IA che suggerisce soluzioni. Il sistema spagnolo "VioGén 2" e il suo impatto sulla libertà di decisione

di *Leonor Sanz Gallardo*

Il contributo analizza criticamente l'impiego dei sistemi di intelligenza artificiale a funzione raccomandativa in contesti giuridici e di sicurezza, soffermandosi sui loro effetti indiretti ma rilevanti sull'autonomia decisionale e sulla tutela dei diritti fondamentali. Muovendo dalla distinzione concettuale tra sistemi di supporto decisionale e decisione automatizzata, l'articolo esamina il sistema spagnolo di valutazione del rischio di violenza di genere "VioGén", con particolare attenzione alla sua evoluzione nella versione "VioGén 2". L'analisi evidenzia come, nella prassi operativa, le raccomandazioni algoritmiche tendano a produrre effetti vincolanti di fatto, attraverso meccanismi di ancoraggio cognitivo, standardizzazione procedurale e opacità metodologica, riducendo l'effettività della supervisione umana. Tali dinamiche incidono sia sulla protezione delle vittime sia sull'ambiente informativo che orienta la decisione giudiziaria. Il lavoro inquadra queste criticità alla luce del diritto europeo e internazionale in materia di intelligenza artificiale e diritti umani, sostenendo la necessità di garanzie di trasparenza, verificabilità e controllo umano.

1. Introduzione: cosa significa "IA che suggerisce soluzioni" nell'ambito giudiziario

Un sistema che formula raccomandazioni algoritmiche è un tipo di intelligenza artificiale che elabora dati mediante un modello matematico o statistico per generare un *output* orientativo, come può essere un punteggio, una categoria, una previsione o un suggerimento. Questo *output* non ha carattere vincolante e non sostituisce la decisione umana; la sua finalità dichiarata è esclusivamente quella di informare chi deve adottare la decisione finale. Questo tipo di sistemi si caratterizza per il fatto di non sostituire il giudizio umano, trasformare i dati in un'interpretazione strutturata della realtà e produrre risultati standardizzati e ripetibili, che consentono di organizzare l'informazione senza eseguire essi stessi alcuna azione né adottare decisioni autonome. A differenza dei sistemi di decisione automatizzata, non attivano misure autonomamente, ma apportano elementi che devono essere valutati da una persona.

Nell'ambito giudiziario, dove la decisione deve essere personale, motivata e frutto di una valutazione critica, questa distinzione risulta essenziale.

Sebbene i sistemi di raccomandazione algoritmica non adottino decisioni né possiedano autorità vincolante, il loro funzionamento produce effetti psicologici, operativi e istituzionali in grado di influenzare, orientare o persino determinare *de facto* la decisione

umana, con un impatto particolare sulla sfera dei diritti fondamentali. Elaborando dati e generando raccomandazioni, tali sistemi possono introdurre errori, *bias* o distorsioni che incidono su diritti come l'uguaglianza, la non discriminazione, la *privacy* o l'accesso alla giustizia, come segnalato sia dall'Alto Commissariato delle Nazioni Unite per i Diritti umani sia dall'Agenzia dell'Unione Europea per i Diritti fondamentali (FRA)[1].

La convergenza tra le principali fonti internazionali, tra cui la Convenzione quadro del Consiglio d'Europa sull'Intelligenza Artificiale (2024)[2], la Raccomandazione dell'UNESCO sull'Etica dell'IA (2021), le Linee guida etiche dell'Unione interparlamentare sull'autonomia umana e la supervisione, nonché la dottrina emergente sul diritto al controllo umano, mostra che la regolazione dei sistemi di IA deve partire dal rispetto e dalla garanzia dei diritti umani, imponendo limiti e salvaguardie chiare. Tutti questi strumenti coincidono nel ritenere che, quando tali sistemi vengono utilizzati in contesti pubblici sensibili, debbano operare con trasparenza, auditabilità, comprensibilità metodologica e reale possibilità di correzione umana, evitando che le loro raccomandazioni si trasformino in automatismi incontrollati che condizionano decisioni istituzionali.

Una volta stabilita questa architettura concettuale e giuridica, che obbliga a preservare il controllo umano effettivo, risulta particolarmente utile analizzare un caso concreto, in cui tali tensioni si manifestano con chiarezza: il sistema spagnolo di valutazione del rischio di violenza di genere noto come "VioGén", operativo dal 2007 e riformato nel 2025, con la denominazione "VioGén 2".

2. Antecedenti: "VioGén" e la sua evoluzione verso "VioGén 2"

Prima di analizzare approfonditamente il sistema "VioGén" ed esporre in che misura esso costituisca un esempio paradigmatico dell'impatto che i sistemi di IA possono avere sulla libertà decisionale nell'ambito giuridico, conviene ricordare che si tratta di uno strumento progettato per valutare il rischio in casi di violenza di genere, la cui operatività può, nella pratica, condizionare l'azione degli agenti responsabili della protezione delle vittime e, con ciò, incidere sui diritti fondamentali in gioco.

VioGén è stato creato dal Ministero dell'interno e messo in funzione il 26 luglio 2007, in attuazione della legge organica del 28 dicembre 2004, n. 1, sulle «*Misure di protezione integrale contro la violenza di genere*». Il suo fondamento normativo immediato è stata l'istruzione n. 10/2007 della Segreteria di Stato per la Sicurezza, che ha approvato il primo protocollo di valutazione del rischio da parte della polizia, successivamente modificato mediante varie istruzioni amministrative.

In tal modo, VioGén è stato concepito con una serie di obiettivi istituzionali: raggruppare tutte le istituzioni pubbliche competenti, integrare in un unico sistema l'informazione rilevante sui casi attivi, realizzare previsioni di rischio basate su fattori empiricamente validati, assegnare misure di protezione e monitoraggio in funzione del livello di rischio, emettere avvisi o allarmi automatici in caso di incidenti rilevanti e creare una rete statale di protezione coordinata e omogenea, disponibile in tutto il territorio nazionale[3].

Il sistema alimenta la propria banca dati mediante denunce alla polizia, interviste alle vittime, perizie forensi, ordini di protezione, precedenti penali e giudiziari, nonché informazioni personali e sociali sia della vittima sia dell'aggressore. Con tutte queste informazioni, VioGén applica strumenti attuariali automatizzati, in particolare la "valutazione di polizia del rischio" (VPR) e la "valutazione dell'evoluzione del rischio" (VPER), basati su modelli statistici che seguono metodologie simili agli approcci epidemiologici utilizzati nella Sanità pubblica[4].

A tal fine, il sistema stabilisce cinque livelli di rischio (non rilevato; basso; medio; alto; estremo), ciascuno dei quali comporta misure di protezione specifiche, che possono variare da un monitoraggio di base a una vigilanza intensiva o alla richiesta di un ordine giudiziario restrittivo. Sebbene l'agente mantenga formalmente la facoltà di adattamento, nella pratica la raccomandazione algoritmica delimita l'insieme iniziale delle misure da considerare.

Dal 2007, VioGén ha registrato quasi tre milioni di valutazioni di rischio, diventando una piattaforma nazionale di coordinamento tra istituzioni di polizia, giudiziarie, penitenziarie e sociali. Lo stesso Ministero dell'interno riconosce che il sistema è uno strumento

1. FRA, *Getting the Future Right – Artificial Intelligence and Fundamental Rights*, Ufficio delle pubblicazioni dell'UE, Lussemburgo, 2020 (https://fra.europa.eu/en/publication/2021/getting-future-right-artificial-intelligence-and-fundamental-rights-summary).

2. www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.

3. https://interior.gob.es/opencms/en/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen-2/.

4. https://s088672f9df8c57fe.jimcontent.com/download/version/1709741880/module/12490599698/name/SISTEMA%20DE%20SEGUIMIENTO%20INTEGRAL%20EN%20CASOS%20DE%20VG.pdf.

«in costante evoluzione», soggetto a revisioni periodiche mediante istruzioni che aggiornano la sua logica attuariale e le sue funzionalità.

L'esperienza accumulata in quasi due decenni mostra che VioGén presenta difetti strutturali che hanno avuto un impatto reale sulla protezione delle vittime. Tra i più significativi, un'analisi citata dall'Human Rights Research Center[5], basata su dati pubblicati dal *New York Times*, ha rivelato che 55 delle 247 donne assassinate dopo una valutazione della polizia erano state classificate come a rischio "basso" o "trascurabile". Questo risultato mostra che lo strumento non fallisce solo in casi marginali, ma in situazioni critiche in cui la valutazione è risultata insufficiente per attivare misure adeguate.

Un caso emblematico è quello di Lobna Hemid, assassinata nel 2022 dopo essere stata valutata da VioGén come "a basso rischio". La classificazione iniziale, percepita come tecnicamente neutra, ha condizionato la risposta istituzionale e contribuito alla mancata adozione delle misure di protezione necessarie.

Un'altra critica ricorrente riguarda l'accettazione automatica del risultato. Secondo l'*audit* di Eticas Foundation[6], gli agenti mantengono il punteggio generato dal sistema nel 95% delle valutazioni, nonostante il protocollo consenta aggiustamenti basati sull'apprezzamento professionale. Di conseguenza, la supervisione umana si riduce a una formalità e la raccomandazione algoritmica diventa una decisione operativa *de facto*.

Il problema è aggravato dal fatto che, in molti casi, le vittime non rivelano informazioni critiche durante le interviste con la polizia a causa della paura, della dipendenza economica o dello *status* migratorio, generando dati incompleti o distorti. Questi *bias* originari si trasmettono all'algoritmo, che sottostima il rischio proprio nelle vittime più vulnerabili.

A ciò si aggiungono le limitazioni inerenti all'approccio attuariale. Come spiega il Ministero dell'interno, i modelli utilizzati in VPR e VPER si basano su indicatori statistici associati alla recidiva violenta, ma senza relazione causale diretta. Di conseguenza, piccole variazioni nelle risposte possono far cambiare il livello di rischio assegnato, senza che i modelli colgano fattori qualitativi o dinamici propri del caso concreto.

Infine, la mancanza di trasparenza del sistema costituisce una delle critiche più ricorrenti. L'*audit* di Eticas Foundation ha rilevato che VioGén «non consente di conoscere la formula esatta, i pesi assegnati né la base empirica completa del modello», e che non è stato sottoposto ad *audit* tecnici esterni indipendenti. Questa opacità viola i principali *standard* internazionali sull'uso etico e responsabile dell'intelligenza artificiale, che esigono tracciabilità, auditabilità e supervisione umana effettiva nei sistemi che incidono sui diritti fondamentali. In questa linea, la Convenzione quadro del Consiglio d'Europa sull'Intelligenza Artificiale, primo trattato internazionale vincolante in materia, impone agli Stati l'obbligo di garantire la tracciabilità dei sistemi e un'adeguata supervisione umana lungo tutto il loro ciclo di vita. La Raccomandazione dell'UNESCO sull'Etica dell'IA (2021) stabilisce allo stesso modo che tali sistemi devono operare secondo criteri di trasparenza, spiegabilità e supervisione umana, a difesa della dignità e dei diritti fondamentali delle persone. Allo stesso modo, le Linee guida etiche dell'Unione interparlamentare sottolineano che ogni IA impiegata in funzioni pubbliche deve preservare l'autonomia umana ed evitare qualsiasi forma di delega occulta delle decisioni, insistendo sulla necessità di garantire un controllo umano continuo ed effettivo.

L'insieme di questi fattori erode le garanzie che devono reggere l'uso dell'IA in contesti di diritti fondamentali. Inoltre, la classificazione algoritmica condiziona l'intensità della protezione fornita: secondo dati citati da *Politico*, solo una donna su sette che richiede protezione alla polizia la riceve effettivamente, in gran parte perché l'assistenza prioritaria è condizionata dal livello di rischio attribuito dal sistema[7].

Le critiche accumulate per oltre un decennio hanno portato il Ministero dell'interno a promuovere, nel 2025, una riforma integrale del sistema, che ha dato luogo a "VioGén 2". Questo aggiornamento si articola attorno a tre linee di azione strettamente collegate ai difetti rilevati nel modello precedente: in primo luogo, l'eliminazione della categoria "assenza di rischio", il cui utilizzo aveva dato luogo a sottovalutazioni particolarmente gravi nella pratica; in secondo luogo, l'integrazione obbligatoria di informazioni qualitative e contestuali nelle interviste di polizia, orientata a

5. www.humanrightsresearch.org/post/spain-s-gender-violence-detection-algorithm-reveals-failures-in-saving-domestic-violence-victims.

6. https://eticasfoundation.org/wp-content/uploads/2024/12/Eticas_Audit_of_VioGen.pdf.

7. M. Heikkilä, *AI: Decoded: Spain's flawed domestic abuse algorithm – Ban debate heats up – Holding the police accountable*, *Politico*, 16 marzo 2022 (www.politico.eu/newsletter/ai-decoded/spains-flawed-domestic-abuse-algorithm-ban-debate-heats-up-holding-the-police-accountable-2/).

ridurre l'incidenza di dati incompleti derivanti dalla paura, dalla dipendenza economica o dalla situazione amministrativa di molte vittime; in terzo luogo, l'introduzione di piani di protezione personalizzati e di una migliore coordinazione interistituzionale, destinate ad evitare risposte eccessivamente dipendenti dalla classificazione algoritmica. Queste linee di riforma sono raccolte nel materiale editoriale sull'aggiornamento del sistema diffuso nel 2025.

Con queste misure, VioGén 2 non intende solo migliorare la precisione tecnica del sistema, ma soprattutto correggere i rischi di automazione occulta rilevati nella pratica, ampliando i margini di supervisione umana e restituendo agli operatori la capacità di esercitare un giudizio più completo e contestualizzato. VioGén 2 rappresenta, quindi, una risposta istituzionale alle sfide poste dall'uso dell'IA nella protezione dei diritti fondamentali.

3. Funzionamento reale: suggerimento o decisione occulta?

Sebbene il sistema VioGén sia stato concepito formalmente come uno strumento di supporto destinato a offrire una guida nell'azione professionale, il suo funzionamento pratico rivela che la raccomandazione algoritmica finisce per diventare, in una parte sostanziale dei casi, una decisione occulta. Questa trasformazione non è intenzionale, ma risponde a meccanismi specifici che si verificano abitualmente nei sistemi di intelligenza artificiale integrati in circuiti istituzionali. Come mostrato precedentemente, sebbene il *design* ufficiale di VioGén stabilisca che l'algoritmo fornisce solo un livello di rischio orientativo e che l'autorità di polizia deve valutare il caso in modo indipendente, la prassi documentata indica che gli operatori seguono la raccomandazione algoritmica quasi senza intervento umano, mantenendo il risultato automatico nel 95% delle valutazioni, secondo l'*audit* di Eticas Foundation.

Ciò avviene perché, nella pratica, la classificazione generata dal sistema funziona come un quadro cognitivo che agisce come una diagnosi preliminare e condiziona il modo in cui si interpretano l'intervista, i segnali di allarme o la narrazione della vittima. Si tratta di un fenomeno tipico dell'effetto di ancoraggio, in cui il primo valore ricevuto diventa il riferimento dominante per ogni analisi successiva. A ciò si aggiunge che il carico di lavoro, la pressione operativa e la necessità di standardizzazione portano gli agenti ad assumere la raccomandazione come punto di partenza operativo, attivando protocolli, priorità e determinando l'intensità del monitoraggio. In questo contesto, la raccomandazione funziona come una pre-decisione, perché è già collegata a un repertorio di misure di polizia preconfigurate.

A livello psicologico, influisce anche il fatto che l'algoritmo non produce un testo interpretativo, bensì una categoria chiusa ("basso rischio", "medio rischio", "alto rischio"), semplificando una realtà complessa in un risultato apparentemente esatto, che acquisisce un'autorità tecnica superiore all'intuizione professionale. La combinazione di ancoraggio cognitivo, standardizzazione organizzativa e mancanza di trasparenza trasforma la raccomandazione algoritmica in un riferimento decisionale che condiziona la prassi di polizia e limita la reale capacità di revisione umana.

Queste tensioni assumono una rilevanza particolare nell'ambito giurisdizionale, poiché i sistemi di IA utilizzati nella giustizia o nella sicurezza sono classificati dall'*AI Act* – regolamento (UE) 2024/1689 –[8] come sistemi ad alto rischio, che richiedono trasparenza, tracciabilità, supervisione umana significativa e valutazioni di impatto sui diritti fondamentali, proprio per evitare condizionamenti indebiti sulla libertà decisionale giudiziaria in contesti in cui il giudice riceve informazioni previamente elaborate da sistemi automatizzati.

4. Impatto sulla libertà decisionale giudiziaria

L'utilizzo di sistemi algoritmici nell'ecosistema della giustizia solleva una questione decisiva: come proteggere la libertà decisionale quando l'informazione che raggiunge il giudice è stata previamente configurata da un sistema automatizzato? In sede giudiziaria, tale libertà esige la possibilità di analizzare il caso senza condizionamenti esterni, valutando i fatti e il rischio con piena autonomia. Tuttavia, quando l'informazione che raggiunge il tribunale proviene da un sistema la cui logica interna non è verificabile, si altera l'ambiente cognitivo in cui il giudice deve deliberare. Questa configurazione preliminare, derivante non da un'intenzione di interferenza, ma dal modo in cui i sistemi di IA organizzano, sintetizzano o filtrano l'informazione preliminare di polizia, può influire sulla ponderazione dell'urgenza, della proporzionalità e del rischio nell'adozione di misure cautelari o di protezione.

Nel quadro europeo e internazionale, l'autonomia decisionale del giudice non è concepita solo come

8. https://artificialintelligenceact.eu/es/ai-act-explorer/.

assenza di pressioni esterne dirette, ma come diritto e dovere di decidere sulla base di informazioni comprensibili e verificabili, senza condizionamenti provenienti da meccanismi automatizzati. Ciò è richiesto dall'art. 6 della Convenzione europea dei Diritti dell'uomo, che esige un tribunale «indipendente e imparziale», garanzia che include l'indipendenza da influenze tecnologiche che possano limitare la reale capacità del giudice di formare un criterio. A sua volta, l'art. 47 della Carta dei Diritti fondamentali dell'Unione europea consacra la «tutela giurisdizionale effettiva», che presuppone la disponibilità di informazioni sufficienti e chiaramente accessibili affinché il giudice possa motivare la sua decisione senza vincoli strutturali. Questa visione è rafforzata dagli *standard* del Consiglio d'Europa, raccolti dal CCJE (Consiglio consultivo dei giudici europei), i cui *pareri* (in particolare il n. 26/2023 sulle tecnologie di supporto nella giustizia) insistono sul fatto che l'indipendenza giudiziaria include la capacità di controllare la razionalità del processo decisionale, il che esige che il giudice possa comprendere, verificare e mettere in discussione l'informazione che incorpora alla propria motivazione.

La Convenzione quadro del Consiglio d'Europa sull'Intelligenza Artificiale (2024) rafforza tale esigenza, imponendo obblighi di trasparenza e adeguata supervisione umana lungo tutto il ciclo di vita dei sistemi algoritmici, impedendo che l'IA operi come un filtro non controllato della realtà che arriva ai tribunali. Nella stessa direzione, la Raccomandazione dell'UNESCO sull'etica dell'IA (2021) esige la primazia dell'*agency* umana in ogni decisione pubblica, avvertendo che i sistemi non trasparenti possono spostare la capacità di giudizio professionale e indebolire la motivazione individualizzata delle decisioni. Questo strumento stabilisce che la supervisione umana deve essere *reale*, *significativa* e *costante*, specialmente quando l'IA incide su diritti essenziali. Dal canto loro, le Linee guida etiche dell'Unione interparlamentare sottolineano che l'IA impiegata in funzioni pubbliche deve preservare l'autonomia decisionale umana ed evitare qualunque forma di delega occulta delle decisioni, insistendo sulla necessità di un controllo umano permanente.

Nell'ambito dell'Unione europea, il regolamento (UE) 2024/1689 (*AI Act*) classifica i sistemi di IA applicati alla giustizia e alla sicurezza nella categoria «ad alto rischio», sottoponendoli ai requisiti più rigorosi del regolamento: trasparenza rafforzata, tracciabilità documentale, supervisione umana significativa, solida *governance* dei dati e valutazioni di impatto sui diritti fondamentali prima della loro messa in servizio. Secondo la sintesi dell'Oxford Institute of Technology and Justice, questa categoria riflette che l'IA utilizzata nella giustizia può influire materialmente sul giusto processo, sull'uguaglianza delle parti e sull'imparzialità effettiva del tribunale, motivo per cui l'UE esige salvaguardie rafforzate per evitare che l'IA deformi l'ambiente fattuale sul quale il giudice delibera.

Da questa prospettiva europea e internazionale, l'irruzione di VioGén non pone solo un problema tecnico, ma costituzionale in senso ampio: la giurisdizione non può operare con piena indipendenza quando la realtà processuale che esamina è già condizionata da un sistema la cui metodologia, logica interna e *bias* non risultano accessibili al giudice. L'indipendenza giudiziaria implica anche indipendenza da infrastrutture tecnologiche non trasparenti, e la tutela giurisdizionale effettiva esige che il giudice possa ricostruire criticamente il caso e motivare la propria decisione senza dipendere da una categoria statistica chiusa o da una valutazione algoritmica non spiegata.

Pertanto, la vera sfida posta dai sistemi di valutazione automatizzata del rischio non risiede in una sostituzione formale del giudice, vietata nel quadro europeo senza eccezioni, ma nella riduzione del controllo epistemico che il giudice deve esercitare sull'informazione rilevante. Preservare la libertà decisionale giudiziaria richiede che l'IA non definisca i contorni della realtà processuale, ma si limiti ad apportare un'informazione comprensibile, verificabile e sottoposta ad *audit*, sempre subordinata al giudizio umano.