Papers di DIRITTO EUROPEO

www.papersdidirittoeuropeo.eu
ISSN 2038-0461

2026, n. 1

Papers di diritto europeo

www.papersdidirittoeuropeo.eu

I contributi sono sottoposti ad un procedimento di revisione tra pari a doppio cieco (*double-blind peer review*).
Non sono sottoposti a referaggio esclusivamente i contributi di professori emeriti, di professori ordinari in quiescenza e di giudici di giurisdizioni superiori e internazionali.

Papers di diritto europeo
ISSN 2038-0461
Registrazione al Tribunale di Verona n. 1875 del 22 luglio 2010

Papers di diritto europeo
www.papersdidirittoeuropeo.eu

Fascicolo 2026, n. 1

INDICE

Violenza algoritmica di genere, tecnologie predittive e diritto internazionale

Agostina Latino*

SOMMARIO: 1. Introduzione: quando l'algoritmo diventa carnefice istituzionale. – 2. La cornice di diritto internazionale: fondamenti strutturali e sfide tecnologiche. – 3. Le forme mutevoli della violenza tecnologica di genere: una tassonomia giuridica. – 4. Sistemi europei di gestione integrata del rischio di violenza di genere a confronto: VioGén, SARA, DASH e Codice Rosso. – 5. Standard esigibile di conformità algoritmica e principio di *non-refoulement* algoritmico nella protezione dalla violenza di genere. – 6. Conclusioni critiche e prospettive *de lege ferenda*.

1. Introduzione: quando l'algoritmo diventa carnefice istituzionale.

L'intersezione tra violenza di genere e tecnologie digitali si delinea come uno dei nodi più intricati e irrisolti del diritto internazionale contemporaneo[1], dove l'ipervelocità dell'innovazione algoritmica, incarnata in strumenti predittivi come i sistemi di valutazione del rischio di recidiva in contesti di violenza domestica, oltre ad amplificare le manifestazioni tradizionali di discriminazione strutturale contro le donne, ne genera di radicalmente nuove, ibride e transnazionali, tali da eludere i paradigmi normativi consolidati e da rivelare, in tutta la loro nudità, le aporie di un ordinamento giuridico ancora intrinsecamente antropocentrico e pre-digitale[2]. In un contesto in cui l'intelligenza artificiale (IA) permea sempre più le *policy* securitarie, dalla profilazione predittiva dei potenziali aggressori al monitoraggio geolocalizzato delle vittime, emerge un paradosso strutturale che questo saggio intende analizzare criticamente, ossia che le tecnologie impiegate per prevenire la violenza di genere possono contribuire a riprodurla, in quanto i sistemi predittivi apprendono da dati storici provenienti da archivi istituzionali (dati di polizia, registri dei servizi sociali, archivi delle procure) già segnati da *bias* e disuguaglianze, che vengono così incorporati e perpetuati. Il paradosso emerge dunque come nodo cruciale, all'esito, da un lato, della promessa di efficienza oggettiva e *data-driven*, capace di ridurre i tassi di violenza domestica e femminicidi attraverso l'elaborazione massiva di variabili statistiche come ad esempio la frequenza di denunce pregresse, le vulnerabilità socio-economiche e i *pattern* comportamentali degli autori; e,

* Professoressa aggregata di Diritto internazionale, Università degli Studi di Camerino.

[1] F. POGGI, *Digital Gender Violence in International Law*, in *European Journal of International Law*, 2024, pp. 567-596; D. K. CITRON, *Hate Crimes in Cyberspace*, Cambridge, 2014, pp. 145-167.

[2] R. BROWNSWORD, *Law*, *Technology and Society: Reimagining the Regulatory Environment*, London, 2019, pp. 3-28.

dall'altro, del rischio sistemico di perpetuazione di *bias* incorporati nei *dataset* di *training*, prevalentemente maschili e storicamente distorti da stereotipi di genere, che culminano in sottostime del pericolo per categorie marginali (in particolare, donne senza figli, donne migranti, donne anziane), in opacità *black-box* dei processi decisionali e nell'erosione del giudizio umano empatico quale canone essenziale per cogliere le sfumature contestuali di abusi invisibili all'occhio algoritmico. In effetti, come dimostra la giurisprudenza della Corte Europea dei diritti dell'uomo (Corte EDU o Corte di Strasburgo) a partire da *Opuz* c. *Turkey*[3], la sottovalutazione istituzionale del rischio può essere un fenomeno strutturale che vìola gli artt. 2 (diritto alla vita) 3 (divieto di tortura e di trattamenti inumani e degradanti) e 14 (divieto di discriminazione) della Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali (CEDU)[4]. La sentenza ha infatti chiarito che l'obbligo di protezione è autonomo rispetto alla volontà della vittima e che l'inerzia istituzionale, anche quando mascherata da formalismi procedurali, produce discriminazione di genere. Questo nucleo argomentativo anticipa molte delle criticità che i sistemi predittivi contemporanei riproducono in forma automatizzata. Emerge dunque una dicotomia, sintomo di una più profonda crisi epistemologica del diritto internazionale pubblico: come conciliare l'obbligo di *due diligence* statale, sancito dalla Convenzione sull'eliminazione di ogni forma di discriminazione nei confronti della donna (*Convention on the Elimination of All Forms of Discrimination against Women*-CEDAW)[5] e dalla Convenzione del Consiglio

[3] Corte europea dei diritti dell'uomo, sentenza del 9 giugno 2009, ricorso n. 33401/02, *Opuz* c. *Turkey*, punto 116. Il caso nasce da sette anni di violenze documentate che Nahide Opuz e sua madre avevano ripetutamente denunciato alle autorità turche. Nonostante l'*escalation* (aggressioni fisiche, minacce di morte, tentativi di investirla con l'auto) le autorità avevano reagito con inerzia sistematica, archiviando i procedimenti e non adottando misure protettive. Nel 2002 il marito uccise la madre di Nahide. La Corte EDU ha ritenuto che lo Stato sapesse o avrebbe dovuto sapere del rischio reale e immediato e non aveva adottato misure ragionevoli per prevenirlo. In questo caso la Corte di Strasburgo ha applicato il c.d. Osman Test, secondo cui le autorità vengono meno all'obbligo positivo di proteggere il diritto alla vita della donna quando le autorità sono a conoscenza delle minacce e queste sono concrete e immediate. Cfr. V. IVONE, *The European Convention on Human Rights and the Case Law of the Strasbourg Court on Domestic Violence*, in V. IVONE, S. NEGRI (a cura di), *Domestic Violence against Women*, Milano, 2019, pp. 90-93.

[4] P. LEACH, *Taking a Case to the European Court of Human Rights*, Oxford, 2017, pp. 120-135.

[5] Nazioni Unite, Convenzione sull'eliminazione di ogni forma di discriminazione nei confronti della donna (*Convention on the Elimination of All Forms of Discrimination against Women-CEDAW*), adottata il 18 dicembre 1979, entrata in vigore il 3 settembre 1981, reperibile *online*. In dottrina, *ex plurimis*, si vedano i contributi di: D. ŠIMONIVIĆ, *Global and Regional Standards on Violence Against Women: The Evolution and Synergy of the CEDAW and Istanbul Conventions*, in *Human Rights Quarterly*, 2014, pp. 590-606; N. A. ENGLEHART, M. K. MILLER, *The CEDAW effect: International law's impact on women's rights*, in *Journal of Human Rights*, 2014, pp. 22-47; A. HELLUM, H. SINDING AASEN, *Women's Human Rights. CEDAW in International, Regional and National Law*, New York, 2013; B. RUDOLF, A. ERIKSSON, *Women's Rights under International Human Rights Treaties. Issues on Rape, Domestic Slavery, Abortion and Domestic Violence*, in *International Journal of Constitutional Law*, 2007, pp. 507-525; L. MURA, *I diritti delle donne e la tutela della diversità nel diritto internazionale*, in *Rivista internazionale dei diritti dell'uomo*, 2000, pp. 47-83; K. TOMASEVSKI, *Women and Human Rights*, London, 1993; E. EVATT, *Eliminating discrimination Against Women: The Impact of the UN Convention*, in *Melbourne University Law Review*, 1991, pp. 435-449; N. BURROWS, *The 1979 Convention on the Elimination of All Forms of*

d'Europa sulla prevenzione e la lotta contro la violenza nei confronti delle donne e la violenza domestica (Convenzione di Istanbul)[6], con l'automazione delegata a privati *big-tech*, i quali operano in un regime di *accountability* asimmetrica, immuni da sanzioni transnazionali proporzionate al loro impatto globale[7]? In effetti, la violenza digitale, dal *cyberstalking* virale al *revenge porn deepfake*, oltre a spostare l'arena del danno (da fisico a virtuale), ne scala l'impatto esponenziale grazie a meccanismi di raccomandazione e viralità *platform-dependent*, sfidando la frammentata stratificazione normativa internazionale e imponendo un ripensamento radicale dei principî di non-discriminazione, trasparenza e riparazione *victim-centered*.

Se l'algoritmo è dunque la nuova grammatica del potere, occorre interrogarsi su chi ne scriva le regole sintattiche e a quale scopo[8]. Questo saggio adotta una prospettiva di diritto internazionale femminista-critica[9], intendendo leggere il testo normativo attraverso la lente di ciò che esclude, di ciò che tace, di ciò che assume come neutro e che neutro non è. In effetti, né il codice binario, né i dati statistici sono neutri e dunque l'architettura della prevenzione algoritmica, se costruita su fondamenta epistemicamente distorte, produce l'apparenza scientifica della (in)giustizia. Questo contributo prende l'abbrivio da

---

Discrimination Against Women, in *Netherland International Law Journal*, 1985, pp. 419-460; C. TINKER, *Human Rights for Women: the UN Convention on the Elimination of All Forms of Discrimination against Women*, in *Human Rights Quarterly*, 1981, pp. 32-346.

[6] Convenzione sulla prevenzione e la lotta contro la violenza nei confronti delle donne e la violenza domestica, adottata a Istanbul dal Comitato dei Ministri del Consiglio d'Europa il 7 aprile 2011, entrata in vigore il 1° agosto 2014, la cui traduzione non ufficiale in italiano è reperibile *online*. In dottrina *ex plurimis*, si vedano i lavori di: S. DE VIDO, M. FRULLI, *Preventing and Combating Violence Against Women and Domestic Violence. A Commentary on the Istanbul Convention*, Cheltenham, 2023; A. LATINO, *Manifestazioni e considerazioni della violenza nei confronti delle donne alla luce della Convenzione di Istanbul*, in *Studi sulla questione criminale*, 2019, pp. 165-186; E. D'URSEL, *La Convention du Conseil de l'Europe sur la prévention et la lutte contre les violences à l'égard des femmes: une révolution silencieuse?*, in *Revue trimestrielle des droits de l'homme*, 2018, pp. 29-49, reperibile *online*; L. GAROFALO, *Alcune considerazioni sulle norme "self-executing" contenute nella Convenzione di Istanbul del 2011*, in *Ordine Internazionale e Diritti Umani*, 2018, pp. 536-543, reperibile *online*; F. POGGI, *Violenza di genere e Convenzione di Istanbul: un'analisi concettuale*, in *Diritti umani e diritto internazionale*, 2017, pp. 51-76; J. A. MCQUIGG RONAGH, *The Istanbul Convention, Domestic Violence and Human Right*s, London, 2017; L. PERONI, *Violence Against Migrant Women: The Istanbul Convention Through a Postcolonial Feminist Lens*, in *Feminist Legal Studies*, 2016, pp. 49-67; G. PASCALE, *L'entrata in vigore della Convenzione di Istanbul sulla prevenzione e la lotta contro la violenza nei confronti delle donne e la violenza domestica*, in *Osservatorio costituzionale*, 2014, pp. 1-12; A. DI STEFANO, *Violenza contro le donne e violenza domestica nella nuova Convenzione del Consiglio d'Europa*, in *Diritti umani e diritto internazionale*, 2012, reperibile *online*; A. VALENTINI, *Recenti sviluppi in seno al Consiglio d'Europa in tema di violenza contro le donne*, in *La Comunità internazionale*, 2012, pp. 77-98.

[7] F. BASSAN, *Digital Platforms and Global Law*, Cheltenham-Northampton, 2021, spec. pp. 133-144.

[8] M. KOSKENNIEMI, *From Apology to Utopia: The Structure of International Legal Argument*, Cambridge, 2005, spec. pp. 500-540.

[9] C. BIDDOLPH, J. KIM, S. PERRY, L. J. SHEPHERD, *Handbook on Gender and Violence*, Cheltenham, 2026; H. CHARLESWORTH, C. CHINKIN, *The Boundaries of International Law: A Feminist Analysis*, Manchester, 2000, spec. pp. 200-256; S. WALBY, *Theorizing Patriarchy*, Oxford, 1990; A. EDWARDS, *Male Violence in Feminist Theory: an Analysis of the Changing Conceptions of Sex/Gender Violence and Male Dominance*, in J. HANMER, M. MAYNARD (edited by), *Women, Violence and Social Control*, London, 1987, pp.13-29.

questa constatazione per sviluppare una tesi critica di diritto internazionale, ossia che i sistemi algoritmici di valutazione del rischio di recidiva nella violenza domestica non si limitano a presentare imperfezioni tecniche correggibili con meri aggiustamenti incrementali del codice[10], in quanto, viceversa, essi incarnano, in forma di previsione statistica, una struttura di discriminazione istituzionale che vìola gli obblighi degli Stati ai sensi del diritto internazionale dei diritti umani. Si tratteggia dunque una sorta di «violenza algoritmica di genere», così si propone di denominarla, che può essere proposta come categoria interpretativa emergente, la cui rilevanza normativa richiede un progressivo consolidamento nella prassi e nella giurisprudenza internazionale.

Sulla base di tali premesse, il presente saggio persegue tre obiettivi interconnessi.

Il primo è costruire un perimetro organico di diritto internazionale entro cui inquadrare giuridicamente le sfide poste dai sistemi predittivi di valutazione del rischio nella violenza domestica. In effetti, la letteratura esistente sul tema tende a biforcarsi in due direzioni che raramente si incontrano: da un lato, studi di diritto internazionale sulla violenza di genere che ignorano la dimensione tecnologica; dall'altro, studi di etica algoritmica e diritto dell'IA che ignorano il diritto internazionale. Questo contributo mira a (tentare di) colmare tale iato, argomentando che la valutazione dei sistemi predittivi, della loro legittimità, dei loro limiti, delle responsabilità che generano, non può essere condotta soltanto attraverso il prisma del diritto europeo dell'IA o del diritto penale nazionale, ma richiede di essere situata nel quadro degli obblighi internazionali degli Stati in materia di violenza di genere, di cui CEDAW e Convenzione di Istanbul costituiscono fonti primarie di riferimento.

Il secondo obiettivo è analizzare comparativamente il funzionamento dei principali sistemi predittivi europei (*in specie*, il *Sistema de seguimiento integral en los casos de Violencia de Género*-VioGén spagnolo[11], il metodo *Spousal Assault Risk Assessment*-SARA[12], il *Domestic Abuse, Stalking and Honour-Based Violence Risk Identification, Assessment and Management*-DASH britannico[13], nonché il Codice Rosso italiano[14] – che non è propriamente un sistema predittivo: cfr. *infra*) attraverso la lente degli obblighi internazionali così ricostruiti, al fine di mostrare che i loro problemi strutturali – dai *bias*

[10] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, 2016, spec. pp. 89-112.

[11] *Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género*, BOE-A-2004-23255. Cfr. il sito https://sistemaviogen.ses.mir.es/publico/viogen/, VioGén gestito dal Ministerio del Interior.

[12] P. R. Kropp, S. D. Hart, C. D. Webster, D. Eaves, *Manual for the Spousal Assault Risk Assessment Guide (SARA)*, Vancouver, 1994.

[13] Home Office-UK, *DASH Risk Identification, Assessment and Management*, London, 2009, aggiornato 2023. La *check-list* dei rischi DASH è reperibile *online*.

[14] Legge 19 luglio 2019, n. 69 (c.d. Codice Rosso). Il Codice Rosso viene qui esaminato non quale sistema predittivo, categoria alla quale non appartiene, ma quale modello procedurale alternativo che consente di valutare se un approccio fondato sulla tempestività istituzionale, anziché sulla previsione algoritmica del rischio, soddisfi in modo più o meno compiuto gli obblighi internazionali applicabili. La comparazione è dunque funzionale, non strutturale.

sistematici nei confronti delle donne migranti e delle donne senza figli all'opacità dei meccanismi di ponderazione, dal fenomeno dell'*over-reliance* degli operatori alle lacune nella copertura della violenza digitale – sono segnali di una tensione irrisolta tra la logica probabilistica dei modelli prognostici e le esigenze di un diritto internazionale che impone valutazioni autonome, non discriminatorie e *victim-centered* del rischio.

Il terzo obiettivo è sviluppare proposte *de lege ferenda* tecnicamente fondate e giuridicamente coerenti. La letteratura critica sull'IA tende infatti a concludersi con inviti generici alla «maggiore trasparenza» o alla «supervisione umana» che, pur condivisibili nella loro direzione, risultano privi di mordente normativo. Questo contributo intende procedere diversamente, identificando strumenti giuridici specifici, quali un protocollo opzionale alla CEDAW sulla violenza digitale, un obbligo di *gender audit* algoritmico con conseguenze giuridiche vincolanti e un regime di supervisione umana qualificata che vada al di là del requisito formale regolamento (UE) 2024/1689 (AI Act)[15], capaci di tradurre le critiche sistemiche in obblighi normativi azionabili.

Il percorso argomentativo prende le mosse dalla costruzione della cornice di diritto internazionale rilevante, analizzando gli obblighi derivanti dalla CEDAW, con particolare attenzione alle raccomandazioni generali nn. 19[16] e 35[17], dalla Convenzione di Istanbul, nonché dall'AI Act e dal regolamento UE 2016/679 (*GeneralData Protection Regulation*, c.d. GDPR)[18], dai Principi Guida ONU su imprese e diritti umani (UNGP, 2011)[19] e dalla Convenzione di Belém do Pará[20], mostrando come questi strumenti, pur nella loro

---

[15] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio (AI Act), che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n, 300/2008, (UE) n, 167/2013, (UE) n, 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828. In dottrina si vedano i contributi dello *special issue* curato da D.R. TROITIÑO, T. KERIKMÄE, *E-governance and the European Union: Agenda for implementation*, *Science Direct*, 2024, reperibile *online*.

[16] General recommendation n. 19 on violence against women, A/47/38 del Comitato CEDAW, (llth session, 1992).

[17] General recommendation n. 35 on gender-based violence against women, updating General Recommendation n. 19 del Comitato CEDAW, CEDAW/C/GC/35 del 26 luglio 2017. Per un commento cfr. S. DE VIDO, *The Prohibition of Violence Against Women as a Customary International Law? Remarks on the CEDAW General Recommendation No. 35*, in *Diritti umani e diritto internazionale*, 2018, pp. 379-396.

[18] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio**,** relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati).

[19] UN Office of the High Commissioner for Human Rights, guiding principles on business and human rights: implementing the United Nations «*protect, respect and remedy*», 2011, cfr. il sito https://digitallibrary.un.org/record/720245?v=pdf . In dottrina, cfr. J. KNOX, *Horizontal Human Rights Law*, in *American Journal of International Law*, 2008, pp. 1-47.

[20] Convenzione interamericana sulla prevenzione, la punizione e l'eradicazione della violenza contro le donne (c.d. Convenzione di Belém) adottata dall'Assemblea generale dell'Organizzazione degli Stati Americani a Belém do Pará, Brasile, il 9 giugno 1994 ed entrata in vigore il 5 marzo 1995, reperibile *online*. In dottrina, si vedano i lavori di: C.M. CARMONA, *La erradicación de la violencia contra la mujer «por tratado»: un análisis comparado del Convenio de Istanbul y de la Convención de Belém do Parà*, in *Revista Europea de Derechos Fundamentales*, 2017, pp. 213-239; C. MEDINA, *The American Convention on Human Rights: Crucial Rights and Their Theory and Practice*, Cambridge, 2016; M. KIMELBLATT,

frammentazione, convergano verso un sistema di obblighi positivi che impone requisiti precisi per la *governance* dei sistemi predittivi applicati alla violenza domestica[21]. Su questa base normativa si innesta una tassonomia giuridica delle forme di violenza tecnologica di genere che distingue tra violenza diretta (*i.e.*, *cyberstalking*, *doxxing*, produzione o diffusione di *deepfake* a contenuto sessuale), violenza mediata (intesa come amplificazione *platform-dependent* di forme preesistenti di violenza fisica o psicologica) e violenza algoritmica strutturale (ossia quella che il presente *paper* intende teorizzare, intendendola quale violenza incorporata nei meccanismi decisionali istituzionali). Tale classificazione è particolarmente importante in quanto produce ricadute operative sull'individuazione del soggetto responsabile, sulla determinazione della giurisdizione competente e sulla selezione degli strumenti di tutela attivabili nel caso concreto. Il quadro normativo e tassonomico così delineato costituisce la premessa dell'analisi comparativa dei quattro summenzionati sistemi europei (VioGén, SARA, DASH, Codice Rosso), condotta al fine di misurarne la compatibilità con gli obblighi internazionali ricostruiti nella prima parte e di identificare le tensioni strutturali che nessun aggiustamento tecnico incrementale è in grado di risolvere. L'analisi converge quindi sulle dimensioni giuridiche dei *bias* algoritmici, approfondendone i meccanismi causali, nonché sul problema del *refoulement* algoritmico per le donne migranti, categoria elaborata in questo contributo quale strumento analitico per descrivere una doppia vulnerabilità che il diritto internazionale non ha ancora tematizzato come autonoma fattispecie normativa[22]. Siffatta doppia vulnerabilità consiste nell'intersezione tra

Reducing Harmful Effects of Machismo Culture on Latin American Domestic Violence Laws: *Amending the Convention of Belém Do Pará to Resemble the Istanbul Convention*, in *The George Washington International Law Review*, 2016, pp. 405-439; L.P. MEJÍA GUERRERO, *La Comisión Interamericana de Mujeres y la Convención de Belém do Pará. Impacto en el Sistema Interamericano de Derechos*, in *Revista interamericana y europea de derechos humanos*, 2012, pp. 189-213.

[21] Sembra peraltro opportuno evidenziare come l'analisi si concentri sui sistemi europeo e interamericano di protezione dei diritti umani, che hanno elaborato la giurisprudenza e gli strumenti convenzionali più rilevanti ai fini della costruzione dello standard di conformità algoritmica proposto, non prendendo in esame il sistema africano, e in particolare il protocollo alla Carta africana dei diritti dell'uomo e dei popoli sui diritti delle donne in Africa (protocollo di Maputo, 2003), che all'art. 4 impone agli Stati obblighi di protezione dalla violenza di genere comparabili per struttura a quelli della Convenzione di Istanbul. Sul punto si veda A. LATINO, *Manifestazioni e considerazioni della violenza nei confronti delle donne*, cit., spec. p. 173 ss. In effetti, il protocollo non viene esaminato in questa sede non già per irrilevanza sistematica, bensì per ragioni di perimetro dell'indagine, già estesa e densa. Si rinvia dunque a una ricerca successiva che potrebbe verificare se e in quale misura la giurisprudenza della Commissione africana e della Corte africana dei diritti dell'uomo e dei popoli abbia elaborato obblighi positivi in materia di tecnologie predittive e violenza di genere, contribuendo a consolidare o a problematizzare lo standard internazionale qui proposto in chiave universale.

[22] Come si avrà modo di approfondire *infra*, la nozione di *non-refoulement algoritmico* viene impiegata nel presente contributo su un piano deliberatamente bivalente, che riflette la natura stessa delle categorie emergenti nel diritto internazionale. Sul piano analitico, essa opera come strumento euristico che consente di descrivere e rendere giuridicamente visibile una specifica forma di fallimento della protezione istituzionale, non ancora tematizzata come autonoma fattispecie normativa. Sul piano normativo, essa trova un ancoraggio progressivo nelle disposizioni positive che si esamineranno, senza tuttavia pretendere di costituirne un'applicazione diretta e consolidata. Tale bivalenza è consapevole e non casuale: la storia del

esposizione alla violenza di genere e marginalità istituzionale derivante dalla condizione migratoria, che si traduce, nei sistemi predittivi, in un deficit di rappresentazione nei dati e in una conseguente sottostima del rischio, con effetti di esclusione dalla protezione. Le conclusioni raccolgono le fila dell'intero percorso e articolano le proposte *de lege ferenda*, restituendo a ciascuna critica una concreta risposta giuridica.

2. La cornice di diritto internazionale: fondamenti strutturali e sfide tecnologiche.

La progressiva emersione della violenza di genere come questione centrale del diritto internazionale contemporaneo consente di cogliere la struttura degli obblighi che oggi gravano sugli Stati, nonché la tensione che attraversa l'intero sistema tra l'eguaglianza formale garantita dal diritto e la persistenza di diseguaglianze strutturali nella realtà sociale. Una delle conquiste più rilevanti di tale evoluzione è il superamento della dicotomia pubblico/privato, che aveva tradizionalmente escluso la violenza domestica dal perimetro della responsabilità statale, come evidenziato nella letteratura femminista di diritto internazionale[23], processo avviato in modo decisivo nel quadro del mandato del Relatore speciale sulla violenza contro le donne[24]. Questo superamento è stato reso possibile dall'applicazione, in tali àmbito, del paradigma della *due diligence*[25], in virtù del quale lo Stato è chiamato a rispondere, oltre che delle violazioni che commette direttamente, anche di quelle che non previene, non indaga e non sanziona, pur avendone tanto la possibilità quanto il dovere[26]. Nell'era dei sistemi predittivi, tale dottrina assume una dimensione inedita e, per certi versi, paradossale perché lo Stato che adotta un sistema algoritmico di valutazione del rischio per adempiere al proprio obbligo di prevenzione si dota di uno strumento istituzionale che può tanto rafforzare quanto indebolire la sua capacità di adempiere tale obbligo. In effetti, quando il sistema sottostima il rischio per determinate categorie di donne, lo Stato non soddisfa l'obbligo di *due diligence* e, omettendo di intervenire pur essendo in condizione di farlo, produce, attraverso una vera

diritto internazionale dei diritti umani mostra che le categorie più feconde, dal *non-refoulement* classico alla *due diligence* in materia di violenza di genere, hanno percorso esattamente questo tragitto, dalla proposta dottrinale all'ancoraggio convenzionale, passando per il progressivo riconoscimento giurisprudenziale. Il presente contributo si colloca nella fase iniziale di questo percorso, con la consapevolezza che spetta alla prassi degli Stati, ai Comitati convenzionali e alle Corti internazionali determinare se e in quale misura la categoria proposta sia in grado di acquisire una propria forza normativa autonoma.

[23] C. CHINKIN, *A Critique of the Public/Private Dimension*, in *European Journal of International Law*, 1999, pp. 387-395, spec. pp. 389-392.

[24] R. COOMARASWAMY, *Report on Violence against Women, its Causes and Consequences*, E/CN.4/1999/68/Add.3, 1999, par. 72.

[25] V. STOYANOVA, *Due Diligence versus Positive Obligations*, in J. NIEMI, L. PERONI, V. STOYANOVA (edited by), *International Law and Violence Against Women: Europe and the Istanbul Convention*, London, 2000, pp. 95-112.

[26] A. OLLINO, *State obligations and due diligence*, in S. DE VIDO, M. FRULLI (edited by), *Preventing and Combating Violence*, Cheltenham, 2023, pp. 136-159; N. MAVRONICOLA, *Positive Obligations in the European Convention on Human Rights*», in *ECHR Studies*, 2021, pp. 45-72.

e propria torsione teleologica, un effetto per cui la tecnologia adottata si trasforma in un meccanismo di elusione dell'obbligo stesso, occultato dalla retorica dell'innovazione securitaria. Peraltro, l'obbligo di *due diligence* non si esaurisce nell'adozione di misure protettive, ma richiede altresì che tali misure siano concretamente accessibili, non discriminatorie e adeguate a garantire la tutela effettiva di tutte le categorie di donne che il diritto internazionale impone di proteggere; tale esigenza si fa ancor più stringente nei contesti in cui l'impiego di sistemi algoritmici rischia di riprodurre o addirittura amplificare le medesime asimmetrie strutturali che la *due diligence* è chiamata a prevenire e correggere.

L'analisi degli strumenti *ad hoc* elaborati nell'ordinamento internazionale a tutela delle donne deve dunque prendere avvio dalla CEDAW[27], che all'art. 1 definisce la discriminazione come «*any distinction, exclusion or restriction made on the basis of sex which has the effect or purpose of impairing or nullifying the recognition, enjoyment or exercise by women […] of human rights and fundamental freedoms*». La scelta di affiancare il criterio dell'effetto a quello dello scopo – scelta consapevole e giuridicamente significativa, come attestano i *travaux préparatoires*[28] – costituisce il fondamento normativo per qualificare i *bias* algoritmici come violazioni della CEDAW. Il criterio dell'effetto opera indipendentemente dall'intenzione soggettiva; ne consegue che un sistema predittivo che sottostimi sistematicamente il rischio per le donne migranti può integrare una violazione dell'art. 1 CEDAW non in ragione di una (consapevole) finalità discriminatoria, bensì in quanto produce un effetto (oggettivamente) discriminatorio, riconducibile alla nozione consolidata di discriminazione indiretta elaborata dalla giurisprudenza internazionale e dai Comitati di monitoraggio. L'estensione di tale categoria ai sistemi algoritmici istituzionali rimane tuttavia ancora poco esplorata in dottrina, nonostante le sue rilevanti implicazioni: l'art. 1 CEDAW impone infatti agli Stati non solo di astenersi da pratiche discriminatorie intenzionali, ma anche di esercitare un controllo attivo sui propri strumenti istituzionali, inclusi quelli algoritmici, al fine di prevenire la produzione di effetti discriminatori strutturali, secondo una logica di *accountability* rafforzata centrata sull'impatto, piuttosto che sull'intenzione. La Corte EDU ha ulteriormente precisato, in *Eremia* c. *Moldova*[29], che l'inerzia

[27] S. NEGRI, *Violence against Women and Domestic Violence in International Law*, in V. IVONE, S. NEGRI (edited by), *Domestic Violence*, cit., pp. 1-24.

[28] R. COOK, S. CUSACK, *Gender Stereotyping: Transnational Legal Perspectives*, Philadelphia, 2011, spec. pp. 7-33; L. A. REHOF, *Guide to Travaux Préparatoires of the United Nations Convention on the Elimination of All Forms of Discrimination against Women*, Dordrecht, 1993.

[29] Corte europea dei diritti dell'uomo, sentenza del 28 maggio 2013, ricorso n. 3564/11, *Eremia* c. *Republic of Moldova.* La sig.ra Eremia subiva violenze dal marito, un agente di polizia, che continuava a minacciarla anche dopo l'emissione di un ordine restrittivo. Le autorità moldave non solo non applicarono l'ordine, ma incoraggiarono la donna a «riconciliarsi» con il coniuge, malgrado anche le figlie avessero assistito ripetutamente alle aggressioni, in quanto lei non era «*not the first nor the last woman to [have been] beaten up by her husband*» (sic!). La Corte EDU ha qualificato tale inerzia come discriminazione istituzionale, poiché rifletteva un atteggiamento sistemico di minimizzazione della violenza domestica.

istituzionale spesso non è il frutto di valutazioni isolate, bensì la manifestazione di una discriminazione strutturale implicita. Tale impostazione rafforza l'idea che anche i sistemi predittivi che producono effetti sistematicamente sfavorevoli per determinate categorie di donne possano configurare una violazione autonoma dell'art. 1 CEDU, indipendentemente dall'intenzione soggettiva degli operatori.

Un secondo elemento decisivo dell'art. 1 è la portata universale della definizione, che non distingue tra sfera pubblica e privata, consentendo di ricondurre la violenza domestica nell'àmbito della responsabilità internazionale dello Stato, secondo un'impostazione ormai consolidata nella dottrina dei diritti umani[30]. Ciò implica che la violenza domestica, inclusa quella che i sistemi algoritmici contribuiscono a perpetuare attraverso opacità e *bias*, rientra pienamente nell'àmbito di applicazione della Convenzione. A ciò si aggiunge il riferimento al godimento effettivo dei diritti, che consente di qualificare come discriminatorie le differenze di trattamento che impediscano alle donne di esercitare concretamente i propri diritti fondamentali, sicché, in tale prospettiva, un sistema predittivo che ostacoli l'accesso effettivo delle vittime migranti alle misure di protezione previste dalla legge, non tanto perché la legge non le garantisca, ma perché l'algoritmo le classifica sistematicamente a rischio insufficiente per attivarle, incide direttamente sul nucleo essenziale dell'art. 1 CEDAW, traducendo in forma applicativa su base statistica una discriminazione strutturale che la Convenzione mira a eliminare.

L'art. 5 CEDAW impone poi agli Stati di «*to modify the social and cultural patterns of conduct of men and women*» al fine di eliminare pregiudizi e pratiche basate sull'idea dell'inferiorità o superiorità di un sesso sull'altro. Questa disposizione rappresenta la dimensione trasformativa della Convenzione che non si limita a proteggere le donne nel sistema esistente, ma impone di rinnovare il sistema che le discrimina. Tale dimensione è cruciale per comprendere la violenza algoritmica nella sua componente epistemologica[31]; in quanto i *bias* nei *dataset* di *training* riflettono i modelli culturali storicamente dominanti nelle istituzioni che hanno prodotto quei dati[32], in specie la tendenza delle forze dell'ordine a minimizzare la violenza domestica, la maggiore propensione dei servizi sociali a intervenire quando vi sono minori, la sfiducia sistemica verso le denunce delle donne straniere. Tutti questi modelli discriminatori, che l'art. 5 impone di eliminare, vengono trascritti nei *dataset* e da lì trasferiti negli algoritmi predittivi, che finiscono per replicare il comportamento istituzionale del passato

[30] V. M. MERINO-SANCHO, *Definitions*, in S. DE VIDO, M. FRULLI (edited by), *Preventing and Combating Violence*, Cheltenham, 2023, pp. 108-134; R. COPELON, *Intimate terror: Understanding domestic violence as torture*, in *Human rights of women: National and international perspectives*, 1994, pp. 116-152.

[31] V. EUBANKS, *Automating Inequality: How High-Tech Tools Profile*, *Police, and Punish the Poor,* New York, 2018, pp. 1-28.

[32] A. BENJAMIN, *Race after Technology*, Cambridge, 2019, spec. pp. 1-32.

conferendogli una patina di oggettività scientifica. L'art. 5 può dunque essere letto come fondamento normativo dell'obbligo degli Stati di garantire che i sistemi algoritmici applicati alla violenza di genere non riproducano, in forma di previsione statistica, i modelli discriminatori che la Convenzione intende sradicare. Detto in altri termini, addestrare un algoritmo su dati istituzionali storicamente distorti, oltre a essere un'opzione tecnica quantomeno discutibile, è una scelta politica che cristallizza e perpetua modelli patriarcali in violazione dell'art. 5.

Il Comitato CEDAW ha progressivamente colmato il silenzio della Convenzione rispetto alla violenza di genere nell'era digitale attraverso la Raccomandazione Generale n. 19 del 1992, successivamente aggiornata dalla n. 35 del 2017[33]. Quest'ultima, adottata venticinque anni dopo in un contesto tecnologico profondamente trasformato, rappresenta un passaggio fondamentale nell'evoluzione degli obblighi convenzionali. Da un lato, essa riconosce espressamente che la violenza di genere può essere perpetrata o facilitata mediante le tecnologie dell'informazione e della comunicazione (par. 20), consentendo di ricondurre anche le forme di violenza algoritmica nell'àmbito di applicazione della Convenzione. Dall'altro, rafforza il contenuto della *due diligence* statale rispetto all'azione di soggetti privati, includendo le imprese tecnologiche tra gli attori rispetto ai quali operano obblighi di vigilanza e prevenzione (par. 24(*b*)). Particolarmente significativo, ai fini della presente analisi, è inoltre il par. 30(*b*)(iii), che richiede agli Stati di fondare le misure preventive su un'analisi sistematica delle cause e delle conseguenze della violenza di genere, escludendo il ricorso a stereotipi relativi alla capacità delle donne di valutare e gestire il proprio rischio. Tale indicazione assume una rilevanza immediata rispetto ai sistemi predittivi, poiché implica l'incompatibilità con gli standard convenzionali di modelli che incorporano e riproducono stereotipi istituzionali, come la sistematica sottovalutazione del rischio per le donne migranti o per le donne prive di figli minori.

La Raccomandazione n. 35 riconosce inoltre che alcune donne sperimentano forme intersezionali di discriminazione (par. 12) e che necessitano di misure specificamente adattate alla loro situazione. Tale prospettiva rivela che i *bias* documentati nei sistemi predittivi sono la manifestazione visibile di un problema strutturale più ampio: le donne con disabilità, le donne anziane in relazioni di lunga durata, le donne LGBTQ+ vittime di violenza del partner dello stesso sesso, sono sistematicamente marginalizzate da algoritmi costruiti attorno al prototipo della vittima giovane, eterosessuale, con figli e cittadina del Paese di residenza. La mancanza di indicatori specifici per queste categorie è l'epitome di una scelta normativa implicita che stabilisce quali donne meritano protezione e quali no, in violazione del par. 12 della Raccomandazione n. 35, che impone

[33] Cfr. *supra* note 16 e 17.

agli Stati di garantire che le misure preventive raggiungano tutte le categorie di donne, *a fortiori* quelle esposte a discriminazioni multiple[34].

Nel panorama europeo, lo strumento internazionale più compiuto in materia di violenza di genere è sicuramente la Convenzione del Consiglio d'Europa sulla prevenzione e la lotta contro la violenza nei confronti delle donne e la violenza domestica sia per l'ampiezza del suo catalogo obbligatorio, sia, soprattutto, per la capacità di tradurre in un quadro normativo coerente l'evoluzione degli obblighi positivi degli Stati in chiave di prevenzione strutturale. La sua architettura, fondata sui quattro assi della prevenzione, protezione, repressione e delle politiche integrate[35], opera come una matrice sistemica che ricompone in un unico disegno le diverse dimensioni della *due diligence* convenzionale. In tale contesto, l'art. 51 assume una funzione di cerniera concettuale perché elabora un vero e proprio standard di razionalità pubblica nella gestione del rischio di violenza domestica. L'art. 51 codifica, infatti, per la prima volta in un trattato internazionale, un obbligo giuridico relativo alla valutazione strutturata del rischio, imponendo alle autorità competenti di procedere a un accertamento qualificato della letalità, della gravità della situazione e della probabilità di reiterazione della violenza. Tale obbligo incarna una forma di razionalità preventiva che vincola gli Stati a dotarsi di strumenti idonei a cogliere la dinamica evolutiva del rischio, evitando che la risposta istituzionale si riduca a un intervento *ex post*. In questo senso, l'art. 51 può essere letto come base funzionale per l'adozione di strumenti strutturati di valutazione del rischio, pur senza implicare necessariamente il ricorso a sistemi algoritmici automatizzati, delimitando rigorosamente il perimetro entro cui tali strumenti possono operare. L'obbligo convenzionale è, in effetti, un obbligo di valutazione, non già di delegazione algoritmica: detto in altri termini, la Convenzione non impone/autorizza la sostituzione dell'apprezzamento umano con un apparato automatizzato opaco, richiedendo, viceversa, che la valutazione sia svolta in modo informato, controllabile e conforme ai parametri di non discriminazione e proporzionalità.

In effetti, la Grande Camera della Corte EDU, nella sentenza *Kurt* c. *Austria*[36], ha chiarito che l'obbligo di valutazione del rischio non può essere adempiuto mediante

---

[34] K. CRENSHAW, *Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color*, in *Stanford Law Review*, 1991, pp. 1241-1299.

[35] A. LATINO, *Manifestazioni e considerazioni*, cit., p. 177 ss.

[36] Corte europea dei diritti dell'uomo (Grande Camera), sentenza del 15 giugno 2021, ricorso n. 62903/15, *Kurt* c. *Austria*. La sig.ra Kurt, vittima di violenza domestica da parte del marito, aveva denunciato ripetuti episodi di violenza fisica, sessuale e psicologica, ottenendo dalle autorità misure urgenti di allontanamento e un ordine di protezione. Alcune settimane dopo, il marito si recò nella scuola frequentata dai figli e uccise uno di essi con un colpo d'arma da fuoco, prima di togliersi la vita. Dinanzi alla Corte EDU, la ricorrente lamentava che le autorità austriache non avessero adottato misure sufficienti a prevenire l'evento, in particolare omettendo di procedere a una valutazione del rischio che considerasse adeguatamente la posizione dei figli e non disponendo misure restrittive ulteriori nei confronti dell'aggressore. La controversia si concentrava dunque sull'estensione degli obblighi positivi dello Stato rispetto alla prevedibilità del rischio per soggetti diversi dalla vittima diretta della violenza domestica.

strumenti statici o meramente formali, richiedendo una valutazione autonoma, immediata e costantemente aggiornata, indipendente dalla volontà della vittima e calibrata sulla dinamica concreta del caso. La Corte ha inoltre affermato che la supervisione umana qualificata costituisce una componente essenziale dell'obbligo positivo di protezione derivante dall'art. 2 CEDU. Particolarmente rilevanti, ai fini della presente analisi, sono tuttavia le opinioni dissenzienti, le quali, pur condividendo l'impianto teorico del test di *due diligence* elaborato dalla maggioranza, ne contestano l'applicazione al caso concreto. Secondo i giudici dissenzienti, la Corte ha adottato una lettura eccessivamente restrittiva del requisito della prevedibilità del rischio, finendo per attribuire un peso sproporzionato ai margini di discrezionalità delle autorità nazionali e per sottovalutare elementi che avrebbero dovuto essere considerati indicativi di una situazione di pericolo elevato. Le opinioni dissenzienti insistono in particolare sull'escalation progressiva della violenza, sulla reiterazione delle minacce, sulla disponibilità di informazioni relative alla pericolosità dell'aggressore e sulla necessità di valutare tali fattori in modo cumulativo, anziché isolato. La critica assume una portata che va oltre il caso concreto, poiché richiama l'attenzione sul rischio che procedure apparentemente corrette sul piano formale producano valutazioni sostanzialmente inadeguate quando non riescano a cogliere la complessità del contesto di violenza. In tale prospettiva, i giudici dissenzienti evidenziano come stereotipi di genere e forme di vittimizzazione secondaria possano incidere sul processo decisionale delle autorità, determinando una sottostima del rischio proprio nei confronti delle vittime maggiormente vulnerabili. Essi propongono pertanto un'applicazione del test di *due diligence* maggiormente sensibile al genere e alle vulnerabilità intersezionali, fondata su una lettura contestuale della violenza domestica e orientata a privilegiare la protezione effettiva della vittima rispetto a una valutazione astratta e frammentata degli indici di rischio[37].

In senso analogo, nel caso *Levchuk* c. *Ukraine*[38], la Corte EDU ha ritenuto incompatibile con gli obblighi positivi derivanti dall'art. 8 (rispetto della vita privata e familiare) una valutazione giudiziaria che, pur riconoscendo il contesto di violenza domestica, non aveva effettuato un'analisi sufficientemente approfondita del rischio

[37] S. DE VIDO, *Verso un test di dovuta diligenza sensibile al genere nei casi di violenza domestica? Sulla recente giurisprudenza della Corte europea dei diritti umani*, in *Diritti umani e diritto internazionale*, 2022, pp. 613-635.

[38] Corte europea dei diritti dell'uomo, sentenza del 3 dicembre 2020, ricorso n. 17496/19, *Levchuk* c. *Ukraine*. La vicenda oggetto del ricorso riguarda la sig.ra Levchuk, vittima di violenza domestica da parte dell'ex partner, con il quale continuava a condividere l'abitazione insieme ai figli. Nonostante i ripetuti episodi di violenza psicologica, minacce e comportamenti intimidatori, documentati anche attraverso interventi delle forze dell'ordine, le autorità ucraine avevano respinto la richiesta di allontanamento dell'uomo dall'abitazione familiare, attribuendo prevalenza al suo diritto di occupazione dell'immobile e senza procedere a una valutazione adeguata del rischio per la ricorrente e per i minori. Dinanzi alla Corte EDU, la ricorrente lamentava che tale decisione l'avesse esposta, insieme ai figli, a una situazione di violenza continuativa. La Corte ha accertato la violazione dell'art. 8 CEDU, ritenendo che le autorità non avessero effettuato una valutazione sufficientemente approfondita del rischio, né adottato misure idonee a garantire una protezione effettiva contro la violenza domestica.

futuro di violenza psicologica e fisica nei confronti della ricorrente e dei figli. Ne deriva che sistemi predittivi incapaci di cogliere l'evoluzione del rischio o che producono classificazioni standardizzate non soddisfano gli standard convenzionali.

Il combinato disposto degli artt. 4(3), 22 e 51 della Convenzione di Istanbul definisce poi, in chiave sistemica, i requisiti sostanziali della valutazione del rischio. L'art. 4(3) impone che essa sia non discriminatoria, in particolare rispetto all'origine nazionale, traducendo nel contesto della violenza di genere il principio generale di uguaglianza nell'accesso alla protezione. L'art. 22 richiede che i servizi di supporto siano accessibili a tutte le vittime, incluse quelle in condizioni di vulnerabilità strutturale, elemento che incide indirettamente anche sulla qualità e sull'effettività della valutazione del rischio. L'art. 51, infine, impone una valutazione strutturata del rischio idonea a orientare l'adozione di misure di protezione adeguate e proporzionate al livello di pericolo concreto.

Ne consegue che i sistemi predittivi affetti da *bias* algoritmici, in particolare quelli che sotto-rappresentano sistematicamente determinate categorie di donne nei *dataset* di addestramento o nei flussi informativi istituzionali, risultano incompatibili con tale triplice requisito convenzionale. Essi, infatti, non soddisfano i requisiti di non discriminazione, accessibilità sostanziale ed effettività, in quanto producono valutazioni distorte che compromettono in modo strutturale la capacità dello Stato di adempiere ai propri obblighi positivi rafforzati.

In tal modo, la Convenzione di Istanbul può essere letta come idonea a delimitare criticamente l'uso delle tecnologie predittive nella gestione del rischio, riaffermando che la prevenzione della violenza di genere non può essere affidata a dispositivi che riproducono o amplificano le asimmetrie strutturali già presenti nella società.

Va ora preso in considerazione il regolamento (UE) 2024/1689 (AI Act)[39], entrato in vigore il 2 agosto 2024 con applicazione progressiva, che introduce per i sistemi di IA ad alto rischio un regime giuridico e che oggi rappresenta uno dei modelli normativi più avanzati nel contesto comparato, pur inserendosi in un quadro globale ancora in evoluzione[40]. Ai sensi dell'art. 6, par. 2, in combinato disposto con l'Allegato III, punto 6, sono classificati ad alto rischio i sistemi destinati ad essere utilizzati dalle autorità di contrasto, tra l'altro, per determinare il rischio per una persona di diventare vittima di reati (lett. *a*) ovvero per determinare il rischio di commissione del reato o di recidiva in relazione a una persona (lett. *d*). Tale classificazione ricomprende pertanto gli strumenti predittivi impiegati nella valutazione del rischio di recidiva nella violenza domestica, collocandoli all'interno della categoria più stringente del quadro europeo di *governance* tecnologica. Va peraltro precisato che l'art. 6, par. 3, introduce un'eccezione applicativa:

[39] Cfr. *supra* nota 15.

[40] A. KARAGIANNI, *Gender in a stereo-(gender) typical EU AI law: A feminist reading of the AI act*, in *Cambridge Forum on AI: Law and Governance*, 2025, pp. 1-12.

un sistema rientrante nell'Allegato III non è considerato ad alto rischio qualora non presenti un rischio significativo di danno per i diritti fondamentali delle persone, ad esempio perché destinato a eseguire un compito procedurale limitato o a migliorare il risultato di un'attività umana già completata, ferma restando la regola secondo cui un sistema che effettui profilazione di persone fisiche è sempre considerato ad alto rischio. Tale clausola di esonero, nel contesto dei sistemi predittivi applicati alla violenza domestica, difficilmente troverà applicazione: strumenti come VioGén producono classificazioni individuali che incidono direttamente sull'attivazione o sul diniego di misure di protezione, configurando un impatto significativo sui diritti fondamentali che esclude strutturalmente il ricorso all'eccezione.

Sono tre i profili del regime delineato dall'AI Act che assumono un peculiare rilievo rispetto ai sistemi predittivi utilizzati nella violenza domestica.

Il primo concerne l'obbligo di supervisione umana di cui all'art. 14. La disposizione non si limita a richiedere la presenza formale di un operatore umano nel circuito decisionale, ma esige una capacità effettiva di comprendere il funzionamento e i limiti del sistema, individuare anomalie e, soprattutto, mantenere il potere di disapplicare l'*output* algoritmico quando il giudizio professionale lo imponga. Il considerando 60 del testo definitivo chiarisce che tale supervisione deve essere «reale ed effettiva», innestando tale requisito nel solco della tradizione europea della *accountability* e della non delegabilità delle funzioni pubbliche essenziali.

Il secondo profilo riguarda l'art. 13, che introduce un modello di trasparenza che opera in modo selettivo, privilegiando gli utenti istituzionali del sistema, *in primis* le forze dell'ordine, e lasciando invece in ombra le persone direttamente incise dagli *output* algoritmici, cioè, per quel che rileva nel nostro perimetro di indagine, le vittime di violenza domestica. Questa configurazione produce una asimmetria informativa strutturale in quanto chi utilizza il sistema per adottare decisioni dispone di informazioni sulla logica, sui limiti e sulle condizioni d'uso dello strumento, mentre chi subisce gli effetti di quelle decisioni resta privo degli elementi minimi necessari per comprenderne la portata, contestarne l'esito o chiedere una revisione fondata su criteri verificabili. In un settore come la violenza domestica, dove la protezione effettiva dipende dalla capacità della vittima di interagire con le istituzioni in condizioni di fiducia, comprensione e *agency*, tale asimmetria risulta difficilmente conciliabile con l'approccio *victim-centered* imposto dalla Convenzione di Istanbul. La Convenzione richiede infatti che le vittime siano poste in condizione di conoscere i presupposti delle decisioni che riguardano la loro sicurezza, di comprenderne la logica e di contestarne eventuali esiti inadeguati o discriminatori. L'assenza di trasparenza verso le vittime, oltre a rappresentare un deficit procedurale, è un elemento che incide sulla qualità della protezione, sulla possibilità di esercitare diritti procedurali effettivi e, in ultima analisi, sulla conformità del sistema agli obblighi internazionali in materia di prevenzione e protezione dalla violenza di genere.

Il terzo profilo di rilievo, assente nella versione provvisoria del regolamento e introdotto nel testo definitivo, è l'art. 27, che impone ai *deployer* pubblici e ai privati che erogano servizi pubblici una valutazione d'impatto sui diritti fondamentali (*Fundamental Rights Impact Assessment* - FRIA) prima di utilizzare un sistema di IA ad alto rischio di cui all'Allegato III. Nel contesto dei sistemi predittivi di valutazione del rischio nella violenza domestica, tale obbligo acquisisce una rilevanza specifica: la FRIA impone di identificare preventivamente i gruppi di persone che possono essere negativamente incisi dall'utilizzo del sistema, di valutare la probabilità e la gravità del rischio discriminatorio e di documentare le misure adottate per mitigarlo. Letto in combinato disposto con gli obblighi di *due diligence* derivanti dalla CEDAW e dalla Convenzione di Istanbul, l'art. 27 configura un presidio *ex ante* che, qualora applicato rigorosamente, potrebbe intercettare le criticità strutturali dei sistemi predittivi prima della loro messa in servizio, anziché rilevarle soltanto attraverso la giurisprudenza *ex post* delle Corti internazionali.

Ai fini dell'analisi rileva poi il GDPR[41] che continua a rappresentare l'asse portante della tutela dei diritti fondamentali nel trattamento dei dati personali, imponendo limiti sostanziali e procedurali che incidono direttamente sulla legittimità dei sistemi predittivi impiegati nella violenza domestica. L'art. 22 sancisce il diritto dell'interessato (nel caso di specie, la vittima) a non essere sottoposto a decisioni basate unicamente su trattamenti automatizzati che producano effetti giuridici significativi o incidano in modo analogo sulla sua sfera giuridica. Tale disposizione, trasposta nel contesto dei sistemi di valutazione del rischio, limita il ricorso a decisioni interamente automatizzate, consentendolo solo alle condizioni e nei limiti previsti dalla stessa norma.

L'art. 9, par. 1, del GDPR rafforza ulteriormente tale tutela vietando, in linea di principio, il trattamento di categorie particolari di dati, tra cui quelli relativi all'origine razziale o etnica, salvo la ricorrenza di specifiche basi giuridiche previste dal paragrafo 2. Tale disposizione assume un rilievo particolare nei sistemi predittivi che utilizzano variabili *proxy* di rischio fondate sulla nazionalità o su indicatori ad essa correlati, poiché tali *proxy*, lungi dall'essere neutrali, possono riprodurre e amplificare *bias* strutturali, generando valutazioni discriminatorie incompatibili con il principio di uguaglianza sostanziale. In questa prospettiva, il GDPR opera come un limite esterno alla progettazione stessa dei modelli predittivi, precludendo il ricorso a correlazioni statistiche che, pur funzionali alla *performance* tecnica, risultano giuridicamente inammissibili.

Completano il quadro gli artt. 13 e 14 del GDPR, imponendo obblighi di informazione che includono la comunicazione di informazioni significative sulla logica del trattamento automatizzato, nonché sulle conseguenze previste di tale trattamento. Tali disposizioni implicano che i criteri, le fonti di dati e le modalità attraverso cui il sistema

[41] Cfr. *supra*, nota 18.

elabora le proprie valutazioni siano resi conoscibili in termini comprensibili e, ove possibile, verificabili.

Alla luce di tali requisiti, nessuno dei sistemi predittivi attualmente in uso, siano essi algoritmici come VioGén o semi-strutturati come DASH e SARA, appare in grado di soddisfare pienamente tali obblighi, vuoi per l'opacità tecnica dei modelli, vuoi per l'assenza di documentazione adeguata, vuoi per la natura stessa dei processi decisionali incorporati. Detto in altri termini, ne deriva che il GDPR non si limita a disciplinare il trattamento dei dati, ma ridefinisce i confini di legittimità dei sistemi predittivi, imponendo che essi siano costruiti e utilizzati in modo tale da garantire trasparenza, non discriminazione e controllo umano effettivo, condizioni che, allo stato, nessuno dei modelli esistenti è in grado di assicurare in modo pienamente conforme.

Il quadro normativo relativo al trattamento automatizzato dei dati personali va completato con un richiamo alla Convenzione del Consiglio d'Europa per la protezione delle persone rispetto al trattamento automatizzato di dati a carattere personale, nella versione modernizzata dal protocollo CETS n. 223, adottato il 18 maggio 2018 (c.d. Convenzione 108+)[42]. Tale strumento assume una rilevanza specifica nel contesto dei sistemi predittivi applicati alla violenza domestica per almeno tre ragioni convergenti.

In primo luogo, la Convenzione 108+ è l'unico strumento internazionale giuridicamente vincolante in materia di protezione dei dati personali aperto all'adesione di Stati non membri del Consiglio d'Europa, configurandosi come strumento di proiezione universale del modello europeo di tutela: nella prospettiva del presente contributo, ciò implica che i suoi standard possono trovare applicazione anche avuto riguardo a Stati che impiegano sistemi predittivi di valutazione del rischio al di fuori dello spazio giuridico europeo, rilevando ai fini della costruzione dello standard internazionale di conformità algoritmica di cui al par. 5 del presente contributo.

In secondo luogo, a differenza del GDPR, la Convenzione 108+ si applica esplicitamente ai trattamenti di dati personali effettuati da autorità di polizia e organi giudiziari, senza le deroghe settoriali che caratterizzano invece la direttiva (UE) 2016/680 (c.d. direttiva *Law Enforcement*)[43]. Tale copertura universale assume rilievo diretto nel contesto dei sistemi predittivi impiegati dalle forze dell'ordine – come VioGén – per i quali il GDPR opera in modo incompleto, lasciando spazi che la Convenzione 108+ è invece strutturalmente idonea a colmare. Segnatamente, l'art. 9 della Convenzione 108+,

[42] Il 18 maggio del 2018 il Consiglio ha adottato un protocollo di modifica (detto anche Convenzione 108+) del testo della Convenzione sulla protezione delle persone rispetto al trattamento automatizzato di dati a carattere personale (STE n. 108) adottata a Strasburgo il 28 gennaio 1981, entrata in vigore il 1° ottobre 1985 all'esito del deposito del quinto strumento di ratifica.

[43] Direttiva (UE) 2016/680 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativa alla protezione delle persone fisiche con riguardo al trattamento dei dati personali da parte delle autorità competenti a fini di prevenzione, indagine, accertamento e perseguimento di reati o esecuzione di sanzioni penali, nonché alla libera circolazione di tali dati e che abroga la decisione quadro 2008/977/GAI del Consiglio.

che disciplina le deroghe ai principî fondamentali di protezione e le condizioni in cui esse possono essere ammesse in materia di sicurezza nazionale e ordine pubblico, subordina qualsiasi eccezione al rispetto di una serie di garanzie cumulative, tra cui il controllo da parte di un'autorità indipendente, il rispetto del principio di proporzionalità e la non discriminazione. Tali garanzie, lette in chiave sistematica, rafforzano gli obblighi già derivanti dal GDPR e dall'AI Act, precludendo che le autorità competenti possano invocare ragioni di sicurezza pubblica per sottrarre i sistemi predittivi di violenza domestica a un controllo indipendente e trasparente.

In terzo luogo, la Convenzione 108+ introduce, nella sua versione modernizzata, un diritto dell'interessato a non essere soggetto a decisioni puramente automatizzate aventi effetti giuridici significativi, diritto che si affianca a quello previsto dall'art. 22 GDPR ampliandone la portata geografica, nonché un diritto a conoscere la logica del trattamento automatizzato. Tali disposizioni, applicate al contesto dei sistemi predittivi di valutazione del rischio, impongono che la vittima di violenza domestica sia posta nella condizione, a monte, di comprendere i criteri attraverso cui è stata effettuata la sua classificazione e, a valle, di contestarla efficacemente, rafforzando così, sul piano della protezione dei dati personali, le garanzie procedurali già richieste dalla Convenzione di Istanbul in relazione all'accesso effettivo alla tutela.

La Convenzione 108+ si configura dunque come strumento complementare e non ridondante rispetto al GDPR: mentre quest'ultimo opera come standard di diritto derivato dell'Unione, vincolante negli Stati membri ma non proiettabile direttamente verso gli ordinamenti terzi, la Convenzione 108+ fornisce la dimensione internazionale e universalizzante dello stesso nucleo di tutela, contribuendo a delineare uno standard globale di protezione dei dati applicabile anche ai sistemi predittivi adottati da Stati che non ricadono nel campo di applicazione del diritto europeo.

Di fatto, nel loro insieme, l'AI Act e il GDPR, pur appartenendo al diritto dell'Unione europea, possono essere considerati come standard regionali rilevanti ai fini dell'interpretazione evolutiva degli obblighi internazionali degli Stati, dando luogo a un doppio circuito di garanzia che opera su piani distinti ma complementari e che produce un effetto di rafforzamento reciproco, in un quadro di crescente interazione tra strumenti normativi europei in materia di tecnologia e diritti fondamentali[44]. Tale doppio livello di tutela, che rappresenta oggi una delle architetture normative più avanzate nella regolazione dei sistemi predittivi, presenta tuttavia zone d'ombra e tensioni irrisolte che incidono sulla protezione effettiva delle vittime di violenza di genere. In effetti, da un lato, l'AI Act istituisce un regime di *governance* tecnologica orientato alla prevenzione dei rischi sistemici e discriminatori, dall'altro, il GDPR assicura un presidio dei diritti fondamentali volto a limitare automatismi decisionali incompatibili con la dignità umana

[44] G. SARTOR, F. LAGIOIA, *The Impact of the AI Regulation on the DSA Framework*, in *European Journal of Law and Technology*, 2023, pp. 1-21.

e con il principio di uguaglianza. La combinazione dei due strumenti contribuisce così a delineare uno standard europeo idoneo a esercitare una significativa influenza sulla regolazione globale dei sistemi predittivi e a ridefinire i confini della responsabilità statale nell'adozione di tecnologie incidenti sulla protezione delle vittime.

Tuttavia, proprio l'interazione tra i due regimi fa emergere criticità strutturali. In primo luogo, la coesistenza di obblighi tecnici (AI Act) e obblighi sui diritti fondamentali (GDPR) non sempre genera un quadro coerente, in quanto i requisiti di trasparenza, supervisione umana e qualità dei dati previsti dall'AI Act possono entrare in contrasto con i limiti posti dal GDPR al trattamento di categorie particolari di dati, soprattutto quando i sistemi predittivi si fondano su *proxy* sensibili o su correlazioni statistiche difficili da giustificare giuridicamente. In secondo luogo, la capacità degli Stati di garantire un controllo umano «significativo» rischia di essere compromessa dalla complessità tecnica dei sistemi, che rende difficile per gli operatori comprendere la logica sottostante agli *output* e, quindi, esercitare un vaglio critico effettivo. Infine, la proiezione internazionale dello standard europeo, pur rappresentando un elemento di forza, può tradursi in una asimmetria applicativa, sicché gli Stati che adottano sistemi predittivi, senza disporre delle infrastrutture tecniche e istituzionali necessarie, rischiano di implementare modelli solo formalmente conformi, ma sostanzialmente incapaci di garantire le tutele richieste.

In questo senso, l'AI Act e il GDPR, pur definendo un modello avanzato di regolazione, rivelano anche le fragilità di un sistema che, pur ambizioso, deve ancora confrontarsi con la complessità dei contesti operativi, con la natura opaca di molti strumenti predittivi e con la necessità di assicurare che la tecnologia non diventi un ulteriore fattore di vulnerabilità per le donne che dovrebbe proteggere.

Il perimetro normativo va, infine, completato prendendo in esame i Principi Guida ONU su imprese e diritti umani[45] e la Convenzione di Belém do Pará[46] che introducono due ulteriori assi normativi che, pur provenendo da tradizioni giuridiche differenti, convergono nel delineare obblighi stringenti per gli attori pubblici e privati coinvolti nello sviluppo e nell'uso dei sistemi predittivi nella violenza di genere.

I Principi Guida ONU su imprese e diritti umani, strumento di *soft law*, delineano un dovere di *human rights due diligence* in capo alle imprese, che non si esaurisce nella valutazione *ex ante* dei rischi, ma che, viceversa, richiede l'identificazione sistematica degli impatti negativi, inclusi quelli derivanti da *bias* discriminatori, la valutazione dei loro effetti sulle persone vulnerabili e un monitoraggio continuo lungo l'intero ciclo di vita del sistema. In tale prospettiva, le imprese che progettano o forniscono strumenti predittivi, inclusi quelli utilizzati da autorità pubbliche, non possono invocare la neutralità tecnica dei modelli, ma sono tenute a dimostrare di aver adottato misure adeguate per

[45] Cfr. *supra*, nota 19.
[46] Cfr. *supra*, nota 20.

prevenire e mitigare gli effetti discriminatori e per gestire i rischi residui attraverso strumenti proporzionati e verificabili.

La Convenzione di Belém do Pará del 1994, primo strumento internazionale a prevedere un meccanismo di petizione individuale specifico per la violenza di genere, ha alimentato una giurisprudenza della Corte interamericana dei diritti umani (Corte Interamericana de Derechos Humanos - Corte IDH) che ha progressivamente ampliato la portata degli obblighi positivi degli Stati[47], , nel quadro più ampio dello sviluppo dei diritti delle donne nel sistema interamericano[48], anche in relazione all'impatto delle tecnologie digitali sulla violenza di genere[49]. Essa impone a tutti i funzionari pubblici, indipendentemente dal settore di appartenenza, un obbligo di conformità agli standard di prevenzione, protezione e indagine che si applicano direttamente anche all'uso dei sistemi predittivi. Ciò significa che l'adozione di strumenti come VioGén, DASH o SARA non può essere considerata una scelta tecnica, astrattamente «neutra», proprio perché ogni funzionario che li utilizza è responsabile della loro conformità agli standard interamericani, inclusi quelli relativi alla non discriminazione, alla protezione effettiva e alla prevenzione della ri-vittimizzazione.

Questi cinque strumenti, rispettivamente, CEDAW, Convenzione di Istanbul, AI Act/GDPR, UNGP e Convenzione di Belém do Pará, non costituiscono un sistema normativo coerente né auto-applicativo, ma formano, piuttosto, un mosaico frammentato[50], caratterizzato da differenze di natura giuridica, àmbito geografico e meccanismi di attuazione, come evidenziato anche nell'analisi comparata dei sistemi regionali di tutela dei diritti umani[51]. La loro integrazione richiede un'operazione interpretativa sistematica che tenga insieme, a un tempo, obblighi di diritto internazionale dei diritti umani, standard regionali, norme europee sulla tecnologia e responsabilità delle imprese. Solo attraverso questa ricomposizione è possibile costruire un quadro di *accountability* multilivello capace di orientare l'uso dei sistemi predittivi verso la finalità condivisa che attraversa tutti gli strumenti, ossia quella proteggere le donne dalla violenza di genere in modo non discriminatorio, trasparente e rispettoso della loro autonomia e delle loro sfaccettate peculiarità.

---

[47] M. ANTKOWIAK, *An Emerging Mandate for International Courts*, in *Virginia Journal of International Law*, 2011, pp. 737-760, spec. 750-755.

[48] J. CELORIO, *The Rights of Women in the Inter-American System of Human Rights*, in *University of Miami Law Review*, 2011, pp. 1192-1211.

[49] A. HEYNS, *Technology and Violence against Women in the Inter-American System*, in *International Journal of Human Rights*, 2023, pp. 567-586.

[50] N. AYALON, *Towards a Global Treaty on Digital Gender Violence*, in *American Journal of International Law Unbound*, 2025, pp. 45-50; M. KOSKENNIEMI, P. LEINO, *Fragmentation of International Law?*, in *Leiden Journal of International Law*, 2002, pp. 553-579.

[51] D. CABAL, *Comparative Analysis of Regional Human Rights Systems*, in *International Journal of Human Rights*, 2022, pp. 345-362.

3. Le forme mutevoli della violenza tecnologica di genere: una tassonomia giuridica.

La classificazione giuridica della violenza tecnologica di genere costituisce un percorso che, muovendo dalla base normativa internazionale ed europea, tenta di dare ordine a un fenomeno che si presenta in forme mutevoli, ibride, spesso sfuggenti alle categorie tradizionali del diritto. La violenza digitale è infatti un eterogeno insieme di pratiche, architetture e decisioni istituzionali che si trasformano e si sovrappongono, e che richiedono, per poter essere comprese e regolate, una classificazione capace di coglierne la natura giuridica, all'esito della specifica manifestazione tecnologica. In effetti, la violenza digitale comprende un insieme poliforme di condotte rese possibili, facilitate o amplificate dalle tecnologie digitali, caratterizzate da elementi strutturali che ne accrescono la pervasività, per il tramite di sovraesposizioni relazionali, viralità, persistenza del contenuto, anonimato dell'autore e estrema difficoltà di rimozione definitiva[52]. Questi tratti, che trasformano il cyberspazio in un ambiente particolarmente favorevole alla cronicizzazione della vittimizzazione, sono stati riconosciuti anche a livello sovranazionale dal *Group of experts on action against violence against women and domestic violence* (GREVIO)[53], che qualifica la dimensione digitale della violenza contro le donne come un insieme di condotte commesse *online* o facilitate dalla tecnologia, spesso trascurate o insufficientemente tipizzate nelle normative interne.

In questo contesto si colloca la direttiva (UE) 2024/1385[54] sulla lotta alla violenza contro le donne e alla violenza domestica, che rappresenta il primo strumento di diritto derivato dell'Unione europea a dettare obblighi di incriminazione specificamente e organicamente riferiti alla violenza di genere commessa nel *web*, con carattere vincolante per gli Stati membri e con un catalogo di fattispecie tipizzate. Tale qualificazione va tuttavia precisata. Sul piano della regolazione delle piattaforme, il regolamento (UE) 2022/2065[55] (*Digital Services Act*, c.d. DSA) aveva già introdotto, in via orizzontale e settorialmente non specifica, obblighi di diligenza graduati in capo agli intermediari digitali – comprensivi di meccanismi di *notice-and-action*, obblighi di trasparenza dei sistemi di raccomandazione e valutazione dei rischi sistemici per i diritti fondamentali – idonei a incidere indirettamente anche sulle forme di violenza digitale di genere. Sul piano

[52] D. K. CITRON, *The Fight for Privacy: Protecting Dignity, Identity, and Love in the Digital Age,* New York, 2022, spec. pp. 12-34.

[53] GREVIO, *General Recommendation No 1 on the digital dimension of violence against women*, 20 ottobre 2021, reperibile *online*.

[54] Direttiva (UE) 2024/1385 del Parlamento europeo e del Consiglio, del 14 maggio 2024, sulla lotta alla violenza contro le donne e alla violenza domestica. Per un commento si veda D. MARRANI, *La dimensione digitale della violenza contro le donne tra diritti umani e cybercriminalità/La dimensión digital de la violencia contra las mujeres entre los derechos humanos y la ciberddelincuencia*, 2025, Napoli, pp. 273-290.

[55] Regolamento (UE) 2022/2065 del Parlamento Europeo e del Consiglio del 19 ottobre 2022 relativo a un mercato unico dei servizi digitali e che modifica la direttiva 2000/31/CE (regolamento sui servizi digitali).

della protezione dei diritti individuali, la Corte EDU aveva già da tempo riconosciuto, a partire da *Volodina* c. *Russia*[56], che la sorveglianza digitale, l'uso di *spyware* e la diffusione non consensuale di dati personali costituiscono modalità specifiche di violenza di genere[57], generando obblighi positivi di prevenzione anche nella dimensione tecnologica. La specificità della direttiva 2024/1385 risiede dunque nell'introdurre, per la prima volta nell'ordinamento dell'Unione, obblighi di incriminazione penale mirati per condotte digitalmente mediate di violenza di genere. La direttiva riconosce esplicitamente che la violenza digitale è una componente strutturale della violenza di genere contemporanea e che la sua regolazione richiede strumenti giuridici dedicati che vadano oltre la logica della moderazione dei contenuti e della responsabilità delle piattaforme. Non sorprende, dunque, che quattro dei sei reati per i quali essa impone agli Stati membri di introdurre una fattispecie penale entro il 2027 presentino una modalità tipicamente *online* della condotta criminosa: si tratta, in specie, della condivisione non consensuale di materiale intimo o manipolato (art. 5), dello *stalking online* (art. 6), delle molestie *online* (art. 7) e dell'istigazione alla violenza o all'odio *online* (art. 8). Si può dunque sostenere che un sistema predittivo che non includa indicatori di violenza digitale è strutturalmente inadeguato rispetto agli obblighi convenzionali.

Occorre quindi procedere con la ricostruzione di una tassonomia giuridica della violenza tecnologica di genere, in quanto tale classificazione costituisce uno strumento dogmatico e operativo per individuare il centro di imputazione della responsabilità, la fonte dell'obbligo violato e il tipo di rimedio attivabile.

Nell'ampio spettro di siffatto polimorfico fenomeno la prima categoria individuabile è quella della violenza digitale diretta, in cui la tecnologia è lo strumento immediato della condotta lesiva. Una prima forma è rappresentata dal *cyberstalking* che,

---

[56] Corte europea dei diritti dell'uomo, sentenza del 14 settembre 2021, ricorso n. 40419/19, *Volodina* c. *Russia.* Il caso trae origine da una relazione violenta tra la sig.ra Volodina e il suo ex partner, caratterizzata da un'*escalation* di abusi non solo fisici e psicologici, ma anche tecnologicamente mediati. Dopo la fine della relazione, l'uomo aveva intrapreso una sistematica attività di controllo e persecuzione digitale nei confronti della ricorrente, installando sul suo telefono cellulare un *software* di sorveglianza (*spyware*) che gli consentiva di monitorarne gli spostamenti, accedere alle comunicazioni private e raccogliere dati personali sensibili. Le informazioni così ottenute venivano utilizzate per intimidire la vittima e, in alcuni casi, diffuse senza il suo consenso, aggravando ulteriormente la situazione di vulnerabilità e paura. Nonostante la presenza di elementi tecnici idonei a dimostrare l'intrusione informatica e le reiterate denunce presentate dalla ricorrente, le autorità russe si erano rifiutate di avviare un'indagine penale effettiva, qualificando i fatti come irrilevanti o non sufficientemente gravi. Dinanzi alla Corte EDU, la ricorrente lamentava che tale inerzia istituzionale avesse determinato una violazione dei suoi diritti fondamentali, in particolare sotto il profilo della protezione contro trattamenti inumani o degradanti e del rispetto della vita privata. La Corte ha riconosciuto che le forme di violenza digitale subite – inclusa la sorveglianza tramite *spyware* e la diffusione non consensuale di dati personali – rientrano nella sfera della violenza domestica e impongono agli Stati obblighi positivi di prevenzione, investigazione e repressione.

[57] É. DUBOUT, *Digital Surveillance and the ECHR*, in *Revue trimestrielle des droits de l'homme*, 2022, pp. 345-360.

con una prevalenza del 70-80% di vittime donne[58], ne costituisce una delle manifestazioni più diffuse, con lo scenario tipico di un ex partner che installa *spyware* sul telefono della vittima, ne monitora la geolocalizzazione, controlla i suoi *account* sui *social media*, e, così facendo, genera uno stato di sorveglianza totale i cui effetti psicologici replicano – e spesso superano – quelli della presenza fisica. Una seconda tipologia è *doxxing*, che consiste nella pubblicazione non consensuale di informazioni private *online* per esporre la vittima a molestie da parte di terzi, sfruttando l'architettura aperta di internet per moltiplicare la violenza senza che il suo autore ne risponda personalmente. Rientrano altresì in questa categoria i *deepfake* pornografici, ossia contenuti audiovisivi sintetici generati mediante algoritmi di *deep learning* che sovrappongono l'identità iconografica di una persona reale su un corpo diverso. Tali pratiche colpiscono prevalentemente le donne e presentano una dimensione politica spesso sottovalutata, in quanto operano come strumenti sistematici di silenziamento delle voci femminili nello spazio pubblico, scoraggiando la partecipazione delle donne alla vita politica, giornalistica e accademica.

La seconda categoria è costituita dalla *violenza digitale mediata*, nella quale la tecnologia non genera necessariamente l'atto lesivo originario, ma ne amplifica portata, durata e intensità attraverso specifiche configurazioni infrastrutturali. In tale ambito assumono rilievo le architetture di piattaforma, in particolare i sistemi di raccomandazione e di *ranking* algoritmico, che incidono sulla visibilità dei contenuti, le procedure di *notice-and-action*, nonché i meccanismi di condivisione e replicabilità che favoriscono la diffusione virale. Tali elementi concorrono a trasformare un singolo episodio di aggressione in un fenomeno continuativo e potenzialmente illimitato di ri-vittimizzazione, inserito nelle dinamiche proprie dell'ecosistema digitale. In questa prospettiva, le tecnologie dell'informazione operano come fattori di amplificazione del danno, determinando una progressiva dissociazione tra l'atto originario e i suoi effetti, che si propagano secondo logiche autonome rispetto al controllo dell'autore. Ne deriva che la responsabilità non può essere ricondotta esclusivamente al soggetto agente, ma assume una dimensione almeno in parte distribuita, coinvolgendo anche le piattaforme digitali nella misura in cui le loro scelte organizzative, tecniche e procedurali contribuiscano causalmente alla diffusione e alla persistenza del contenuto lesivo. In tale quadro si inserisce il già menzionato regolamento (UE) 2022/2065 (DSA), il quale, pur non essendo specificamente rivolto alla violenza di genere, introduce un sistema di obblighi di *due diligence* graduati in capo ai prestatori di servizi intermediari. In particolare, il DSA prevede meccanismi di *notice-and-action* (art. 16), obblighi di trasparenza dei sistemi di raccomandazione (art. 27) e, per le piattaforme *online* di dimensioni molto grandi, obblighi di valutazione e mitigazione dei rischi sistemici (artt.

[58] European Institute for Gender Equality, *Combating Cyber Violence against Women and Girl*s, Publications Office of the EU, 2022, spec. pp. 23-45, reperibile *online*; K. SOUTHWORTH, *Intimate Partner Violence, Technology, and Stalking*, in *Violence against Women*, 2007, pp. 842-856.

34-35), inclusi quelli relativi alla diffusione di contenuti illegali e agli effetti negativi sui diritti fondamentali. Tali disposizioni assumono rilievo anche con riferimento alla violenza digitale mediata, nella misura in cui contribuiscono a delineare un quadro di responsabilità regolatoria volto a limitare l'amplificazione dei contenuti lesivi all'interno dell'ecosistema digitale.

La terza categoria, che il presente contributo propone, è quella della violenza algoritmica strutturale, la quale si innesta nell'architettura stessa dei sistemi istituzionali preposti alla protezione delle donne. A differenza della violenza diretta, essa non è riconducibile a un aggressore individuale e, diversamente dalla violenza mediata, non si limita ad amplificare una violenza preesistente, ma si produce nella normale operatività di sistemi tecnologici la cui dimensione discriminatoria è incorporata nella logica del codice e nei criteri di funzionamento. In questa prospettiva, la lesione non deriva né da un soggetto privato, né dalle dinamiche di amplificazione proprie delle piattaforme, ma dall'integrazione di *bias* e criteri discriminatori all'interno di sistemi automatizzati impiegati da istituzioni pubbliche, quali modelli predittivi, strumenti di valutazione del rischio e sistemi di supporto decisionale in materia di protezione, custodia, misure restrittive o accesso ai servizi. L'elemento di novità giuridica risiede nel fatto che lo Stato non si configura più (soltanto) come garante inadempiente, bensì come soggetto attivo nella produzione del rischio, in quanto adotta, finanzia e utilizza tali strumenti. Ne deriva uno spostamento paradigmatico della responsabilità, che si colloca non più sul piano dell'omissione, ma su quello della produzione «istituzionale» del rischio. In altri termini, la specificità della violenza algoritmica strutturale consiste nella coincidenza tra l'autore della discriminazione e il titolare dell'obbligo di prevenzione, configurando una categoria concettuale che, per la sua struttura, non trova precedenti nelle tipologie tradizionali di violenza tecnologicamente facilitata. Proprio per tale ragione, essa si presenta come una sfida normativa in larga misura inesplorata per il diritto internazionale dei diritti umani, imponendo una riconsiderazione dei modelli di imputazione della responsabilità statale. In tale prospettiva, assumono rilievo, da un lato, la già menzionata Raccomandazione generale n. 35 del Comitato CEDAW, che rafforza gli obblighi di *due diligence* rispetto alla violenza di genere anche in presenza di attori non statali o strumenti tecnologici, e, dall'altro, l'AI Act, che disciplina i sistemi di intelligenza artificiale ad alto rischio in funzione della tutela dei diritti fondamentali e della prevenzione della discriminazione.

Nel solco della tripartizione proposta, la chiave esegetica che si propone ha come fulcro, non tanto la descrizione fenomenologica delle singole condotte, quanto la riconduzione delle diverse manifestazioni della violenza tecnologica di genere a modelli differenziati di imputazione della responsabilità. La tassonomia, infatti, consente di leggere il fenomeno, più che quale mera sommatoria di pratiche digitali, come un sistema articolato in cui mutano progressivamente il rapporto tra autore, mezzo e fonte dell'obbligo giuridico, con effetti diretti sui criteri di attribuzione della responsabilità e

sui rimedi attivabili. Non sfugge, a chi scrive, che tale operazione, tuttavia, pur offrendo una chiave ordinante efficace, espone anche a criticità teoriche e applicative, di cui si cercherà di dare conto.

Con riferimento alla violenza digitale diretta, la tecnologia si configura quale strumento immediato della condotta lesiva, consentendo di mantenere una sostanziale continuità con i paradigmi classici del diritto penale. L'agente utilizza l'infrastruttura digitale come prolungamento della propria azione, sicché la lesione dei beni giuridici è direttamente riconducibile alla sua condotta secondo un modello di responsabilità prevalentemente individuale e fondato su un nesso causale lineare. Tuttavia, tale apparente continuità è solo parziale. Fenomeni quali il *doxxing* o i *deepfake* introducono dinamiche di diffusione incontrollabile e di moltiplicazione del danno che eccedono la sfera di dominio dell'autore, determinando una frattura tra azione e conseguenze. Ne deriva una prima criticità, poiché il diritto continua a imputare la responsabilità secondo schemi lineari, mentre il danno si sviluppa secondo logiche reticolari e cumulative. Ciò solleva interrogativi circa l'estensione dell'imputazione oggettiva, il rischio di sovra-responsabilizzazione dell'agente iniziale e, al contempo, la difficoltà di intercettare contributi lesivi successivi, non facilmente tipizzabili.

La violenza digitale mediata segna un passaggio ulteriore, poiché la tecnologia non si limita più a fungere da mezzo, ma diviene parte integrante del processo lesivo. La violenza assume qui una natura co-prodotta, risultante dall'interazione tra la condotta dell'autore e le caratteristiche strutturali delle piattaforme digitali. Le architetture algoritmiche e le logiche di funzionamento delle piattaforme contribuiscono all'amplificazione, alla persistenza e alla viralità del contenuto offensivo, trasformando un atto originario in una sequenza potenzialmente illimitata di ri-vittimizzazione. In tale contesto, la responsabilità si configura come distribuita e condivisa fra una pluralità di attori – peraltro di natura eterogenea (individui e piattaforme). Tuttavia, proprio questa distribuzione costituisce il principale profilo critico della categoria in quanto, da un lato, vi è il rischio di una diluizione della responsabilità, per cui la molteplicità dei soggetti coinvolti rende difficile individuare un centro di imputazione effettivo; dall'altro, emerge la difficoltà di distinguere tra neutralità tecnica e contributo causale della piattaforma. In effetti, il passaggio da una causalità lineare a una causalità sistemica mette sotto pressione i criteri tradizionali di imputazione, mentre gli strumenti normativi fondati sulla *due diligence,* come quelli previsti dal DSA, rischiano di risultare insufficienti a garantire una tutela effettiva, soprattutto quando le scelte algoritmiche, pur formalmente lecite, producono effetti discriminatori o lesivi su larga scala. Va considerato, inoltre, il rischio opposto, ossia quello di una eccessiva estensione della responsabilità delle piattaforme, con possibili effetti di *overblocking* e conseguente compressione della libertà di espressione.

La violenza algoritmica strutturale rappresenta, infine, il livello più avanzato e problematico della tassonomia, poiché mette in discussione i presupposti stessi della responsabilità giuridica tradizionale. In questa ipotesi, la violenza non è riconducibile né a un agente individuale né all'amplificazione di una condotta preesistente, ma è incorporata nella struttura dei sistemi automatizzati impiegati dalle istituzioni pubbliche. Modelli predittivi, strumenti di valutazione del rischio e sistemi di supporto decisionale possono, infatti, riprodurre e consolidare *bias* discriminatori, incidendo su diritti fondamentali quali la libertà personale, l'accesso alla protezione o la parità di trattamento. Come sottolineato, il dato decisivo è rappresentato dalla coincidenza tra il soggetto autore della lesione e il titolare dell'obbligo di prevenzione: lo Stato non è più un garante esterno che omette di intervenire, ma diviene esso stesso produttore istituzionale del rischio. La responsabilità si sposta così dal piano dell'omissione a quello dell'azione, configurandosi come responsabilità strutturale e sistemica, che trova i propri riferimenti giuridici e la base normativa nel diritto internazionale dei diritti umani. In tale prospettiva, strumenti come la Raccomandazione generale n. 35 del Comitato CEDAW e l'AI Act assumono rilievo centrale, nella misura in cui impongono obblighi di prevenzione *ex ante*, valutazione dei rischi e controllo dei *bias.*

Alla luce di quanto precede, la tripartizione elaborata è volta a individuare tre distinti modelli di responsabilità: individuale, nella violenza diretta; distribuita, nella *violenza mediata*; strutturale e pubblica, nella violenza algoritmica. Tale articolazione, al netto delle criticità evidenziate, si colloca nel solco di una progressiva trasformazione del diritto, che si sposta dalla repressione della condotta alla regolazione dei sistemi, dalla colpevolezza individuale alla gestione del rischio tecnologico, fino alla messa in discussione del ruolo dello Stato come garante. Ne deriva una chiave di lettura capace di descrivere le forme mutevoli della violenza tecnologica di genere e, a un tempo, di orientare la risposta giuridica in termini coerenti con la complessità del fenomeno, evitando tanto derive panpenalistiche, quanto lacune di tutela, cercando altresì di favorire un approccio multilivello in grado di integrare responsabilità individuale, regolazione delle piattaforme e *accountability* istituzionale, senza sacrificare né l'effettività della tutela (in specie: delle donne vittime di violenza), né le garanzie fondamentali dell'ordinamento.

4. Sistemi europei di gestione integrata del rischio di violenza di genere a confronto: VioGén, SARA, DASH e Codice Rosso.

Nel panorama europeo di *risk governance* della violenza di genere, i sistemi VioGén, SARA, DASH e il modello procedurale del Codice Rosso rappresentano quattro configurazioni profondamente diverse del rapporto tra conoscenza del rischio, ruolo dell'operatore umano e responsabilità dello Stato. La loro analisi, condotta in due tempi

– dapprima attraverso una illustrazione sistematica di ciascuno strumento, successivamente mediante un confronto trasversale alla luce degli obblighi internazionali applicabili – consente di mostrare che i problemi strutturali di questi sistemi patiscono segnali di una tensione irrisolta tra la logica probabilistica dei modelli prognostici e le esigenze di un diritto internazionale che impone valutazioni autonome, non discriminatorie e *victim-centered* del rischio.

Innanzitutto, occorre prendere in considerazione il *Sistema de seguimiento integral en los casos de Violencia de Genero* (VioGén) istituito in Spagna nel 2007, in attuazione della *Ley Organica 1/2004 de Medidas de Proteccion Integral contra la Violencia de Generoe,* oggetto di ampia analisi nella letteratura giuridica spagnola[59], e gestito dal Ministero dell'Interno, che, nel 2025, anche all'esito delle critiche avanzate dall'*audit* esterno di *Eticas Foundation*[60], ha adottato il *Protocolo de Actuación*, implementando una nuova piattaforma tecnica denominata Sistema VioGén-2. Al momento dell'adozione di tale protocollo, VioGén registrava 831.500 casi attivi, configurandosi come il più esteso sistema centralizzato di monitoraggio della violenza di genere nell'Unione europea. Il funzionamento del sistema si articola attorno alla *Valoracion Policial del Riesgo* (VPR), un questionario strutturato composto da 42 indicatori che produce automaticamente una classificazione del rischio su cinque livelli – *desestimado*, *bajo*, *medio*, *alto*, *extremo* – aggiornabile attraverso la *Valoracion Policial de la Evolucion del Riesgo* (VPER). La scala quinaria consente una gestione coordinata dei casi e favorisce l'interoperabilità tra forze di polizia e servizi di protezione su tutto il territorio nazionale. Sul piano architetturale, VioGén si configura come un sistema prevalentemente basato su *automated risk scoring*, in cui la classificazione del rischio è generata in via primaria dall'elaborazione algoritmica degli indicatori, come evidenziato anche nella letteratura sul ruolo della polizia predittiva nella prevenzione della violenza di genere[61], mentre l'intervento umano assume una funzione prevalentemente correttiva e successiva, senza tradursi in una valutazione professionale strutturata antecedente alla formazione del punteggio. Il meccanismo di ponderazione degli indicatori, che determina il peso relativo di ciascun fattore nel calcolo del livello di rischio, è coperto da segreto industriale da parte del fornitore tecnologico, con conseguente opacità della logica decisionale interna del sistema. La classificazione VPR non configura un provvedimento amministrativo autonomamente impugnabile, ma uno strumento di supporto alla decisione poliziesca che si colloca in una zona grigia tra atto tecnico e determinazione suscettibile di incidere in modo rilevante sulla protezione della vittima. In un caso documentato dal GREVIO, una

[59] *Ex plurimis* cfr. E. CUESTA-GARCÍA, *VioGén: Origen, Evolución y Perspectivas*, in *Revista de Derecho Penal*, 2022, pp. 234-258; A. MONTESINOS GARCÍA, *Los algoritmos que valoran el riesgo de reincidencia: En especial, el sistema Viogen*, in *Revista de derecho y proceso penal*, 2021, pp. 19-55.

[60] Eticas Foundation, *Audit Externo del Sistema VioGén*, 2022, reperibile *online*.

[61] M. Á. PRESNO LINERA, *Policía predictiva y prevención de la violencia de género: el sistema VioGén*, in *Revista de Internet, Derecho y Política*, 2023, pp. 1-13.

vittima classificata *bajo* è stata successivamente aggredita dall'ex partner dopo il diniego delle misure protettive; tuttavia, proprio in ragione della natura ibrida dello strumento, né la vittima, né i suoi difensori hanno avuto la possibilità di contestare nel merito la valutazione algoritmica[62].

Le analisi indipendenti e i rapporti del GREVIO evidenziano criticità strutturali sia nella validazione degli indicatori sia nella capacità del sistema di intercettare situazioni di rischio in contesti di marginalità istituzionale. Essi documentano, inoltre, un elevato tasso di *over-reliance* da parte degli operatori, fenomeno riconducibile alla c.d. «deferenza algoritmica», per cui gli agenti tendono ad aderire automaticamente al punteggio restituito dal sistema anche in presenza di elementi contestuali di segno contrario. In tali condizioni, lo strumento non si limita a supportare il giudizio professionale, ma, di fatto, lo sostituisce, determinando una sostanziale eterodirezione della decisione pubblica. Una simile dinamica risulta incompatibile con il requisito di supervisione umana effettiva previsto dall'art. 14 dell'AI Act, il quale impone che l'intervento umano sia reale, informato e idoneo a prevenire o correggere esiti distorti del sistema. Ne consegue che l'affidamento acritico all'*output* algoritmico, oltre a essere una criticità operativa, incide direttamente sulla legittimità giuridica dell'impiego di tali sistemi.

In secondo luogo, rileva il *Spousal Assault Risk Assessment* (SARA), sviluppato da Kropp *et al.* negli anni Novanta[63] e ampiamente diffuso in diversi contesti europei, che si inserisce nel paradigma dello *Structured Professional Judgment* (SPJ), caratterizzato dall'utilizzo di *checklist* standardizzate di fattori di rischio integrate da una valutazione discrezionale del professionista, senza ricorso a modelli di *scoring* automatizzato. A differenza di VioGén, infatti, il SARA non produce un punteggio algoritmico, ma offre una struttura metodologica di 20 fattori di rischio, organizzati in àmbiti quali la storia di violenza, la presenza di fattori psicosociali e la situazione della vittima, che orientano il giudizio del professionista senza sostituirlo[64].

Questa architettura SPJ presenta rilevanti vantaggi sotto il profilo della trasparenza e dell'intellegibilità, in quanto il processo valutativo è esplicitabile e verificabile, il professionista mantiene la piena responsabilità della decisione finale e l'esito può essere illustrato alla vittima e ai suoi difensori senza mediazioni algoritmiche opache. L'assenza di dipendenza da *dataset* storici attenua inoltre, sul piano strutturale, il rischio di discriminazione indiretta nei confronti delle categorie sottorappresentate nelle banche dati istituzionali.

---

[62] GREVIO, *Baseline Evaluation Report: Spain*, 2020, parr. 78-83, reperibile *online*.

[63] Cfr. *supra*, nota 12.

[64] E. STARK, *Coercive Control: How Men Entrap Women in Personal Life*, Oxford, 2007, spec. pp. 345-389.

Tuttavia, tale modello presenta un limite speculare rispetto ai sistemi automatizzati, riconducibile alla variabilità inter-operatore e alla dipendenza dalla qualità della formazione. In assenza di programmi formativi specialistici e continuativi, il giudizio professionale può essere influenzato da *bias* individuali, inclusi stereotipi di genere o culturali, con effetti potenzialmente analoghi a quelli prodotti dai *bias* algoritmici. Ne deriva che l'efficacia del SARA è strettamente condizionata dalla capacità dello Stato di assicurare standard elevati e uniformi di formazione, nonché meccanismi di controllo e aggiornamento continuo delle competenze degli operatori.

Vi è poi il *Domestic Abuse, Stalking and Honour-Based Violence Risk Identification, Assessment and Management* (DASH), diventato standard nazionale in Inghilterra e Galles nel 2009 e aggiornato nel 2023[65], che adotta una struttura ibrida che combina una *checklist* standardizzata di indicatori con una componente di valutazione professionale, collocandosi a metà strada tra i sistemi di *automated scoring* e il paradigma SPJ. L'elemento di maggiore rilievo del DASH, nella sua versione aggiornata, è l'integrazione di indicatori specifici sulla violenza digitale e sul controllo coercitivo. La versione 2023 include infatti specifici *item* sull'installazione di *software* di monitoraggio sui dispositivi della vittima, sulla pubblicazione non consensuale di immagini *online* e sull'uso di piattaforme digitali per minacciare o controllare. Tale integrazione risponde alla crescente interazione tra dimensione fisica e digitale della violenza domestica contemporanea, caratterizzata da forme di esposizione continuativa della vittima e da effetti lesivi difficilmente reversibili. Benché la struttura ibrida del DASH garantisca la permanenza di una componente di giudizio professionale che consente di contestualizzare i dati raccolti, purtuttavia la sua efficacia è condizionata dalla necessità di aggiornamenti costanti degli indicatori e da adeguati programmi di formazione degli operatori, senza i quali la capacità del sistema di rappresentare le trasformazioni della violenza digitale rischia di degradarsi nel tempo.

Infine, nel contesto italiano, la legge 19 luglio 2019, n. 69 (c.d. Codice Rosso) adotta un approccio radicalmente diverso rispetto ai tre sistemi precedentemente esaminati, poiché non introduce alcun meccanismo di valutazione automatizzata o strutturata del rischio, ma concentra l'intervento sulla tempestività della risposta istituzionale. Il suo asse portante è costituito dall'obbligo di riferimento immediato alla procura entro 48 ore dalla denuncia (art. 1) e dall'obbligo del pubblico ministero di assumere informazioni dalla persona offesa entro tre giorni dalla comunicazione della notizia di reato (art. 1, comma 2). L'adozione di questi termini procedurali è direttamente riconducibile sia alla condanna decisa dalla Corte di Strasburgo nel caso *Talpis* c. *Italia*[66],

[65] Cfr. *supra* nota13.

[66] [66] Corte europea dei diritti dell'uomo, sentenza del 2 marzo 2017, ricorso n. 41237/14, *Talpis* c. *Italia.* Il caso riguarda la sig.ra Talpis, una donna di origine moldava residente in Italia, vittima di reiterate violenze domestiche da parte del marito, caratterizzate da aggressioni fisiche, minacce e abusi in un contesto di consumo abituale di alcol da parte dell'uomo. La ricorrente aveva denunciato più volte i

che aveva identificato nella latenza istituzionale tra denuncia e risposta giudiziaria una violazione degli obblighi positivi di tutela derivanti dagli artt. 2 (diritto alla vita) e 3 (divieto di trattamenti inumani o degradanti) nonché l'art. 14 (divieto di discriminazione), riconoscendo che l'inerzia istituzionale si inseriva in un contesto di sottovalutazione strutturale della violenza domestica nei confronti delle donne, sia al caso *Landi* c. *Italia*[67], in cui la Corte EDU ha accertato la violazione dell'art. 2 ritenendo che le autorità non avessero adempiuto agli obblighi positivi di protezione, in particolare per non aver reagito con la dovuta diligenza alle segnalazioni di pericolo e per non aver effettuato una valutazione autonoma e proattiva del rischio. Il Codice Rosso si configura quindi, almeno nella sua genesi, come risposta normativa a una sistematica inadempienza procedurale dello Stato italiano, prima ancora che come politica preventiva.

Sul versante penale sostanziale, la legge introduce nuove fattispecie incriminatrici, tra cui la diffusione illecita di immagini o video sessualmente espliciti (art. 612-ter c.p., c.d. *revenge porn*).

Le valutazioni sull'efficacia complessiva del Codice Rosso sono contrastanti: se, da un lato, i dati ISTAT[68] indicano un aumento significativo delle denunce, dall'altro, i rapporti del GREVIO[69] e dell'Associazione Antigone[70] rilevano che tale incremento quantitativo non si è tradotto in una riduzione significativa dei femminicidi.

---

comportamenti violenti alle autorità italiane, senza tuttavia ottenere una risposta tempestiva ed efficace. In particolare, a seguito di una grave aggressione, la donna si era rivolta alle forze dell'ordine e ai servizi sanitari, ma il procedimento penale era rimasto inattivo per un periodo significativo e non erano state adottate misure protettive adeguate. Nei mesi successivi, la situazione di violenza era ulteriormente degenerata fino a sfociare, nel 2013, in un attacco particolarmente grave: il marito, in stato di ebbrezza, aveva aggredito la ricorrente con un coltello. Il figlio della coppia, intervenuto per difendere la madre, era stato ucciso dall'aggressore, mentre la donna riportava gravi ferite. Solo dopo tale episodio le autorità avevano adottato misure restrittive nei confronti dell'uomo. Dinanzi alla Corte EDU, la ricorrente lamentava che l'inerzia e i ritardi delle autorità italiane nell'intervenire, nonostante le ripetute segnalazioni di rischio, avessero contribuito in modo determinante all'esito letale della vicenda. Per un commento in dottrina si veda E. BREMS, *Talpis v Italy and the ECHR*, in *Human Rights Law Review*, 2018, pp. 567-572.

[67] Corte europea dei diritti dell'uomo, sentenza del 7 aprile 2022, ricorso n. 10929/19, *Landi* c. *Italia*. Il caso trae origine dalla vicenda della sig.ra Landi, vittima di reiterate violenze domestiche da parte del marito, affetto da disturbi psichiatrici e già segnalato alle autorità per comportamenti aggressivi. La ricorrente aveva denunciato più volte le violenze, ma tali denunce erano state in parte ritirate e, in ogni caso, non avevano determinato l'adozione di misure protettive efficaci. Nonostante le segnalazioni dei carabinieri, che avevano evidenziato la pericolosità dell'uomo e suggerito l'adozione di misure restrittive, la procura non aveva disposto interventi adeguati né aveva proceduto a una valutazione strutturata del rischio. La situazione è degenerata nel 2018, quando l'uomo ha ucciso il figlio della coppia e tentato di uccidere la ricorrente. Solo dopo l'evento letale sono state adottate misure incisive nei confronti dell'aggressore. Dinanzi alla Corte europea dei diritti dell'uomo, la ricorrente ha lamentato come l'inazione delle autorità italiane, nonostante la prevedibilità del rischio, avesse contribuito in modo determinante all'esito tragico. Per un commento si veda C. BELCASTRO, *La Corte EDU si pronuncia nuovamente sugli obblighi positivi nei casi di violenza domestica: le sentenze Landi, De Giorgi e M.S. c. Italia a confronto*, in *Ordine internazionale e diritti umani*, 2022, pp. 1024-1034, reperibile *online*.

[68] ISTAT, *La violenza contro le donne 2022-2023*, 2024, spec. pp. 12-34.

[69] GREVIO, *Baseline Evaluation Report: Italy*, 2019, cit., parr. 45-67.

[70] Associazione Antigone, *XXI Rapporto*, 2025, reperibile *online*, spec. p. 289.

In buona sostanza, il Codice Rosso, pur evitando i rischi di discriminazione automatizzata propri dei sistemi predittivi, si caratterizza per l'assenza di uno standard nazionale uniforme di valutazione del rischio, lacuna già segnalata dal Piano Strategico Nazionale 2021-2023[71] come priorità irrisolta. Questa carenza produce una maggiore dipendenza dalla prassi operativa e dalla sensibilità del singolo operatore, con conseguenti effetti sull'omogeneità e sulla prevedibilità delle risposte istituzionali.

Occorre ora procedere con la comparazione tra i quattro sistemi che, letta attraverso gli obblighi derivanti da CEDAW, Convenzione di Istanbul, AI Act, GDPR, Principi Guida ONU su imprese e diritti umani e Convenzione di Belém do Parà, consente di cogliere come ciascun modello incorpori una specifica concezione di vulnerabilità e, soprattutto, una diversa capacità di prevenire o riprodurre forme di discriminazione strutturale. Si può già anticipare che il quadro che emerge è quello di sistemi che, nella migliore delle ipotesi, soddisfano parzialmente gli obblighi internazionali e che, nella peggiore, li violano sistematicamente.

In effetti, VioGén costituisce l'esempio più emblematico di un sistema predittivo fondato sulla trasformazione della traccia documentale in rischio futuro. La sua architettura algoritmica produce classificazioni che riflettono più la storia istituzionale dei dati che la vulnerabilità reale delle vittime. Le donne migranti, le donne senza figli minori, le donne anziane e, più in generale, tutte le categorie caratterizzate da una minore interazione con le istituzioni vengono sistematicamente classificate a rischio inferiore, fenomeno ampiamente evidenziato anche nella letteratura sui sistemi di *risk assessment* applicati alle donne migranti[72]: la sottodenuncia storica – ossia il fenomeno per cui i dati ufficiali non riflettono la reale dimensione del fenomeno a fronte del fatto che molte vittime non si rivolgono alle autorità o non lasciano «tracce» istituzionali – diventa, nel modello, assenza di rischio. Questo esito è il prodotto di un intreccio di *bias* strutturali –

[71] Presidenza del Consiglio dei ministri, *Piano Strategico Nazionale sulla Violenza Maschile contro le Donne 2021–2023*, 17 novembre 2021, reperibile *online*, spec. p. 25 ss.

[72] C. LAVERY, A. SHEEHAN, *Risk Assessment Tools and Migrant Women*, in *International Journal of Comparative and Applied Criminal Justice*, 2023, pp. 178-193.

bias di *dataset* storico[73], *bias* di selezione[74] e *bias* di misurazione[75] – che convergono nel generare un *disparate impact* vietato dal diritto internazionale. Come già evidenziato, la CEDAW, nella sua interpretazione consolidata – in particolare alla luce della raccomandazione generale n. 35 – non richiede la dimostrazione dell'intenzionalità discriminatoria: è sufficiente l'effetto prodotto. Ne consegue che anche sistemi privi di finalità discriminatorie (tanto, a monte, del legislatore, quanto, a valle del singolo operatore) possono integrare una violazione convenzionale qualora generino effetti di discriminazione vietata. Ancora, sebbene l'art. 35 della Convenzione di Istanbul imponga agli Stati di dotarsi di procedure di valutazione del rischio che tengano conto dei fattori di vulnerabilità multipla e l'art. 4 vieti qualsiasi discriminazione quanto alla tutela da garantire alle vittime, VioGén produce esiti che, sistematicamente, assicurano meno protezione alle donne in condizione di maggiore vulnerabilità intersezionale: un risultato frontalmente incompatibile con entrambe le disposizioni. Per quel che concerne l'AI Act, ai sensi dell'Allegato III, punto 6, i sistemi di valutazione del rischio rientrano tra i sistemi di intelligenza artificiale ad alto rischio, ai quali si applicano gli obblighi di gestione del rischio di cui all'art. 9 e di supervisione umana significativa di cui all'art. 14. In tale prospettiva, l'*over-reliance* documentata degli operatori di VioGén non costituisce una mera criticità operativa, ma incide direttamente sulla conformità giuridica del sistema, in quanto svuota di contenuto il requisito della supervisione umana effettiva, riducendola a un controllo meramente formale. Quando l'intervento umano si traduce in una ratifica automatica dell'*output* algoritmico, il sistema opera, di fatto, come un meccanismo decisionale automatizzato, in contrasto con la *ratio* del regolamento e con le condizioni di legittimità dei sistemi ad alto rischio. Sotto il profilo del GDPR, l'opacità del meccanismo di ponderazione, protetto da segreto industriale, compromette la possibilità per l'interessata di comprendere, contestare e ottenere la revisione della valutazione algoritmica, inscrivendosi nella più ampia problematica della *black box society*, in cui

[73] Il *bias* di *dataset* storico emerge quando i dati utilizzati per addestrare sistemi di IA e *machine learning* incorporano le disuguaglianze, i pregiudizi e le gerarchie sociali presenti nel contesto in cui quei dati sono stati prodotti. Poiché i *dataset* non sono mai «neutrali» in quanto rappresentano il modo in cui istituzioni e società hanno classificato, interpretato e trattato le persone nel passato, un modello addestrato su tali informazioni finisce per riprodurre, e spesso intensificare, quelle stesse distorsioni nelle decisioni future.

[74] Il *bias* di selezione si manifesta quando il campione di dati o i soggetti inclusi in un'analisi non rispecchiano realmente la popolazione a cui lo studio intende riferirsi. La distorsione nasce già nella fase di ingresso: l'arruolamento avviene in modo parziale, non casuale o non rappresentativo, e questo compromette la solidità metodologica dell'indagine, rendendo le conclusioni valide solo per quel gruppo ristretto e non estendibili alla popolazione più ampia, soprattutto all'esito di peculiarità compromesse nel caso in cui alcuni gruppi risultino sovraesposti, altri quasi assenti, alterando la fotografia complessiva del fenomeno.

[75] Il *bias* di misurazione si verifica quando gli strumenti, le procedure o i criteri utilizzati per raccogliere o valutare i dati introducono una deviazione sistematica rispetto alla realtà osservata. Non si tratta di errori casuali, ma di distorsioni strutturate, prodotte da strumenti non calibrati, protocolli incoerenti o metodi applicati in modo disomogeneo, che generano differenze sistematiche tra gruppi o variabili, anche per ignoranza dei diversi modelli culturali.

l'asimmetria informativa tra chi produce e chi subisce le decisioni automatizzate incide direttamente sull'effettività dei diritti[76]. Tale configurazione si pone in tensione con gli artt. 13 e 14 (che impongono la comunicazione di informazioni significative sulla logica del trattamento), con l'art. 22 (che limita il ricorso a decisioni basate unicamente su trattamenti automatizzati), nonché con il principio di trasparenza di cui all'art. 5, par. 1, lett. a). Ne deriva che l'assetto tecnico del sistema che, da un lato, solleva profili di incompatibilità con la disciplina europea sulla protezione dei dati, e, dall'altro, contribuisce a produrre un deficit di tutela procedurale che incide direttamente sull'effettività della protezione. Infine, alla luce dei Principi Guida ONU su imprese e diritti umani, in particolare dei principi 15 e 17, che, come già ricordato, delineano un dovere di *human rights due diligence* volto a identificare, prevenire e mitigare gli impatti negativi sui diritti fondamentali lungo l'intero ciclo di vita dei sistemi tecnologici, il modello di VioGén evidenzia una carenza strutturale nella gestione dei *bias* e dei rischi discriminatori, che non trova giustificazione nel richiamo alla presunta neutralità tecnica dello strumento. In tale prospettiva, il funzionamento del sistema configura una forma di violenza algoritmica strutturale, nella quale rischio e discriminazione risultano incorporati nella stessa architettura decisionale. In questo quadro, lo Stato assume un ruolo attivo nella produzione istituzionale del rischio attraverso l'adozione e l'impiego di un dispositivo che riproduce e amplifica le asimmetrie esistenti sotto forma di valutazioni apparentemente oggettive. Le conseguenze di tale configurazione si estendono alla responsabilità statale in relazione agli obblighi positivi di protezione e di non discriminazione, incidendo direttamente sull'effettività della tutela e determinando una compressione della protezione proprio nei confronti delle categorie maggiormente esposte a marginalità istituzionale.

Dal canto suo, il SARA, appartenendo al paradigma dello *Structured Professional Judgment*, si colloca in una posizione di maggiore compatibilità con gli obblighi internazionali. L'assenza di *automated scoring* soddisfa infatti strutturalmente l'obbligo di supervisione umana significativa previsto dall'art. 14 AI Act, nella misura in cui il sistema non produce un punteggio algoritmico vincolante bensì una struttura metodologica che orienta senza determinare. Ancora, il ragionamento valutativo resta esplicitabile e comprensibile, garantendo un livello di trasparenza compatibile con gli artt. 13-14 GDPR e con il diritto della persona offesa a comprendere la logica della valutazione che la riguarda. Inoltre, l'assenza di dipendenza da *dataset* storici riduce strutturalmente il rischio di discriminazione indiretta nei confronti delle donne migranti e di altre categorie sottorappresentate nelle banche dati istituzionali, in linea con gli obblighi della CEDAW, della Convenzione di Istanbul e della Convenzione di Belém do Parà. Quest'ultima, in particolare, impone agli Stati misure speciali per le donne in

[76] F. PASQUALE, *The Black Box Society*, Cambridge, 2015, spec. pp. 1-18.

situazione di vulnerabilità multipla (art. 9), che il SARA, a differenza dei sistemi automatizzati, può tenere in considerazione attraverso il giudizio professionale. Tuttavia, il funzionamento del SARA solleva questioni rilevanti sul piano della conformità agli obblighi internazionali, nella misura in cui l'affidamento al giudizio professionale implica che la qualità della valutazione dipenda direttamente dagli standard istituzionali di formazione e aggiornamento degli operatori. In tale prospettiva, le criticità già evidenziate nella letteratura sulla valutazione del rischio nei casi di violenza domestica[77] assumono rilievo giuridico, poiché incidono sulla capacità del sistema di garantire decisioni non discriminatorie e adeguate al rischio concreto. Ne deriva che l'effettività del SARA come strumento conforme agli standard internazionali è strettamente condizionata all'adempimento, da parte dello Stato, degli obblighi di formazione specialistica e continua previsti dall'art. 15 della Convenzione di Istanbul, quali componenti essenziali della *due diligence* in materia di prevenzione della violenza di genere.

Per quel che concerne il DASH, esso si configura come uno degli strumenti europei più avanzati in termini di conformità agli standard convenzionali, principalmente per due ragioni. La prima è la sua struttura ibrida, che preserva una componente di valutazione professionale coerente con il requisito di supervisione umana previsto dall'art. 14 dell'AI Act e contribuisce a garantire maggiore trasparenza nel processo valutativo. La seconda è rappresentata dall'integrazione di indicatori specifici relativi alla violenza digitale e al controllo coercitivo, che risulta in linea con gli obblighi internazionali in materia di prevenzione della violenza di genere. Se da un lato, infatti, la Convenzione di Istanbul definisce la violenza contro le donne in termini ampi (art. 3, lett. *b*), come interpretato dalla dottrina e dalla prassi, mentre l'art. 51 impone l'adozione di procedure strutturate di valutazione del rischio, dall'altro, la Raccomandazione Generale n. 35 del Comitato CEDAW (par. 20) ha chiarito che la violenza di genere può essere commessa o facilitata attraverso tecnologie digitali. Dunque, in tale contesto, il DASH aggiornato rappresenta uno dei pochi sistemi che incorporano operativamente tali dimensioni. Permangono, tuttavia, elementi di vulnerabilità, posto che in assenza di aggiornamenti costanti e di formazione continua degli operatori, la capacità del sistema di intercettare l'evoluzione delle forme digitali di controllo coercitivo rischia di indebolirsi, incidendo sulla capacità dello Stato di garantire una protezione effettiva, in linea con gli obblighi positivi elaborati ai sensi dell'art. 3 CEDU.

Infine, il c.d. Codice Rosso si colloca al di fuori del paradigma predittivo, privilegiando una logica di tempestività procedurale volta a rafforzare la risposta istituzionale ai casi di violenza domestica. Tale impostazione risulta coerente con gli obblighi positivi di protezione elaborati dalla giurisprudenza della Corte europea dei

[77] A. HOLT, *Risk Assessment in Domestic Violence Cases*, in *Criminology and Criminal Justice*, 2023, pp. 345-362.

diritti dell'uomo ai sensi dell'art. 2 CEDU, nonché con gli obblighi di intervento tempestivo e coordinato previsti dalla Convenzione di Istanbul, in particolare agli artt. 49-51. Sebbene l'assenza di strumenti algoritmici escluda, in linea di principio, i rischi di discriminazione automatizzata associati ai sistemi predittivi, permangono, tuttavia, rilevanti criticità strutturali. In particolare, il sistema non prevede uno standard nazionale uniforme di valutazione del rischio, con la conseguenza che l'identificazione della vulnerabilità e la definizione delle misure di protezione restano in larga misura affidate alla prassi operativa e alla discrezionalità degli operatori. Tale assetto può incidere sulla coerenza e sull'uniformità della protezione sul territorio nazionale e solleva profili di tensione rispetto agli obblighi di non discriminazione, anche alla luce della CEDAW, che richiede un accesso effettivo e non differenziato alla tutela. In assenza di criteri strutturati e condivisi, la risposta istituzionale risulta esposta al rischio di disparità applicative, con effetti potenzialmente rilevanti sul piano dell'uguaglianza sostanziale. Inoltre, quantunque l'introduzione di fattispecie penali quali il *revenge porn* (peraltro non scevra di difficoltà applicative nel caso in cui la diffusione del materiale sia riconducibile a soggetti estranei al rapporto sessuale[78]), segnali una crescente attenzione alla dimensione digitale della violenza di genere, tuttavia, tali interventi operano prevalentemente sul piano repressivo e non incidono in modo diretto sulla fase di valutazione preventiva del rischio, dinamica che riflette una più ampia tendenza all'affidamento alla risposta penale nella gestione della violenza domestica, la cui efficacia in termini di prevenzione è stata ampiamente problematizzata[79]. In questo senso, la mancanza di strumenti idonei a intercettare precocemente le forme di violenza digitale, incluse quelle riconducibili al controllo coercitivo mediato dalla tecnologia, limita la capacità del sistema di sviluppare una risposta pienamente preventiva, quale invece emerge nei modelli che integrano esplicitamente tali indicatori nella fase di *risk assessment*.

5. Standard esigibile di conformità algoritmica e principio di *non-refoulement* algoritmico nella protezione dalla violenza di genere.

Dall'analisi condotta attraverso la lettura trasversale dei quattro sistemi emerge, con riferimento alla questione dei *bias*, come essa costituisca un nodo giuridico centrale che attraversa l'intera architettura degli obblighi internazionali applicabili. I meccanismi causali dei *bias* – riconducibili, in particolare, ai *dataset* storici, ai processi di selezione e alle modalità di misurazione – producono effetti discriminatori ampiamente documentati

[78] P. BECCARI, *Le prime difficoltà applicative della nuova fattispecie di "revenge porn" in caso di diffusione del materiale da parte di soggetti estranei al rapporto sessuale*, in *Sistema penale*, 2022, pp. 5-23.

[79] L. GOODMARK, *A Troubled Marriage: Domestic Violence and the Legal System*, New York, 2012, spec. pp. 45-52.

dalla letteratura critica sull'intelligenza artificiale[80], che gli strumenti internazionali vietano espressamente.

Tale divieto trova fondamento, in primo luogo, nella CEDAW, nella sua consolidata interpretazione incentrata sul criterio dell'effetto[81], nonché, in ambito europeo, nei requisiti di gestione del rischio sistemico previsti dall'AI Act (art. 9[82]), negli obblighi di trasparenza e intellegibilità sanciti dal GDPR (artt. 5, 13-14, 22[83]). Analogamente, i Principi Guida ONU su imprese e diritti umani delineano un dovere di *due diligence* lungo l'intera filiera tecnologica[84], mentre la Convenzione di Istanbul impone garanzie di protezione strutturata e non discriminatoria (artt. 4 e 51[85]). Infine, la Convenzione di Belém do Pará richiede l'adozione di misure specifiche per le donne in condizione di vulnerabilità multipla (art. 9).

Se dunque la comparazione suggerisce che nessun modello, considerato isolatamente, soddisfa pienamente il quadro degli obblighi internazionali, l'esame sviluppato consente tuttavia di articolare sei requisiti cumulativi che un sistema di valutazione del rischio nella violenza di genere deve soddisfare per essere considerato conforme al diritto internazionale vigente. Tali requisiti, che assumono la forma di uno standard giuridico ricavabile in via interpretativa dalle fonti esaminate, presentano carattere cumulativo, con la conseguenza che il soddisfacimento parziale di uno o più di essi non esclude la responsabilità dello Stato per la violazione degli altri.

Il primo requisito attiene alla necessità di una valutazione strutturata del rischio capace di superare i limiti della mera tracciabilità istituzionale. In tal senso, se l'art. 51 della Convenzione di Istanbul impone agli Stati l'adozione di procedure di valutazione del rischio sistematiche e organizzate, la Raccomandazione Generale CEDAW n. 35 (par. 28) chiarisce che tali procedure devono tenere conto della sottodenuncia storica e delle barriere istituzionali all'accesso. Ne deriva che modelli di valutazione fondati

---

[80] R. MARTÍNEZ ALMANZA, S. PÉREZ DOMÍNGUEZ, *Género y herramientas algorítmicas: una revisión sistemática sobre los retos y el impacto que plantean en el ámbito judicial y penitenciario*, in *Revista de Internet, Derecho y Política*, 2026, pp. 1-15, reperibile online; R. RICHARDSON, *Dirty Data, Bad Predictions*, in *New York University Law Review Online*, 2019, pp. 193-228, reperibile online; S. NOBLE, *Algorithms of Oppression*, New York, 2018, spec. pp. 10-30; V. EUBANKS, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, cit.; J. SKEEM, C. LOWENKAMP, *Risk, Race, and Recidivism*, in *Criminology,* 2016, pp. 680-712, reperibile online; S. BAROCAS, A. D. SELBST, *Big Data's Disparate Impact*, in *California Law Review*, 2016, pp. 671-732.

[81] Raccomandazione generale n. 35, cit., par. 10; R. COOK, S. CUSACK, *Gender Stereotyping* cit., spec. pp. 7-33.

[82] M. VEALE, F. ZUIDERVEEN BORGESIUS, *Demystifying the Draft EU Artificial Intelligence Act - Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach*, in *Computer Law Review International*, 2021, pp. 97-112, reperibile online.

[83] B. GOODMAN, S. FLAXMAN, *European Union Regulations on Algorithmic Decision-Making*, in *AI Magazine*, 2017, pp. 50-57.

[84] UNGP, principi 13, 15 e 17; P. MUCHLINSKI, *Human Rights Due Diligence and the UN Guiding Principles on Business and Human Rights*, in *Business and Human Rights Journal*, 2020, pp. 130-148.

[85] GREVIO, *6th General Report on GREVIO'S Activities, January to December 2024*, reperibile online.

esclusivamente su dati documentali pregressi rischiano di trasformare la logica della *data-driven prediction* in un meccanismo di cristallizzazione delle asimmetrie esistenti, anziché in uno strumento di correzione delle stesse. In una prospettiva conforme agli obblighi internazionali, la valutazione del rischio deve invece integrare indicatori idonei a intercettare forme di vulnerabilità non pienamente rappresentate nei dati, inclusi i segnali di controllo coercitivo e le condizioni di marginalità istituzionale.

Il secondo requisito è la supervisione umana significativa, effettiva e non meramente formale. L'art. 14 AI Act, letto in combinato disposto con il considerando 60, che specifica che la supervisione deve essere «reale ed effettiva», impone che il professionista mantenga la capacità di comprendere, valutare e, se necessario, sovvertire l'*output* del sistema. Questo requisito non è soddisfatto ove si registri il fenomeno dell'*over-reliance*, il quale trasforma il controllo umano in una mera ratifica della decisione algoritmica, laddove, viceversa, la supervisione umana richiesta dall'AI Act non può essere ridotta a un momento formale di convalida, dovendo preservare la possibilità di un giudizio autonomo fondato sul caso concreto.

Il terzo requisito è quello di trasparenza e intellegibilità del processo valutativo. Il GDPR (*ex* artt. 13-14 e art. 22) impone che la logica sottostante al trattamento automatizzato sia comunicata all'interessata in forma comprensibile, problematica ulteriormente complicata dai limiti strutturali dell'accessibilità cognitiva del processo decisionale nei sistemi predittivi, ampiamente evidenziati nella letteratura sul *predictive policing*[86]. In effetti, il c.d. diritto a una spiegazione, nell'accezione elaborata criticamente dalla dottrina[87], richiede che la vittima e i suoi avvocati possano comprendere le ragioni dell'attribuzione di un determinato livello di rischio e, ove necessario, contestarlo. L'opacità dei modelli commerciali protetti da segreto industriale non soddisfa questo requisito e genera uno schermo procedurale incompatibile con le garanzie del giusto processo nella dimensione della protezione dalla violenza.

Il quarto requisito è la *due diligence* tecnologica lungo l'intera filiera di fornitura, postulati dai Principî 13, 15 e 17 degli UNGP che impongono alle imprese di identificare, prevenire, mitigare nonché rendere conto degli impatti negativi sui diritti umani prodotti dai propri sistemi, inclusi quelli derivanti dalla loro adozione da parte di enti statali. Lo Stato, dal canto suo, ha l'obbligo di non adottare tecnologie che producono effetti discriminatori sistemici sui diritti fondamentali, nonché di assicurarsi che i contratti di appalto pubblico con fornitori tecnologici incorporino clausole di *human rights impact assessment*. La letteratura sull'automazione ha ampiamente documentato come

[86] R. LEINS, *Predictive Policing and the Problem of Explainability*, in *Law, Innovation and Technology*, 2021, pp. 1-26.

[87] S. WACHTER, B. MITTELSTADT, C. RUSSELL, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, in *Harvard Journal of Law & Technology*, 2017, pp. 841-887, reperibile *online*.

l'adozione acritica di sistemi commerciali di *predictive policing* riproduca e amplifichi le disuguaglianze strutturali preesistenti[88].

Il quinto requisito è l'integrazione di indicatori specifici sulla violenza digitale e sul controllo coercitivo. In effetti, poiché la Raccomandazione Generale CEDAW n. 35 (par. 20) riconosce che le forme digitalmente mediate di violenza e controllo costituiscono violenza di genere a pieno titolo, un sistema di valutazione del rischio che non incorpori indicatori sulla sorveglianza digitale, sulle minacce via piattaforme, sul *doxing* e sull'installazione di *stalkerware* è strutturalmente incapace di rilevare una quota crescente della violenza domestica contemporanea, lasciando esposte proprio le vittime di forme di violenza più recenti e meno visibili ai dispositivi istituzionali tradizionali.

Il sesto e ultimo requisito è la protezione rafforzata delle vittime in condizione di vulnerabilità multipla. Se, a livello regionale, la Convenzione di Belém do Parà (art. 9) impone agli Stati di prestare particolare attenzione alle donne che si trovano in condizione di vulnerabilità derivante da più fattori concorrenti, tra cui la condizione migratoria, l'assenza di *status* regolare, la dipendenza economica e l'isolamento sociale, a livello universale, la Raccomandazione Generale CEDAW n. 33 (par. 26) ha esteso questo obbligo all'accesso alla giustizia, richiedendo che i meccanismi di protezione siano progettati in modo da non discriminare le donne con limitata capacità di interagire con le istituzioni. Un sistema conforme deve quindi incorporare meccanismi di compensazione attiva del deficit di tracciabilità istituzionale, anziché trattarlo come mera assenza di rischio *tout court*[89].

In questa prospettiva, il requisito della protezione rafforzata delle vittime in condizione di vulnerabilità multipla assume una rilevanza ulteriore se letto alla luce della nozione di *non-refoulement* algoritmico, nozione qui impiegata in senso analogico e funzionale rispetto al principio classico di *non-refoulement*, ossia senza configurarne un'estensione diretta sul piano del diritto positivo, bensì come categoria utile a descrivere nuove forme di esposizione al rischio derivanti da processi decisionali istituzionali mediati da sistemi algoritmici. In effetti, il concetto di *non-refoulement algoritmico*

[88] S. NOBLE, *Algorithms of Oppression*, cit., 10-30; V. EUBANKS, *Automating Inequality*, cit., pp. 145-175; C. O'NEIL, *Weapons of Math Destruction*, cit., pp. 10-120.

[89] Preme sottolineare come la distinzione tra *lex lata* e *lex ferenda* nella costruzione dei sei requisiti cumulativi non è sempre netta, e chi scrive ne è consapevole. Alcuni requisiti, in particolare quello relativo alla supervisione umana significativa e quello concernente la trasparenza e l'intellegibilità del processo valutativo, trovano un fondamento sufficientemente solido nelle disposizioni vigenti dell'AI Act e del GDPR, nonché nella giurisprudenza della Corte europea dei diritti umani in materia di obblighi positivi. Altri, segnatamente il requisito della *due diligence* tecnologica lungo l'intera filiera di fornitura e quello della protezione rafforzata delle vittime in condizione di vulnerabilità multipla, si collocano in una zona di confine in cui l'interpretazione evolutiva delle fonti esistenti si avvicina alla proposta normativa. Tale zona di confine riflette una caratteristica strutturale del diritto internazionale dei diritti umani, nel quale gli obblighi positivi si sono storicamente consolidati attraverso un processo incrementale di stratificazione interpretativa che ha trasformato proposte dottrinali in prassi convenzionale. I sei requisiti sono dunque presentati come uno standard emergente proprio per segnalare che il loro grado di consolidamento è variabile e che il percorso verso il pieno riconoscimento giurisprudenziale è ancora in corso.

richiede la più attenta elaborazione, poiché non appartiene al lessico consolidato del diritto internazionale ma emerge dall'incrocio tra tre tradizioni giuridiche distinte: il diritto internazionale dei rifugiati, il diritto antidiscriminatorio e la critica giuridico-femminista ai sistemi di intelligenza artificiale. Sotto il primo profilo, com'è noto, il principio di *non-refoulement* nel diritto internazionale dei rifugiati proibisce agli Stati di espellere, respingere o comunque allontanare una persona verso un territorio in cui rischia di subire gravi violazioni dei propri diritti fondamentali. Nella sua accezione classica, sancita dall'art. 33 della Convenzione di Ginevra del 1951 e sviluppata dalla giurisprudenza della Corte EDU nell'interpretazione dell'art. 3 (divieto di tortura, trattamento inumano o degradante) CEDU, il principio opera come limite alla discrezionalità statale in materia migratoria, imponendo una valutazione individualizzata del rischio *ex ante* qualsiasi misura di allontanamento. La dottrina internazionalistica ha progressivamente esteso il perimetro applicativo del principio sino a ricomprendere il *refoulement* indiretto o per interposta autorità che si verifica quando uno Stato trasferisce una persona (richiedente asilo o migrante) verso un Paese terzo (Paese intermedio) che, a sua volta, la espelle o la invia verso un altro Stato in cui la sua vita o la sua libertà sarebbero minacciate (Paese di origine o altro territorio insicuro)[90]. La nozione di *non-refoulement* algoritmico si situa in questa linea di sviluppo estensivo benché operi su un piano strutturalmente diverso, non riguardando la decisione di espellere o trattenere una persona, bensì la decisione di proteggerla o non proteggerla dalla violenza domestica. L'analogia con il *non-refoulement* classico nasce dal fatto che se il *refoulement* tradizionale espone la vittima a una situazione di pericolo determinata dall'azione statale, il *refoulement* algoritmico espone la vittima a una situazione di pericolo determinata dall'omissione qualificata dello Stato, che delega a un sistema predittivo la valutazione del rischio e accetta acriticamente l'*output* discriminatorio di tale sistema. In entrambi i casi, il danno per la vittima deriva dall'effetto prodotto da un meccanismo istituzionale che lo Stato ha adottato o tollerato.

La dimensione specifica che la nozione illumina è quella intersezionale tenuto conto del fatto che le donne migranti, e in particolare le donne in condizione di irregolarità amministrativa, sono sistematicamente classificate a rischio inferiore dai sistemi predittivi non a causa di una loro minore esposizione alla violenza, posto che, viceversa, i dati empirici indicano semmai il contrario, come documentato da tempo risalente dalla letteratura sull'intersezionalità tra violenza di genere e condizione migratoria[91], ma a causa della loro minore interazione con le istituzioni, che si traduce in un deficit di tracce documentali sui cui il modello costruisce il punteggio di rischio. La *semi-legalità* della

[90] *Ex plurimis* sia consentito il rinvio a A. LATINO, *The Principle of Non-Refoulement Between International Law and European Union Law*, in P. ACCONCI, D. DONAT-CATTIN, A. MARCHESI, G. PALMISANO, V. SANTORI (edited by), *International Law and the Protection of Humanity Essays in honor of Judge Flavia Lattanzi*, Leiden, 2016, p. 131 ss.

[91] K. CRENSHAW, *Mapping the Margins*, cit., pp. 1243-1252.

condizione migratoria irregolare[92]; diventa così, nel linguaggio del modello predittivo, un *proxy* di assenza di rischio, con effetti di esclusione sistematica dalla protezione che il diritto internazionale vieta espressamente. La Corte IDH in alcune sue pronunce, *ex plurimis*, *González y otras* («*Campo Algodonero*»)[93], *López* c. *Honduras*[94] e *Véliz Franco y otros* c. *Guatemala*[95], ha affermato che la prevedibilità del rischio può derivare anche

[92] A. KUBAL, *Conceptualizing Semi-Legality in Migration Research*, in *Journal of Ethnic and Migration Studies*, 2013, pp. 555-573.

[93] Corte interamericana dei diritti umani, *González y otras («Campo Algodonero»)* c. *México*, Serie C núm. 205, 16 de noviembre de 2009, reperibile *online*. La vicenda oggetto del ricorso riguarda il ritrovamento, nel 2001, dei corpi di Claudia Ivette González, Esmeralda Herrera Monreal e Laura Berenice Ramos Monárrez a Ciudad Juárez, nello Stato di Chihuahua, al confine settentrionale del Messico con gli Stati Uniti, in un contesto già segnato da un diffuso fenomeno di sparizioni e femminicidi. Benché la scomparsa delle tre giovani fosse stata subito denunciata dai familiari, le autorità avevano omesso di attivare tempestivamente le ricerche e avevano condotto indagini frammentarie e inefficaci, senza, a monte, adottare misure adeguate a prevenirne la morte. La Corte IDH ha stabilito che, in presenza di un contesto strutturale di femminicidio, lo Stato ha un obbligo rafforzato di prevenzione, anche senza un rischio individuale specifico già formalizzato.

[94] Corte interamericana dei diritti umani, *Luna López* c. *Honduras*, Serie C n. 269, 21 de abril 2012, reperibile *online*. Il caso trae origine dall'assassinio di Carlos Antonio Luna López, difensore ambientalista e consigliere comunale della *Corporación Municipal de Catacamas*, nel dipartimento di Olancho, Honduras, avvenuto il 18 maggio 1998. Luna López era una figura di rilievo nel movimento per la protezione delle risorse naturali locali, in un Paese in cui le riforme legislative degli anni Novanta – in particolare la *Ley de Modernización Agrícola* e la nuova legge mineraria del 1998 – avevano alimentato tensioni crescenti tra comunità locali, imprese estrattive e interessi fondiari. Nei mesi precedenti la morte, egli aveva sporto denuncia presso il Ministerio Público contro un imprenditore del settore forestale che lo aveva minacciato di morte a causa delle sue denunce sullo sfruttamento illegale del legname. Il Pubblico Ministero non aprì alcun fascicolo formale, qualificando l'episodio come irrilevante dopo una riunione informale. Le autorità non adottarono alcuna misura effettiva di protezione. Alcune settimane dopo, Luna López fu assassinato. La Corte IDH dichiarò la responsabilità internazionale dello Stato per la violazione dell'obbligo di garantire il diritto alla vita, accertando che le autorità erano a conoscenza del rischio reale e immediato per la vita di Luna López e non avevano adottato misure efficaci di protezione. La Corte rilevò altresì la violazione dell'integrità personale dei familiari della vittima. Il caso si inscrive in un contesto strutturale nel quale, tra il 1991 e il 2011, si erano registrate in Honduras almeno sedici uccisioni di difensori ambientalisti: la Corte valorizzò questa dimensione sistemica per affermare che la prevedibilità del rischio può derivare non solo da segnali individuali specifici, ma dall'esistenza stessa di un *pattern* sistemico di violenza. È in questa dimensione strutturale – e non nella specifica pertinenza alla violenza di genere, che il caso non riguarda – che il precedente viene qui richiamato, a sostegno dell'argomentazione secondo cui uno Stato che continui ad avvalersi di strumenti istituzionali sistematicamente incapaci di intercettare il rischio per determinate categorie vulnerabili non può invocare la mancanza di prevedibilità del danno, essendo tale prevedibilità già inscritta nelle caratteristiche strutturali del contesto.

[95] Corte interamericana dei diritti umani*, Véliz Franco y otros* c. *Guatemala*, Serie C núm. 277, 19 de mayo de 2014, reperibile *online*. La fattispecie sottoposta al vaglio della Corte IDH prende le mosse dall'uccisione della giovane María Isabel Véliz Franco, scomparsa nel 2001 e successivamente ritrovata senza vita in Guatemala, in un contesto caratterizzato da un diffuso fenomeno di violenza contro le donne e da elevati livelli di impunità. Nonostante la denuncia tempestiva della scomparsa da parte della madre, le autorità non avevano attivato ricerche immediate né adottato misure adeguate per localizzare la vittima, giustificando l'inerzia con stereotipi di genere (in particolare, suggerendo che la ragazza fosse scappata volontariamente o avesse un comportamento «immorale»). Le indagini successive sono state tardive, carenti e non hanno identificato i responsabili. Dinanzi alla Corte IDH, i familiari hanno lamentato la responsabilità dello Stato per la mancata prevenzione e per le gravi carenze investigative. La Corte ha accertato la violazione dei diritti alla vita, all'integrità personale, alle garanzie giudiziarie e alla protezione giudiziaria, rilevando che l'inerzia delle autorità si inseriva in un contesto strutturale di violenza di genere. Essa ha ribadito che, in presenza di tale contesto, lo Stato è tenuto a un obbligo rafforzato di prevenzione e a una risposta investigativa tempestiva ed efficace.

da un contesto strutturale di violenza sistemica, con la conseguenza che, applicato ai sistemi predittivi, uno Stato che continui a utilizzare modelli che sottostimano sistematicamente il rischio per le donne migranti non può invocare la mancanza di prevedibilità del danno, poiché la discriminazione algoritmica costituisce parte integrante di quel contesto che esso è tenuto a correggere.

La nozione di *non-refoulement* algoritmico consente altresì di dare contenuto specifico all'obbligo di *due diligence* statale in materia di violenza di genere in rapporto alle tecnologie algoritmiche, sebbene, ai fini della qualificazione giuridica, resti tuttavia necessario dimostrare un nesso sufficientemente diretto tra l'adozione o l'uso del sistema e l'inadempimento degli obblighi positivi di protezione da parte dello Stato. In tale prospettiva, l'obbligo di *due diligence* comprende anche il divieto per lo Stato di adottare misure che, pur non intenzionalmente discriminatorie, producano effetti di esclusione sistematica di gruppi vulnerabili dall'accesso alla protezione. La nozione di *non-refoulement* algoritmico consente di estendere tale principio ai sistemi predittivi, evidenziando come l'impiego di tecnologie che classificano sistematicamente a rischio inferiore le donne migranti irregolari integri una violazione di siffatto obbligo di *due diligence*. Tale violazione si traduce in una forma di responsabilità internazionale dello Stato, nella misura in cui l'adozione e l'uso di strumenti algoritmici idonei a generare effetti discriminatori comportano un inadempimento degli obblighi positivi di protezione e di non discriminazione, indipendentemente dall'intenzionalità del legislatore o del fornitore tecnologico.

Sul piano del diritto positivo, la nozione di *non-refoulement* algoritmico può essere ancorata a un insieme convergente di disposizioni. In primo luogo, la CEDAW (art. 2, lett. d-e) impone agli Stati sia di astenersi da qualsiasi atto o pratica discriminatoria nei confronti delle donne, sia di garantire una protezione effettiva contro atti discriminatori, anche quando derivanti dal funzionamento di apparati istituzionali. In secondo luogo, la Convenzione di Istanbul (artt. 4 e 22) vieta ogni forma di discriminazione nell'accesso alla protezione e richiede che i servizi di supporto siano accessibili a tutte le vittime, indipendentemente, tra l'altro, dalla condizione migratoria. In ambito europeo, il GDPR (art. 22) limita il ricorso a decisioni basate unicamente su trattamenti automatizzati che producano effetti giuridici o incidano in modo analogo sulla sfera dell'interessato; in tale contesto, la nozione di «effetto significativo» può essere interpretata in senso ampio, includendo anche le decisioni relative all'attivazione o al diniego di misure di protezione. Infine, l'AI Act, ai sensi dell'allegato III, punto 6, qualifica come ad alto rischio i sistemi di intelligenza artificiale utilizzati per la valutazione del rischio relativo a comportamenti illeciti, categoria che ricomprende anche gli strumenti impiegati nella valutazione del rischio di recidiva nella violenza domestica.

La nozione di *non-refoulement* algoritmico si configura dunque come una categoria giuridica emergente che consente di dare forma normativa a una specifica modalità di

fallimento della protezione internazionale, riconducibile all'impiego di sistemi decisionali incapaci di intercettare il rischio per determinate categorie di vittime. Essa rende visibile uno scarto strutturale tra obbligo giuridico e capacità istituzionale, nel quale la tutela prevista dal diritto internazionale permane sul piano formale, ma non si traduce in accesso effettivo alla protezione. Tale scarto, come si è cercato di porre in luce, non deriva da una mera inefficienza operativa, bensì dal modo in cui il rischio viene costruito, classificato e distribuito all'interno dei dispositivi istituzionali che incorporano criteri selettivi fondati su dati incompleti, storicamente distorti o incapaci di rappresentare le condizioni di vulnerabilità.

In questa prospettiva, la valutazione del rischio assume una funzione costitutiva della protezione stessa: le scelte relative alla selezione dei dati, alla definizione degli indicatori e alla ponderazione dei fattori rilevanti determinano chi viene riconosciuto come soggetto meritevole di tutela e chi, invece, resta ai margini del sistema. L'architettura algoritmica diviene così un luogo di produzione normativa implicita, in cui si definiscono, attraverso logiche tecniche, i confini dell'accesso ai diritti. Quando tali logiche si fondano su rappresentazioni parziali o distorte, il risultato è una distribuzione asimmetrica della protezione, che colpisce in modo particolare le categorie meno visibili nei dati, come le donne migranti in condizione di marginalità istituzionale.

L'insieme dei requisiti cumulativi individuati consente di delineare uno standard emergente di conformità algoritmica nel diritto internazionale dei diritti umani, già riconoscibile sul piano interpretativo e destinato (quantomeno, lo si auspica) a consolidarsi attraverso l'evoluzione giurisprudenziale. In questo quadro, l'impiego di sistemi che generano effetti discriminatori sistematici rileva direttamente ai fini della responsabilità internazionale dello Stato, in quanto l'esposizione al rischio deriva dal funzionamento stesso del dispositivo adottato. La violazione si radica così nel rapporto tra struttura del sistema e distribuzione degli esiti, investendo non solo la fase dell'intervento, ma quella, precedente e decisiva, della qualificazione del rischio.

È in questo spazio che la nozione di *non-refoulement* algoritmico converge con la categoria della violenza algoritmica strutturale. La selezione implicita delle vittime meritevoli di protezione, combinata con l'esclusione di quelle non riconosciute dai modelli predittivi, dà luogo a una forma di produzione istituzionale del danno, nella quale la violenza si manifesta come effetto sistemico di un assetto decisionale che organizza, distribuisce e legittima l'accesso differenziato alla tutela. In tale configurazione, la violenza, lungi dall'esaurirsi nell'evento lesivo, si inscrive nelle condizioni strutturali che ne rendono possibile la reiterazione, collocandosi al crocevia tra tecnologia, potere pubblico e diseguaglianze sociali.

6. Conclusioni critiche e prospettive *de lege ferenda*.

Le considerazioni che seguono si collocano in parte sul piano del diritto vigente e in parte in una prospettiva *de lege ferenda*, nella consapevolezza che alcune delle soluzioni prospettate richiedono un ulteriore sviluppo normativo. La distinzione tra le due dimensioni richiede una precisazione metodologica: laddove il contributo ricava obblighi positivi in via interpretativa dagli strumenti vigenti (CEDAW, Convenzione di Istanbul, AI Act, GDPR), tale operazione si muove sul piano della *lex lata* – ossia del diritto positivo attualmente in vigore – pur trattandosi di un'interpretazione evolutiva e non ancora consolidata nella prassi delle Corti internazionali. Laddove invece il contributo propone strumenti normativi nuovi (un protocollo opzionale alla CEDAW, un regime di responsabilità rafforzato per i fornitori tecnologici, un fondo di compensazione), esso si colloca esplicitamente sul piano della *lex ferenda*, pur nella consapevolezza che tale distinzione non è sempre netta in quanto alcuni degli obblighi qui proposti come *de lege ferenda* potrebbero già essere ricavabili, in via interpretativa, dal diritto vigente, e, ad avviso di chi scrive, il loro progressivo riconoscimento nella giurisprudenza internazionale è tanto prevedibile quanto auspicabile.

In particolare, al fine di tratteggiare delle osservazioni conclusive, sembra opportuno innanzitutto evidenziare come l'analisi degli strumenti normativi e della prassi delle Corti di EDU e IDH sviluppata in questo contributo converga verso tre principî che definiscono gli standard minimi di compatibilità di un sistema predittivo con il diritto internazionale dei diritti umani.

Il primo principio è quello della non-delega: lo Stato non può delegare l'adempimento del proprio obbligo di valutazione del rischio a un sistema automatizzato che sistematicamente non adempia quell'obbligo nei confronti di determinate categorie di vittime, sicché, da un lato, la valutazione del rischio deve essere autonoma rispetto ai *pattern* storici del *dataset* e, dall'altro, deve essere capace di adeguarsi al rischio reale della vittima specifica attraverso la supervisione umana qualificata.

Il secondo principio stigmatizza l'indifferenza proibita in quanto la discriminazione istituzionale implicita nei confronti delle categorie di donne sistematicamente sotto-rappresentate nei sistemi predittivi è una forma di passività istituzionale che la Corte EDU ha qualificato come violazione autonoma dell'art. 14 CEDU e la Corte IDH come violazione degli artt. 7(*b*) e 7(*c*) della Convenzione di Belém do Pará che impongono agli Stati obblighi di *due diligence*, richiedendo, da un lato, l'adozione di misure efficaci per prevenire, investigare e sanzionare la violenza contro le donne e, dall'altro, l'integrazione di strumenti normativi e istituzionali idonei a eliminarla.

Il terzo principio riguarda l'aggiornamento contestuale, posto che i sistemi predittivi che non includono indicatori di violenza digitale e che non sono in grado di adattare la valutazione al contesto strutturale di violenza sistemica di genere documentato non soddisfano l'obbligo convenzionale di valutazione del rischio in modo adeguato e proporzionato alla realtà della violenza domestica nell'era digitale.

Il percorso argomentativo sviluppato in queste pagine si colloca lungo tre decenni di evoluzione normativa e si misura con vicende concrete in cui il sistema ha mostrato limiti riconoscibili e, in molti casi, evitabili[96]. Tale evoluzione ha condotto alla progressiva elaborazione, da parte delle Corti di Strasburgo e di San José, di un *corpus* di obblighi positivi che ha significativamente ristretto gli spazi per gli Stati di invocare la dimensione privata della violenza domestica quale causa di esonero dalla responsabilità. Questo assetto, tuttavia, continua a operare prevalentemente *ex post* nella dimensione della riparazione, intervenendo a valle della violazione e spesso a fronte di esiti già irreversibili. Nel contesto dell'era algoritmica, la questione si sposta *ex ante* sul terreno della prevenzione in quanto la capacità del diritto internazionale di anticipare il rischio diviene centrale rispetto a sistemi tecnologici che operano con livelli di velocità, scala e complessità difficilmente compatibili con i tradizionali meccanismi di accertamento della responsabilità. In tale prospettiva, l'effettività della tutela richiede un ripensamento degli strumenti giuridici in chiave anticipatoria, orientato a intercettare e neutralizzare i fattori di rischio prima che si traducano in violazioni dei diritti fondamentali.

Il quadro tratteggiato consente peraltro di avanzare tre critiche al sistema vigente.

La prima criticità si colloca sul piano epistemologico e riguarda il modo in cui il rischio viene costruito e reso conoscibile nei sistemi di valutazione fondati su dati storici. In tali modelli, il rischio coincide con ciò che è stato registrato nel passato istituzionale, con il risultato che la previsione riflette la distribuzione delle denunce e delle interazioni con le istituzioni. L'utilizzo dei dati di denuncia come *proxy* del rischio rappresenta, in questa prospettiva, una scelta intrinsecamente politica, che eleva a parametro oggettivo il prodotto di pratiche istituzionali selettive e storicamente situate. La tracciabilità istituzionale diventa così il criterio dominante di visibilità del rischio, mentre le esperienze che sfuggono ai circuiti formali restano escluse dalla rappresentazione. Una volta incorporata nei sistemi predittivi, tale impostazione contribuisce a stabilizzare e legittimare le disuguaglianze esistenti, traducendo in previsione tecnica ciò che è il risultato di rapporti di potere. Il sapere prodotto dall'algoritmo assume la forma di una oggettività apparente, che maschera la natura selettiva e normativa delle scelte sottostanti e consolida, in chiave anticipatoria, i pregiudizi che il diritto internazionale è chiamato a disarticolare[97].

La seconda criticità è di natura democratica e riguarda la struttura dei processi decisionali che definiscono il funzionamento dei sistemi di valutazione del rischio. Le scelte che ne determinano l'architettura – selezione delle variabili, criteri di ponderazione, soglie di attivazione – sono assunte in spazi tecnici e amministrativi chiusi alla

[96] C. MACKINNON, *Regulating AI: A Feminist Roadmap*, in *Harvard Law Review*, 2025, pp. 1-42; H. CHARLESWORTH, *Not Waving but Drowning: Gender Mainstreaming and Human Rights in the United Nations*, in *Harvard Human Rights Journal*, 2005, pp. 1-18.

[97] S. NOBLE, *Algorithms of Oppression*, cit., pp. 1-50; V. EUBANKS, *Automating Inequality*, cit., pp. 175-210.

partecipazione delle donne che tali sistemi sono destinati a proteggere, lacuna già evidenziata nella letteratura sulla partecipazione delle vittime nei processi di *risk assessment*[98]. In questo modo, categorie, indicatori e parametri incorporano rappresentazioni della violenza, della vittima e del rischio costruite senza confronto con le esperienze vissute, trasformando la progettazione tecnica in un esercizio di potere istituzionale. L'esclusione sistematica delle vittime produce un duplice effetto: restringe il perimetro di ciò che è riconosciuto come rischio e consolida gerarchie di visibilità già presenti nei dati istituzionali, sicché le forme di violenza meno documentate, meno denunciate o più difficili da classificare vengono espulse dal campo della rilevanza giuridica già nella fase di progettazione, con ricadute dirette sull'accesso alla protezione. Il dispositivo finisce così per selezionare le vittime prima ancora di attivare la tutela, organizzando in modo differenziale la distribuzione delle misure protettive. In tale configurazione, la dimensione democratica investe la costruzione stessa del rischio in quanto l'assenza di partecipazione incide sui criteri attraverso cui il pericolo viene definito, misurato e reso operativo. La progettazione algoritmica si configura come un luogo di produzione normativa implicita, nel quale si stabiliscono le condizioni di accesso alla protezione, con effetti che si riflettono direttamente sulla capacità dello Stato di adempiere agli obblighi di prevenzione e di non discriminazione[99].

La terza criticità si colloca nella prospettiva femminista e riguarda la costruzione implicita della vittima all'interno dei sistemi di valutazione del rischio. Il soggetto che tali dispositivi rendono visibile coincide con un modello storicamente prodotto e socialmente legittimato di vittima «credibile», conforme a parametri di rispettabilità e vulnerabilità inscritti in un ordine patriarcale – madre, preferibilmente giovane, economicamente dipendente e appartenente alla comunità nazionale. In tal modo, la selezione operata dal sistema contribuisce a definire il rischio stesso, stabilendo quali esperienze di violenza assumano rilevanza giuridica e quali restino ai margini[100]. In questo contesto, la crescente enfasi sulla regolazione dei sistemi predittivi orienta il discorso giuridico verso una razionalità tecnico-gestionale, nella quale la violenza di genere viene trattata come problema di ottimizzazione delle risposte istituzionali. Tale impostazione produce un effetto di depoliticizzazione, riducendo la visibilità delle condizioni strutturali e delle relazioni di potere che sostengono la violenza e favorendo una lettura securitaria focalizzata sugli strumenti di intervento[101]. In questa prospettiva, la *governance* algoritmica del rischio contribuisce a stabilizzare le gerarchie esistenti,

[98] S. LEWIS, *Participation of Survivors in Risk Assessment*, in *Violence against Women*, 2023, pp. 1245-1263.

[99] C. D'IGNAZIO, L. F. KLEIN, *Data Feminism,* Cambridge-MA, 2020, pp. 123-167.

[100] H. CHARLESWORTH, C. CHINKIN, *The Boundaries of International Law*, cit. pp. 1-18; K. CRENSHAW, *Mapping the Margins*, cit., pp. 1241-1252; C. MACKINNON, *Toward a Feminist Theory of the State*, Cambridge-MA, 1989, pp. 161-170.

[101] H. CHARLESWORTH, *Not Waving but Drowning*, cit., pp. 1-18.

inscrivendo nella logica del sistema una definizione selettiva della vittima e rafforzando la riproduzione delle disuguaglianze legittimando l'approccio securitario a scapito di quello trasformativo, più in linea con le istanze del diritto internazionale.

Di fatto, tre lacune strutturali attraversano il quadro normativo vigente: *in primis*, la frammentazione in quanto CEDAW, Convenzione di Istanbul, AI Act, GDPR e DSA affrontano aspetti parziali senza un quadro di riferimento unitario; in secondo luogo, l'asincronia, posto che l'AI Act è entrato in vigore circa quindici anni dopo la comparsa dei sistemi predittivi, quando questi erano già radicati nella prassi istituzionale; infine, l'asimmetria di *accountability* fra gli Stati, passibili di condanna davanti alle corti internazionali, e le imprese tecnologiche, che sviluppano sistemi discriminatori sono soggette principalmente a standard volontari.

In tale prospettiva, il principale rimedio *de lege ferenda* consiste nell'elaborazione e nell'adozione di un protocollo opzionale alla CEDAW dedicato alla violenza digitale di genere, quale strumento idoneo a colmare la frammentazione normativa esistente e a ridefinire in chiave trasformativa gli obblighi statali nell'ecosistema digitale. Un simile protocollo si configurerebbe come sede, a un tempo, di codificazione e di sviluppo progressivo di uno standard internazionale di conformità algoritmica fondato sui diritti delle persone umane, capace di integrare le diverse dimensioni normative attualmente disperse.

Sotto il profilo sostanziale, esso dovrebbe, in primo luogo, consacrare una definizione giuridica di violenza digitale di genere comprensiva della dimensione della violenza algoritmica strutturale, riconoscendone la natura istituzionale e sistemica. In secondo luogo, dovrebbe introdurre obblighi vincolanti di progettazione, sviluppo e utilizzo responsabile dei sistemi predittivi, imponendo la validazione scientifica indipendente, l'obbligatorietà di *gender audit* periodici e l'accesso effettivo ai dati rilevanti, accompagnati dalla pubblicazione trasparente dei risultati e da meccanismi automatici di sospensione o ritiro dei sistemi non conformi.

Il protocollo dovrebbe inoltre riconoscere alle vittime un diritto sostanziale alla comprensione e alla contestazione delle decisioni algoritmiche, traducendo in termini operativi il diritto all'informazione sulla logica del trattamento e rafforzandolo attraverso obblighi di intellegibilità effettiva e verificabile. A tali garanzie dovrebbe affiancarsi l'istituzione di meccanismi di ricorso accessibili, rapidi ed effettivi, idonei a consentire (più che) la riparazione del danno e inoltre (soprattutto) la revisione tempestiva delle valutazioni di rischio.

In una prospettiva più ampia, il protocollo dovrebbe ridefinire il contenuto della *due diligence* statale alla luce delle tecnologie predittive, qualificando l'adozione e l'uso di sistemi algoritmici come ambito specifico di responsabilità internazionale e imponendo agli Stati un obbligo di prevenzione *ex ante* dei rischi discriminatori sistemici. In questo quadro, l'obbligo di supervisione umana qualificata assumerebbe una dimensione

sostanziale, implicando che il decisore umano disponga di accesso a tutte le informazioni rilevanti, incluse quelle non elaborate dal sistema, e possieda competenze specialistiche nella valutazione del rischio nella violenza domestica, unitamente a una legittimazione istituzionale effettiva a discostarsi dall'*output* algoritmico.

Infine, un assetto normativo coerente richiederebbe l'introduzione di un regime di responsabilità rafforzato per gli attori coinvolti nello sviluppo e nell'uso di tali sistemi, comprensivo di una presunzione di responsabilità in caso di mancato adeguamento agli standard di *audit* e della creazione di un fondo di compensazione per le vittime di violenza digitale, alimentato da contributi obbligatori. In tal modo, il protocollo contribuirebbe a spostare stabilmente il baricentro della tutela dalla risposta alla violazione verso la costruzione di condizioni istituzionali idonee a prevenirla. Tale proposta richiede tuttavia un'analisi più articolata sul piano del diritto internazionale della responsabilità degli Stati, che il presente contributo si limita ad abbozzare nelle sue linee essenziali, con l'avvertenza che una trattazione esaustiva richiederebbe uno spazio autonomo. In effetti, sul piano del diritto positivo vigente, la responsabilità internazionale dello Stato per l'impiego di sistemi predittivi discriminatori può essere ricostruita attraverso tre diversi vettori normativi, tra loro cumulabili. Il primo è il classico regime di responsabilità per fatto illecito internazionale, disciplinato dal Progetto di Articoli sulla responsabilità degli Stati per atti internazionalmente illeciti (*Articles on Responsibility of States for Internationally Wrongful Acts* – ARSIWA, 2001)[102], elaborato dalla Commissione del diritto internazionale, il cui art. 2 richiede la sussistenza di un comportamento (azione od omissione) attribuibile allo Stato ai sensi del diritto internazionale e che costituisca violazione di un obbligo internazionale. Nel contesto dei sistemi predittivi di violenza domestica, l'adozione e l'impiego continuato di uno strumento che produce sistematicamente effetti discriminatori nei confronti di determinate categorie di donne – in violazione degli obblighi derivanti dalla CEDAW, dalla Convenzione di Istanbul e dal divieto di non discriminazione – integra, ad avviso di chi scrive, gli elementi dell'illecito di cui all'art. 2 ARSIWA. La sussistenza del nesso di attribuibilità non pone particolari difficoltà, in quanto il sistema predittivo è adottato, finanziato e utilizzato da organi dello Stato (forze dell'ordine, servizi di protezione), il cui comportamento è attribuibile allo Stato ai sensi dell'art. 4 ARSIWA. Il secondo vettore è la responsabilità per omissione qualificata, che assume rilevanza quando lo Stato, pur disponendo degli strumenti giuridici e delle risorse tecniche per correggere i *bias* discriminatori documentati, omette di farlo. La giurisprudenza della Corte EDU, a partire da *Opuz* c. *Turkey*, ha stabilito che l'inerzia istituzionale dinanzi a un rischio conoscibile e prevenibile integra una violazione autonoma degli obblighi positivi convenzionali, indipendentemente dall'intenzione

[102] Articles on Responsibility of States for Internationally Wrongful Acts 2001, text adopted by the Commission at its fifty-third session, in 2001, and submitted to the General Assembly, annex to General Assembly resolution 56/83 of 12 December 2001, and corrected by document A/56/49(Vol. I)/Corr.4.

discriminatoria. Il terzo vettore è la *due diligence* tecnologica, elaborata in via di sviluppo progressivo dalla prassi dei Comitati convenzionali e dai Principi Guida ONU su imprese e diritti umani, che impone allo Stato non solo di astenersi dall'adottare sistemi discriminatori, ma anche di sottoporre i fornitori tecnologici a obblighi di verifica preventiva, *audit* periodici e rendicontazione pubblica. Sul piano *de lege ferenda*, la proposta di introdurre una presunzione di responsabilità in caso di mancato adeguamento agli standard di *audit* trova precedenti parziali nel regime dei sistemi di IA ad alto rischio dell'AI Act (artt. 9 e 17), che già invertono parzialmente l'onere della prova imponendo al fornitore la documentazione della conformità. L'estensione di tale logica al piano internazionale comporterebbe che, in presenza di documentazione indipendente di effetti discriminatori sistemici (come gli *audit* di *Eticas Foundation* su VioGén), lo Stato sia tenuto a dimostrare di aver adottato misure correttive adeguate, con onere della prova rovesciato rispetto al modello classico di responsabilità per fatto illecito. Quanto alla proposta di un fondo di compensazione per le vittime di violenza digitale, essa trova precedenti nel Fondo volontario delle Nazioni Unite per le vittime di tortura[103] e nell'obbligo imposto a tutti gli Stati membri dell'Ue dalla menzionata direttiva 2004/80 di istituire sistemi nazionali di indennizzo per le vittime di reati intenzionali violenti[104], sebbene nessuno di questi strumenti abbia carattere specificamente orientato alla violenza algoritmica di genere. La fattibilità politico-giuridica di tale proposta dipende dalla disponibilità degli Stati a istituire un meccanismo a contribuzione obbligatoria – sul modello dei fondi ambientali internazionali – nel quale i fornitori tecnologici privati partecipino proporzionalmente ai profitti derivanti dall'uso istituzionale dei loro sistemi predittivi. L'ostacolo principale non è di natura giuridica bensì politica: la resistenza degli Stati a vincolare strutturalmente fondi pubblici a obblighi di compensazione nei confronti di vittime di discriminazione algoritmica. Ciononostante, la proposta conserva valore come principio orientativo del sistema, coerente con la logica trasformativa che attraversa l'intero impianto argomentativo.

Gli ulteriori spunti che emergono dall'analisi condotta, ai fini della configurazione di un diritto internazionale capace di incidere in modo effettivamente trasformativo sulla violenza algoritmica di genere, possono essere sistematizzati lungo tre direttrici principali, corrispondenti ad altrettanti nuclei critici.

In primo luogo, si impone la questione dell'*accountability*, intesa non solamente come responsabilità *ex post*, bensì come costruzione di un sistema giuridico di obblighi positivi, preventivi e riparatori. In tale prospettiva, appare necessario superare l'attuale frammentazione normativa e la prevalenza di strumenti di *soft law* o di autoregolazione,

[103] United Nations Voluntary Fund for Victims of Torture, AG Res. 36/151, 101st plenary meeting 16 December 1981.

[104] In Italia, questa tutela è garantita dal Fondo di rotazione per la solidarietà alle vittime dei reati intenzionali violenti gestito dal Ministero dell'Interno e amministrato dalla Consap, cfr. il sito https://www.consap.it/fondo-vittime-mafia-estorsione-usura-reati-violenti-e-orfani-per-crimini domestici/.

che si sono rivelati strutturalmente inadeguati a fronteggiare fenomeni di violenza algoritmica caratterizzati da opacità tecnica, asimmetrie informative e transnazionalità. Ne deriva l'esigenza di elaborare strumenti giuridicamente vincolanti che impongano agli Stati obblighi di *due diligence* rafforzata nella prevenzione, investigazione e sanzione delle forme di violenza digitale di genere, nonché obblighi diretti, o, quantomeno, mediati, alle piattaforme e agli attori privati che progettano, sviluppano e gestiscono sistemi algoritmici. Ciò implica, tra l'altro, l'introduzione di standard internazionali in materia di trasparenza e sottoponibilità a procedure di controllo e revisione (*audit*), non discriminazione e accesso effettivo ai rimedi, nonché meccanismi di controllo indipendenti e sanzioni proporzionate ed effettive.

In secondo luogo, si rende imprescindibile una prospettiva di trasformazione strutturale, che consenta di affrontare le radici sistemiche della violenza algoritmica di genere. Tale violenza, infatti, deve essere letta come espressione e riproduzione, in ambiente digitale, di diseguaglianze preesistenti di natura economica, giuridica e socioculturale. In questo senso, un diritto internazionale realmente trasformativo dovrebbe imporre agli Stati obblighi di politica pubblica orientati alla rimozione di tali diseguaglianze, andando oltre un approccio meramente securitario o repressivo. Ciò include, *inter alia*, politiche di redistribuzione economica, il rafforzamento dell'autonomia giuridica e materiale delle donne, l'educazione ai diritti digitali e alla parità di genere, nonché la promozione di modelli culturali non discriminatori. In tale ottica, la regolazione degli algoritmi si configura come parte integrante di una più ampia agenda di giustizia sociale e di uguaglianza sostanziale.

In terzo luogo, emerge con forza la dimensione della partecipazione, quale condizione di legittimità e di efficacia delle politiche e delle norme volte a contrastare la violenza algoritmica di genere. Le donne devono essere riconosciute come soggetti politici titolari del diritto di partecipare in modo pieno, effettivo e informato ai processi decisionali che riguardano la progettazione, l'implementazione e la *governance* dei sistemi algoritmici. Ciò implica l'istituzionalizzazione di meccanismi partecipativi inclusivi e intersezionali, capaci di dare voce alle diverse esperienze di vulnerabilità e di discriminazione, nonché il riconoscimento del sapere situato come risorsa epistemica essenziale per l'elaborazione di soluzioni normative adeguate. In questa prospettiva, la partecipazione si configura come elemento costitutivo di un paradigma di *governance* democratica della tecnologia, fondato sull'autodeterminazione e sulla co-produzione delle condizioni di protezione.

Nel loro insieme, tali tre livelli – *accountability*, trasformazione strutturale e partecipazione – delineano i contorni di un diritto internazionale che, lungi dal limitarsi a reagire agli effetti della violenza algoritmica, aspiri a prevenirla e a disarticolarne le cause profonde, ponendosi come strumento di emancipazione e di giustizia di genere nell'ecosistema digitale globale.

Se si volesse individuare una figura concettuale idonea a restituire l'approdo teorico e normativo di questo lavoro, essa potrebbe essere quella della bilancia intesa come dispositivo di allocazione e regolazione dei diritti, intrinsecamente esposto a un processo continuo di riassetto e taratura. Nell'era algoritmica, tale dispositivo si confronta con la tensione tra l'integrazione di tecnologie predittive nei processi decisionali e la necessità di preservare standard di eguaglianza sostanziale e non discriminazione, a fronte del rischio strutturale che tali tecnologie incorporino e riproducano asimmetrie preesistenti. La continua ricerca del punto di equilibrio sollecita una riflessione profonda sul ruolo delle tecnologie nei processi decisionali che incidono sui diritti fondamentali. L'analisi svolta ha mostrato come l'affidamento acritico a sistemi predittivi possa produrre esiti incompatibili con gli obblighi positivi degli Stati in materia di protezione dalla violenza di genere, fino a configurare fenomeni riconducibili alla nozione, qui proposta, di *non-refoulement* algoritmico. In tale prospettiva, l'errore sistemico assume una dimensione qualitativamente diversa: esso si traduce in una forma di esclusione istituzionale dalla protezione, nella quale la vittima resta esposta al rischio in conseguenza dell'accettazione, da parte dello Stato, di *output* algoritmici incapaci di intercettarne la vulnerabilità.

La bilancia diviene così il luogo simbolico in cui si gioca la prevenzione di nuove forme di violenza istituzionale mediate dalla tecnologia. In essa si riflette il passaggio verso un diritto internazionale orientato alla prevenzione *ex ante*, nel quale l'errore algoritmico discriminatorio acquisisce piena rilevanza giuridica e richiede meccanismi di controllo effettivi e tempestivi.

In questo quadro, la rimodulazione richiede un presidio umano qualificato quale elemento intrinseco alla legittimità del processo decisionale. Il giudizio umano, per risultare conforme agli standard internazionali, deve essere informato, responsabile e sensibile al contesto concreto, ossia capace di integrare nel processo decisionale elementi contestuali e intersezionali che sfuggono ai modelli statistici. In tal modo, esso preserva lo spazio del discernimento e impedisce che la razionalità probabilistica si traduca in schemi decisionali rigidi, incapaci di confrontarsi con la complessità dei casi concreti.

In effetti, seppure il codice può essere progressivamente migliorato sotto il profilo dell'equità, della trasparenza e della sottoponibilità a verifica, permane purtuttavia un limite strutturale connesso alla sua intrinseca incapacità di esaurire la complessità delle situazioni in cui sono in gioco l'integrità, la dignità e la vita delle persone. In particolare, nei contesti di violenza algoritmica di genere, l'automatizzazione delle decisioni espone al rischio di nuove forme di invisibilizzazione ed esclusione, che il diritto internazionale è chiamato a riconoscere e prevenire.

Ne consegue che un diritto internazionale autenticamente trasformativo, oltre a disciplinare l'uso delle tecnologie, deve governarne gli effetti sulla distribuzione della tutela e del rischio. In questa prospettiva, l'evocata bilancia assume il significato di un dispositivo chiamato a misurare le conseguenze dell'innovazione tecnologica alla luce

dei principi di eguaglianza sostanziale, non discriminazione e tutela effettiva dei diritti umani. All'interno di tale quadro, la nozione di *non-refoulement* algoritmico emerge come principio-limite dell'ecosistema digitale contemporaneo, volto a impedire che dispositivi concepiti per identificare e prevenire la violenza contribuiscano, attraverso meccanismi di esclusione sistematica, a perpetuarla. La sua rilevanza risiede nella capacità di rendere visibili quelle forme di vulnerabilità che restano occultate dai modelli predittivi e di riaffermare che la protezione dei diritti fondamentali non può essere subordinata alle logiche statistiche della classificazione algoritmica. Solo così la bilancia della protezione internazionale può continuare a svolgere la propria funzione di garanzia, evitando che gli strumenti destinati a prevenire il danno finiscano per concorrere alla sua produzione.

ABSTRACT: Il saggio muove da un paradosso strutturale del diritto internazionale contemporaneo: i sistemi algoritmici di valutazione del rischio adottati dagli Stati per proteggere le vittime di violenza domestica possono trasformarsi in meccanismi istituzionali di produzione del danno che intendevano prevenire. A partire da questa premessa, il contributo elabora due categorie interpretative originali, la «violenza algoritmica di genere» e il «*non-refoulement* algoritmico», e le applica a una comparazione sistematica di quattro modelli europei (VioGén, SARA, DASH, Codice Rosso), condotta alla luce di un quadro normativo multilivello che integra CEDAW, Convenzione di Istanbul, AI Act, GDPR e Principî Guida ONU su Imprese e Diritti Umani. L'analisi, svolta con metodologia femminista-critica, mostra come le tensioni strutturali rilevate in ciascun sistema riflettano una contraddizione irrisolta tra la logica probabilistica dei modelli predittivi e le esigenze di un diritto internazionale che impone valutazioni autonome, non discriminatorie e *victim-centered* del rischio. Il contributo individua sei requisiti cumulativi di conformità algoritmica al diritto internazionale vigente e avanza proposte *de lege ferenda*, tra cui un protocollo opzionale alla CEDAW sulla violenza digitale di genere e un regime vincolante di *gender audit* algoritmico con conseguenze giuridiche per gli Stati inadempienti.

PAROLE CHIAVE: violenza algoritmica di genere; sistemi predittivi di valutazione del rischio; *non-refoulement* algoritmico; CEDAW; Convenzione di Istanbul; AI Act; VioGén; discriminazione intersezionale.

Algorithmic Gender Violence, Predictive Technologies and International Law

ABSTRACT: *This article departs from a structural paradox in contemporary international law: the algorithmic risk assessment systems deployed by States to protect domestic violence victims can themselves become institutional mechanisms producing the very harm they were designed to prevent. From this premise, the article develops two original interpretive categories, «algorithmic gender violence» and «algorithmic non-refoulement», and applies them to a systematic comparison of four European models (VioGén, SARA, DASH and the Italian Codice Rosso), assessed against a multilevel normative framework integrating CEDAW, the Istanbul Convention, the AI Act, the GDPR and the UN Guiding Principles on Business and Human Rights. Drawing on feminist-critical methodology, the analysis demonstrates that the structural tensions identified across each system reflect an unresolved contradiction between the probabilistic logic of predictive models and the demands of an international legal order requiring autonomous, non-discriminatory and victim-centred risk assessments. The*

article identifies six cumulative requirements of algorithmic compliance with existing international law and advances de lege ferenda proposals, including an Optional Protocol to CEDAW on digital gender-based violence and a binding algorithmic gender audit obligation with legal consequences for non-compliant States.

KEYWORDS: algorithmic gender violence; predictive risk assessment; algorithmic non-refoulement; CEDAW; Istanbul Convention; AI Act; VioGén; intersectional discrimination.