degli abusi, la raccolta tempestiva di prove digitali (*ivi,* par. 29), la responsabilità delle piattaforme (*ivi*, par. 39), nonché la cooperazione con le agenzie di *law enforcement*, procuratori e il sistema giudiziario (*ivi,* par. 38.3). Infine, lo strumento di *soft law* evidenzia la necessità che le aziende tecnologiche collaborino con la società civile e le autorità giudiziarie mettendo a disposizione strumenti tecnologici avanzati per identificare i perpetratori di violenza digitale, interrompere le reti di diffusione illecita di materiale e prevenire la vittimizzazione secondaria.

3. Come si è detto, la seconda raccomandazione è basata (anche) sugli standard ricavati dalla Convenzione di Istanbul, adottata quindici anni fa ed entrata in vigore nel 2014, che tuttavia non fa espresso riferimento nel suo testo né alla violenza di genere nello spazio digitale, né al versante positivo dell'IA e al suo utilizzo per la prevenzione della violenza di genere e la tutela delle vittime. Ciononostante, taluni orientamenti in materia di prevenzione, supporto alle forze dell'ordine e rimozione dei contenuti lesivi tramite IA si possono rintracciare nella prima Raccomandazione Generale del GREVIO del 2021 — l'organo di monitoraggio della Convenzione di Istanbul — dedicata alla dimensione digitale della violenza, che ha invero riportato nel campo di applicazione dello strumento giuridico queste forme pervasive di violenza di genere. In tale occasione, il GREVIO ha raccomandato, tra l'altro di: promuovere l'inclusione dell'alfabetizzazione digitale e della sicurezza online nei programmi scolastici ufficiali e a tutti i livelli di istruzione; incoraggiare il settore delle TIC e gli intermediari internet, comprese le piattaforme di social media, a compiere uno sforzo attivo per evitare pregiudizi di genere nella progettazione di prodotti intelligenti, applicazioni per dispositivi mobili e videogiochi, nonché nello sviluppo dell'intelligenza artificiale; e, rispettivamente, a creare meccanismi di monitoraggio interno volti a garantire l'inclusione di prospettive incentrate sulle vittime, promuovendo al contempo una maggiore consapevolezza della prospettiva e delle esperienze delle utenti, in particolare di quelle esposte a forme intersezionali di discriminazione o a rischio di subirle (*ivi*, par. 51). La Raccomandazione Generale ha inoltre suggerito agli intermediari, inclusi i fornitori di servizi internet, i motori di ricerca e le piattaforme di social media, di mettere a disposizione istruzioni d'uso facilmente accessibili per segnalare i contenuti lesivi e richiederne la rimozione (*ivi*, par. 53). Per quanto concerne la violenza digitale, una delle raccomandazioni contenute nel testo ha evidenziato la necessità di «intensificare gli sforzi di rafforzamento delle competenze (*capacity-building*) per i professionisti della giustizia penale e delle forze dell'ordine, al fine di dotarli delle perizie e delle risorse necessarie su come applicare i quadri giuridici esistenti per affrontare la dimensione digitale della violenza contro le donne, nonché di sviluppare le loro capacità medico-legali e informatico-forensi nella raccolta e nella messa in sicurezza delle prove elettroniche, evitando di causare la vittimizzazione secondaria e la ri-traumatizzazione della persona offesa» (*ivi,* par. 55(2)).

4. Spostando l'attenzione sul diritto dell'UE, il principale e recente sviluppo in materia è rappresentato dalla Direttiva (UE) 1385/2024 sulla lotta alla violenza contro le donne e alla violenza domestica (Direttiva VAW). La Direttiva VAW ha per basi giuridiche gli articoli 82(2) e 83(1) del Trattato sul Funzionamento dell'Unione europea (TFUE). La Direttiva

VAW, che criminalizza alcuni comportamenti, segnatamente *cyber stalking*, molestie digitali, diffusione non consensuale di immagini intime e manipolate, incitamento digitale della violenza e dell'odio sulla base del genere (su cui si veda diffusamente A. DI STASI, R. ESPINOSA CALABUIG (a cura di), *Cyberviolenza di genere e nuove "frontiere" normative e giurisprudenziali: la direttiva UE 2024/1385*, Napoli, 2025) contiene diverse disposizioni utili per comprendere come l'IA possa essere impiegata per prevenire la violenza di genere e la violenza domestica e per proteggere le vittime (ad es. applicazioni basate sull'IA rivolte alle vittime), ma anche per condurre le indagini. Inoltre, anche nell'ambito della prevenzione, è opportuno segnalare il ruolo che l'alfabetizzazione digitale può avere nell'utilizzo di app e *chatbot*, che non sono necessariamente di immediata e facile fruizione. Così, ad esempio, ai sensi dell'articolo 14(1) Direttiva VAW, «gli Stati membri provvedono affinché la vittima possa segnalare gli atti di violenza contro le donne o di violenza domestica alle autorità competenti attraverso canali accessibili, di facile utilizzo e prontamente disponibili», i quali devono includere, «almeno per i reati informatici» rientranti nell'ambito di applicazione dell'atto di diritto derivato, «la possibilità di segnalare online o tramite altre TIC accessibili e sicure, fatte salve le norme procedurali nazionali relative alla formalizzazione della segnalazione online». L'articolo impone inoltre agli Stati membri di garantire «la possibilità di segnalare online o tramite altre TIC accessibili e sicure includa la possibilità di presentare elementi di prova». In merito alle indagini, la Direttiva VAW obbliga gli Stati membri a garantire che «le persone, le unità o i servizi incaricati dell'indagine e dell'azione penale per i reati di violenza contro le donne o di violenza domestica dispongano di competenze sufficienti in materia e di efficaci strumenti investigativi per indagare e perseguire efficacemente detti atti, in particolare per raccogliere, analizzare e procurarsi prove elettroniche nei casi di criminalità online» (articolo 15(1)). Riguardo alla facoltà (gli Stati «possono») per gli Stati membri di adottare linee guida destinate alle autorità competenti nei procedimenti penali per i casi di violenza contro le donne o violenza domestica, comprese direttive per la magistratura inquirente, l'articolo 21 stabilisce taluni requisiti, tra cui indicazioni su come (b) «raccogliere e conservare le prove pertinenti, comprese le prove online». L'articolo 23, *Misure di rimozione di determinato materiale online*, prevede la possibilità per le autorità competenti di emettere ordini vincolanti per rimuovere o disabilitare l'accesso a materiale accessibile al pubblico online di cui all'articolo 5(1), lettere a) e b), e agli articoli 7 e 8 della Direttiva VAW, ma non fa riferimento al rilevamento di tale materiale. L'articolo 29 VAW obbliga gli Stati a garantire linee telefoniche di assistenza valide su tutto il territorio nazionale, attive 24 ore su 24 e sette giorni su sette, per fornire informazioni e consulenza alle vittime, e li incoraggia a fornire le linee telefoniche di assistenza «anche attraverso altre TIC sicure e accessibili, comprese le applicazioni online». In termini di prevenzione, l'articolo 34(8) stabilisce che «gli Stati membri provvedono affinché tali misure preventive includano lo sviluppo di competenze di alfabetizzazione digitale, tra cui competenze critiche del mondo digitale e il pensiero critico, per permettere agli utenti di individuare e affrontare i casi di violenza online, cercare assistenza e prevenire detta violenza».

Altri strumenti giuridici sono rilevanti nell'analizzare il ruolo che l'IA può svolgere nella prevenzione della violenza di genere contro le donne e domestica: il Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale (*IA Act*); il Regolamento (UE) 2022/2065 del Parlamento europeo e del Consiglio, del 19 ottobre 2022, relativo a un mercato unico dei servizi digitali (*Digital Services Act* - DSA); il Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con

riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati (GDPR). Così, l'*AI Act* stabilisce, tra l'altro, norme armonizzate per l'immissione sul mercato, la messa in servizio e l'uso dei sistemi di IA nell'Unione; il divieto di talune pratiche di IA; requisiti specifici per i sistemi di IA ad alto rischio e obblighi per gli operatori di tali sistemi. Ai sensi del considerando n. 27 del Regolamento, per «diversità, non discriminazione ed equità» si intende che «i sistemi di IA sono sviluppati e utilizzati in modo da includere soggetti diversi e promuovere la parità di accesso, l'uguaglianza di genere e la diversità culturale, evitando nel contempo effetti discriminatori e pregiudizi ingiusti vietati dal diritto dell'Unione o nazionale». L'UE incoraggia un uso dell'IA fondato su tali principi. Nessuna disposizione specifica affronta tuttavia l'uso dell'IA per prevenire la violenza di genere e domestica, non essendo l'atto finalizzato a introdurre regole per l'IA ritenuta a rischio minimo o nullo. Inoltre, nel Regolamento, il riferimento all'impatto negativo dell'IA sulle donne è limitato al preambolo. Quest'ultimo riconosce che i sistemi di IA possono comportare rischi per i diritti fondamentali, inclusa la parità di genere, e che «se progettati e utilizzati in modo inadeguato, tali sistemi possono essere particolarmente intrusivi e violare il diritto all"istruzione e alla formazione, nonché il diritto alla non discriminazione, e perpetuare modelli storici di discriminazione, ad esempio nei confronti delle donne, di talune fasce di età, delle persone con disabilità o delle persone aventi determinate origini razziali o etniche o un determinato orientamento sessuale» (considerando n. 56). Tuttavia, nonostante questo strumento possa avere implicazioni positive nel contrasto alla violenza informatica basata sul genere, manca un riconoscimento esplicito del fatto che i sistemi di IA possano essi stessi costituire forme di violenza di genere che colpiscono in modo sproporzionato donne e ragazze (ad es. i *deepfake*), così come manca il riconoscimento del raccordo tra l'*AI Act* e la Direttiva VAW (sul punto si veda S. DE VIDO, *EU Law in Light of the Istanbul Convention: Legal Implications after Accession*, EU Publications Office Luxembourg, 2025). Il *Digital Services Act* pone al centro la sicurezza e i diritti degli utenti, creando un ambiente online più sicuro. Tra i suoi obiettivi, stabilisce norme su specifici obblighi di diligenza mirati per talune categorie di fornitori di servizi intermediari. Pertanto, la rilevazione di contenuti illeciti rientra nella valutazione dei rischi che i fornitori di piattaforme online di grandissime dimensioni devono effettuare. Ai sensi dell'Articolo 34, «i fornitori di piattaforme online di grandissime dimensioni e di motori di ricerca online di grandissime dimensioni individuano, analizzano e valutano con diligenza tutti i rischi sistemici nell'Unione derivanti dalla progettazione o dal funzionamento del loro servizio e dei relativi sistemi, compresi i sistemi algoritmici, o dall'uso fatto dei loro servizi». Utilizzando un'interpretazione sistematica che tenga conto dell'ordinamento giuridico dell'UE, si può sostenere che le piattaforme debbano essere in grado di individuare i rischi di violenza di genere presenti nei loro sistemi, compresi quelli algoritmici. Infine, il GDPR è cruciale parlando di app, *chatbot* e raccolta di prove tramite sistemi digitali. Ad esempio, ai sensi dell'Articolo 5 GDPR, i dati devono essere raccolti per finalità determinate, esplicite e legittime, e non successivamente trattati in modo incompatibile con tali finalità (si veda, *inter alia*, G.M. RUOTOLO, *Intelligenza Artificiale e trattamento dei dati tra diritto internazionale e ordinamenti interni*, in *AI Law*, 2/2025).

5. Ci si deve ora chiedere in che ambiti l'IA possa essere utile per la prevenzione e la risposta tempestiva alla violenza di genere contro le donne. Con riferimento al rilevamento dei contenuti illeciti online, vi sono state indagini sui metodi computazionali per identificare e affrontare i danni compiuti da comportamenti che avvengono nel mondo digitale (si veda ad esempio A. ARORA ET AL., *Detecting Harmful Content on Online Platforms: What Platforms Need*

vs. Where Research Efforts Go, in *ACM Computing Surveys*, 56/2023, Issue 3, p. 1 ss.). Tuttavia, quanto è facile individuare contenuti illeciti in rete? Esistono sistemi in grado di automatizzare e potenziare i processi di moderazione dei contenuti su vasta scala? L'IA, in altri termini, può sostituire l'essere umano nell'individuazione dei rischi e degli abusi potenziali? In uno studio condotto sui Termini e Condizioni di 42 piattaforme online, gli autori e le autrici hanno riscontrato che «la violenza, i contenuti espliciti e i discorsi d'odio sono gli argomenti più menzionati, a dimostrazione della notevole attenzione ad essi riservata nonché dell'importanza di individuare le violazioni delle clausole relative a tali temi. La sollecitazione sessuale e la propaganda politica sono gli argomenti con il minor numero di menzioni» (*ibid.*) L'abuso sessuale su minori registra una frequenza di citazioni analoga a quella della disinformazione, del bullismo, delle molestie e dello spam, pur essendo contemplato da un numero inferiore di piattaforme. Nello studio, il termine «violenza» è utilizzato in modo molto generico, quindi può certamente includere la violenza di genere, ma anche altre forme di violenza non collegate al genere.

Una possibilità per rilevare i contenuti illeciti online consiste nel consentire agli/alle utenti (come fanno molte piattaforme) di segnalare manualmente ai fornitori di servizi i contenuti inappropriati e (potenzialmente) illegali per la loro rimozione (si veda, in riferimento alla normativa europea, G. MORGESE, *Moderazione e rimozione dei contenuti illegali online nel diritto dell'UE*, in *Federalismi.it*, 1/2022, p. 80 ss.). Tuttavia, ciò può causare un ulteriore trauma alle vittime responsabili della segnalazione, oltre al timore di ritorsioni che potrebbe scoraggiare l'uso stesso della segnalazione. Uno dei metodi chiave per superare la segnalazione volontaria consiste nel valutare se i contenuti caricati corrispondano a contenuti problematici già noti. L'*hashing* percettivo — l'uso di un algoritmo di tracciamento che produce un estratto, una chiave di *hash* o un'impronta digitale di varie forme di contenuti multimediali — rileva i contenuti illegali (S. HAWKES ET AL, *Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools*, in IEEE Security & Privacy Magazine, 2024, p. 1 ss.) Come spiega Somers (C. SOMERS, *Ensuring Online Safety - The Role of Artificial Intelligence in Combatting Illegal Content Online*, blogpost, 2023):

> Gli algoritmi di *hashing* sono utilizzati per estrarre da un'immagine una firma alfanumerica sintetica e univoca, definita anche *hash* o impronta digitale dell'immagine. Tale impronta digitale appena generata può quindi essere confrontata con le impronte digitali di altre foto caricate su un servizio online. Ciò rende possibile cercare attivamente contenuti illegali noti e adoperarsi per rimuoverli. Ad esempio, Microsoft utilizza PhotoDNA sui propri servizi per scansionare le immagini illegali. Sebbene l'*hashing* sia una tecnica rapida e affidabile, il suo limite principale risiede nel fatto che può essere utilizzata solo per ricercare contenuti già noti. Non è una tecnica efficace per identificare contenuti nuovi e sconosciuti.

L'IA, in altri termini, risulta efficace se fondata su solidi basi di dati sottostanti. Questo risulta problematico in alcuni ambiti del web. I membri di ampie e oscure comunità online ascrivibili alla cosiddetta "Manosfera", ad esempio, accolgono, diffondono e coltivano percezioni e idee false, stereotipate, dannose e misogine sulle donne, spesso incentrate sulla loro natura, sulla sessualità, sull'autonomia corporea e sull'autodeterminazione (M. HOLM, *The manosphere and anti-gender backlash on social media,* in Z. LEFKOFRIDI (ed.), *Elgar Encyclopedia of Gender and Politics,* Cheltenham, 2026, p. 424 ss.). Vengono utilizzati codici e *meme* che non corrispondono necessariamente alle banche dati esistenti e che possono apparire innocui a prima vista. Inoltre, un meccanismo di scansione dei dati noto come *client-side scanning* è stato di recente proposto da aziende tecnologiche e governi come soluzione per rilevare contenuti illegali nelle comunicazioni con crittografia *end-to-end* (si veda qui G. RALTON, *Proposed*

mechanisms to detect illegal content can be easily evaded, study finds, blog post, 2022) Il *client-side scanning* si riferisce a «sistemi che scansionano i contenuti dei messaggi, quali immagini, testi e video, alla ricerca di corrispondenze con una banca dati di contenuti illegali precedentemente noti, prima che il messaggio venga crittografato e inviato al destinatario previsto» (*ibid.*)

6. L'IA può inoltre essere utile come strumento di analisi e valutazione del rischio. La valutazione del rischio è contemplata dall'articolo 51 della Convenzione di Istanbul e consiste in un processo decisionale attraverso il quale viene determinata la migliore linea d'azione stimando, identificando, qualificando e quantificando il rischio. La sua finalità, come riportato in uno studio EIGE del 2019, è quella di «ridurre il danno per le vittime femminili di violenza nelle relazioni intime e per i loro figli», identificando «tutti i livelli di rischio, ossia standard, medio ed elevato, nonché i bisogni specifici delle vittime» (EIGE, *Risk Assessment and Management of Intimate Partner Violence in the EU*, 2019, p. 19). A prima vista, sembrerebbe che solo il sistema della giustizia penale sia tenuto ad adoperare la valutazione del rischio nell'adozione di decisioni relative, ad esempio, a misure di protezione o ordini di allontanamento. Tuttavia, come sottolineato dal GREVIO nella sua *Mid-term Horizontal Review of GREVIO Baseline Evaluation Reports* (2022), «tutti i professionisti interessati, non solo le forze dell'ordine, sono tenuti a valutare e ad adottare misure per gestire i rischi per la sicurezza di una specifica vittima caso per caso, compreso il rischio di violenze reiterate e letali e, se necessario, a fornire un supporto coordinato» (*ivi*, par. 451). Taluni strumenti di valutazione del rischio sono stati elaborati per tutti i primi operatori (*first responders*), come l'*Ontario Domestic Assault Risk Assessment* (ODARA), utilizzabile da una vasta gamma di operatori, tra cui: operatori dei centri antiviolenza, addetti ai servizi di assistenza alle vittime, avvocati, operatori sanitari e assistenti sociali. La condivisione dei dati e la cooperazione tra forze dell'ordine, strutture sanitarie e servizi sociali costituiscono elementi essenziali per un'efficace strategia preventiva. Una «valutazione individuale» per individuare le esigenze di protezione delle vittime è prevista inoltre dall'articolo 16 della Direttiva VAW, «in aggiunta ai requisiti per la valutazione individuale di cui all'articolo 22 della direttiva 2012/29/UE», e deve essere effettuata nella fase più precoce possibile, «ossia al momento del primo contatto con le autorità competenti o non appena possibile dopo il primo contatto».

Un esempio di valutazione del rischio è il Sistema di Monitoraggio Integrato nei Casi di Violenza di Genere (*Sistema VioGén*), un'applicazione informatica creata dalla Segreteria di Stato per la Sicurezza (SES) del Ministero dell'Interno spagnolo, operativa in via continuativa in Spagna dal 26 luglio 2007 (J.L. GONZÁLEZ-ÁLVAREZ, J.J. LÓPEZ-OSSORIO, C. URRUELA, M. RODRÍGUEZ-DÍAZ, *Integral Monitoring System in Cases of Gender Violence. VioGén System*, in *Behavior & Law Journal*, 4(1)/2018, p. 29 ss.). Il sistema, tecnicamente un'applicazione web integrata nella Rete SARA spagnola, gestisce oltre 510.000 casi. Le informazioni raccolte comprendono dati sulla commissione di reati e infrazioni legati alla violenza di genere, precedenti penali e la situazione penitenziaria dei presunti autori in merito a permessi o scarcerazioni. Contiene inoltre dati identificativi, informazioni personali (filiazione, famiglia, situazione lavorativa, ecc.) e dati sull'assistenza e il supporto alle vittime (ad es. tipo di aiuto, uso di rifugi). L'accesso e l'uso sono riservati a specifici professionisti, tra cui membri dell'ordine giudiziario, del Pubblico Ministero, specialisti delle forze dell'ordine a tutti i livelli, amministrazioni penitenziarie, unità di coordinamento sulla violenza, unità di valutazione forense integrata, servizi assistenziali e servizi sociali degli enti locali. Agli utenti viene assegnato un profilo specifico con diversi livelli di autorizzazione per garantire la sicurezza e la conformità alle leggi sulla protezione dei dati. I casi sono attivi quando sono oggetto di

attenzione di polizia, inattivi quando non la richiedono più ma possono essere riattivati, e sono considerati a basso rischio (archiviati) solo in specifiche circostanze legali, come un'assoluzione irrevocabile o una condanna definitiva eseguita. Il Sistema *VioGén* risulta utile per il coordinamento e lo scambio di informazioni, la valutazione del rischio, il monitoraggio e la protezione, la prevenzione e gli allarmi. L'algoritmo stima e predice il rischio che le vittime corrano il pericolo di recidiva in ciascun caso segnalato. Il sistema impiega un modello probabilistico di valutazione del rischio composto da 37 variabili, quali la gravità dei precedenti atti violenti, la presenza e la natura delle minacce formulate dall'aggressore, i precedenti di violenza dell'aggressore e il grado di vulnerabilità della donna e degli eventuali minori coinvolti. L'algoritmo del sistema analizza i dati attraverso regole predefinite e criteri di valutazione del rischio per assegnare uno specifico livello di rischio, che può variare da «nessun rischio percepito» fino a «rischio estremo». Il livello di protezione di polizia offerto alla vittima viene definito in base alla categoria di rischio assegnata. Le valutazioni del rischio possono anche essere modificate e adeguate manualmente dagli agenti, se necessario. A seconda del livello di rischio, il protocollo specifica le misure di protezione di polizia necessarie per incrementare la sicurezza della vittima, le quali aumentano di intensità al crescere del livello di rischio. Ad esempio, nel caso di aggressori recidivi, un livello elevato di rischio di recidiva comporta ulteriori misure di protocollo, tra cui una maggiore sorveglianza, il monitoraggio dell'aggressore e comunicazioni alle autorità giudiziarie e alle vittime stesse. Il ruolo degli agenti di polizia rimane cruciale in questo tipo di valutazione del rischio: la loro esperienza nella gestione dei casi di violenza contro le donne consente loro di interpretare il contesto della situazione e di adeguare i livelli di valutazione del rischio quando necessario, integrandovi ciò che il sistema algoritmico potrebbe omettere. Di conseguenza, il sistema funge da prezioso strumento di supporto, ma non si sostituisce al lavoro di agenti di polizia adeguatamente formati.

7. Da ultimo, occorre interrogarsi sull'uso dell'IA da un punto di vista etico-giuridico. In materia di trattamento dei dati personali, il GDPR costituisce lo strumento giuridico vigente di riferimento; pertanto, le app e gli strumenti di valutazione del rischio devono conformarsi a tali standard. Così, ad esempio, in *VioGén* gli utenti accedono al sistema con un nome utente e una chiave personale e intrasferibile che consente l'audit delle loro attività, ma che presenta limitazioni, sia in termini di informazioni accessibili, sia di funzionalità attivabili. In secondo luogo, gli algoritmi possono anche riprodurre stereotipi, pregiudizi e falsi miti profondamente radicati. Lo stesso *VioGén* è stato criticato per il fatto di basarsi su algoritmi che hanno rafforzato gli stereotipi di genere e i pregiudizi esistenti, in particolare nella valutazione delle vittime e dei carnefici, oltre per la mancanza di trasparenza (v. D. MORONDO TARAMUNDI, *Prevenzione versus predizione del rischio: l'uso del programma semi-automatizzato di decisione VioGén in Spagna*, in *Diritto & Questioni Pubbliche*, XXV/2025, p. 69 ss.; si veda anche A. KARAGIANNI, *The EU Artificial Intelligence Act through a Gender Lens*, Bonn, 2025, p. 4). I criteri di valutazione del rischio potrebbero non cogliere la complessità dei singoli casi, «portando a classificazioni errate ed esiti potenzialmente dannosi per le vittime/sopravvissute, in particolare se donne» (*ibid.*). La formazione sull'uso del sistema *VioGén* si rivela dunque essenziale, ma «non può sostituire una formazione sufficiente e adeguata sulla violenza di genere» (D. MORONDO TARAMUNDI, *op.cit.*, p. 79). In terzo luogo, vi sono potenziali insidie e questioni etiche collegate all'uso della tecnologia, come le app per dispositivi mobili e i *chatbot*, per supportare le persone sopravvissute a violenze domestiche e *stalking*. Come evidenziato in un progetto dedicato proprio all'uso di questi sistemi,

chiamato ISEDA (un *chatbot* ideato per sostenere le donne vittime di abusi e violenza domestica, su cui si veda K. BUTTERBY, N. LOMBARD, *Developing a chatbot to support victim-survivors who are subjected to domestic abuse: considerations and ethical dilemmas,* in *Journal of Gender-Based Violence*, 9(1)/2025, p. 153 ss.), si riscontrano: a) problemi di sicurezza e protezione: le persone sopravvissute temono di essere tracciate dall'aggressore, di vedere i propri dati personali esposti o non protetti e di non essere in grado di cancellare completamente i dati dal proprio dispositivo; b) problemi di accessibilità: non tutte le vittime possono accedere alla tecnologia digitale a causa di vincoli socio-economici o se appartengono a gruppi minoritari come persone neurodivergenti, persone sorde, sopravvissute anziane o con difficoltà di alfabetizzazione, in quanto le tecnologie potrebbero non essere fruibili nelle loro forme attuali; c) mancanza di emulazione umana: la tecnologia, come nel caso dei *chatbot*, fatica ad emulare caratteristiche umane reali quali l'empatia, la capacità di leggere il linguaggio del corpo o mantenere il contatto visivo. Si suggerisce che la tecnologia debba fungere da canale d'accesso ai servizi umani e non da sostituto del contatto umano; d) problemi di manutenzione: la tecnologia stessa potrebbe non essere adeguatamente aggiornata, il che può condurre a problemi quali informazioni obsolete o collegamenti ipertestuali interrotti, come riscontrato in una valutazione delle applicazioni esistenti in materia di violenza domestica.
In quarto luogo, vi sono considerazioni etiche legate alle disuguaglianze strutturali e intersezionali presenti nelle nostre società (su IA e diritti umani, in particolare il principio di non discriminazione, si veda, *inter alia*, A. ODDENINO, *Intelligenza artificiale e tutela dei diritti fondamentali: alcune notazioni critiche sulla recente proposta di Regolamento della UE, con particolare riferimento all'approccio basato sul rischio e al pericolo di discriminazione algoritmica*, in A. PAJNO, F. DONATI, A. PERRUCCI (a cura di), *Intelligenza artificiale e diritto: una rivoluzione. Diritti fondamentali, dati personali e regolazione,* Bologna, p. 135 ss.; G.M. RUOTOLO, A. STIANO (a cura di), *Discriminazioni automatiche. Come il diritto può aiutare l'intelligenza artificiale e viceversa*, Torino, 2026; G. PIZZOLANTE, *Regolamento sull'intelligenza artificiale, metodi di gestione del rischio e tutela dei diritti umani*, in *Questa Rivista*, 2026, p. 235 ss.). Per accedere ad app e *chatbot* è necessaria un'alfabetizzazione digitale unitamente al possesso di un dispositivo aggiornato; fattori quali la povertà, le condizioni sociali, la violenza economica e lo status di migrante possono ostacolarne l'accesso, esacerbando ulteriormente radicati sistemi di oppressione. Gli sviluppatori devono essere consapevoli del fatto che funzionalità come i servizi di localizzazione, spesso utilizzati per trovare supporto a livello locale, possono essere sfruttati dagli aggressori per tracciare la vittima. Infine, sussistono considerazioni etiche relative allo status della vittima. In situazioni di pericolo imminente, un *chatbot* non è uno strumento idoneo, in quanto dovrebbe essere programmato per indirizzare immediatamente l'utente ai servizi di emergenza. La tecnologia è in grado di fare questo? Inoltre, nell'elaborazione di un *chatbot* o di una app, il linguaggio adoperato dovrebbe essere attento alle questioni di genere e sensibile ai traumi (*trauma-informed*), e sviluppato in collaborazione con professioniste/i dei servizi di supporto alle donne (E. DE GRAW, *Using AI to Track Gender-Based Violence,* in *Think Global Health Blog*, 2025).

8. L'analisi condotta evidenzia la natura intrinsecamente ambivalente dell'IA: se da un lato costituisce un veicolo di riproduzione di stereotipi e nuove forme di sopraffazione (dalle violazioni via *deepfake* all'abuso sessuale basato su immagini), dall'altro offre strumenti tecnologici inediti per la prevenzione, la rimozione dei contenuti lesivi e la valutazione del rischio di recidiva. Dal punto di vista giuridico, l'evoluzione normativa a livello regionale – dalle Raccomandazioni del Consiglio d'Europa CM/Rec(2026)1 e CM/Rec(2026)2 alla

Direttiva UE 2024/1385 – dimostra un progressivo adeguamento alla dimensione digitale della violenza. Tuttavia, strumenti come l'*hashing* percettivo o i modelli predittivi alla base del sistema *VioGén* mostrano evidenti limiti strutturali: non riescono a intercettare linguaggi d'odio di nuova formazione (come i codici della *Manosfera*) e rischiano di riprodurre i medesimi *bias* di genere che dovrebbero contrastare. Dall'analisi discende una conclusione forse ovvia, ma da tenere bene in mente quando si auspicano riforme giuridiche che incentivano l'uso dell'IA (sul confronto e dialogo tra fatto e norma, si veda G. DELLA MORTE (a cura di), *L'intelligenza artificiale tra fatto e norma*, Milano, 2026): l'intelligenza artificiale non può sostituire completamente il fattore umano, ma va intesa come ausilio, supporto integrativo, che richiede, anche per il suo utilizzo, una adeguata formazione, sensibile al genere e al trauma, degli attori che entrano in gioco nelle azioni di prevenzione e contrasto della violenza di genere nei confronti delle donne.

SARA DE VIDO