Camera di Commercio Cagliari - Oristano

CAMERA ARBITRALE

L'INTELLIGENZA ARTIFICIALE NELL'ARBITRATO E NELLE PROFESSIONI: OPPORTUNITÀ, RISCHI E COMPLIANCE

Strumenti operativi e normativi per l'integrazione dell'IA nella giustizia arbitrale

12 giugno 2026

CAMERA DI COMMERCIO
DI CAGLIARI-ORISTANO

Via Carducci 23
Oristano

Cybersicurezza e IA nel contesto dell'arbitrato:

dal threat landscape all'implementazione di security by design e resilienza.

Relazione dell'Avv. Francesco Paolo Micozzi

Ordine dei dottori Commercialisti ed Esperti Contabili di Oristano

ORDINE degli INGEGNERI della Provincia di Oristano

Ministero della Giustizia

PROGRAMMA DELL'EVENTO

Saluti istituzionali

Avv. Enrico Maria Meloni - Presidente dell'Ordine degli Avvocati di Oristano

Dott.ssa Giuseppina Uda – Presidente dell'Ordine dei Commercialisti di Oristano

Ing. Cristian Licheri – Presidente dell'Ordine degli Ingegneri della Provincia di Oristano di Oristano

Dott.ssa Ilaria Licheri – Segretario dell'Ordine dei Dottori Agronomi e dei Dottori Forestali della Provincia di Oristano

Dott. Davide Boneddu – Presidente dell'Ordine dei Geologi della Sardegna

Presentazione del Convegno

Dott.ssa Grazia Corradini – Presidente del Consiglio e della Camera Arbitrale presso la Camera di Commercio di Cagliari-Oristano

Apertura lavori

Prof. Avv. Fabrizio Montaldo – Avvocato cassazionista e Professore aggregato di Diritto processuale civile e Ordinamento giudiziario presso Unica - Vice - Presidente del Consiglio della Camera Arbitrale presso la Camera di Commercio di Cagliari – Oristano

Avv. Enrico Maria Meloni - Avvocato cassazionista, Presidente dell'Ordine degli Avvocati di Oristano e Vice Presidente Vicario del Consiglio della Camera Arbitrale presso la Camera di Commercio di Cagliari-Oristano

Cybersicurezza e IA nel contesto dell'arbitrato: dal threat landscape all'implementazione di security by design e resilienza.
Relatore Avv. Francesco Paolo Micozzi, avvocato cassazionista

Intelligenza Artificiale oltre l'apparenza: capire i meccanismi per evitare l'illusione della conoscenza.
Relatore Prof. Giorgio Giacinto, Professore Ordinario di Ingegneria Informatica presso l'Università di Cagliari

Quando l'algoritmo sbaglia: responsabilità civile del professionista, tutela delle parti e valutazione del danno nell'arbitrato.
Relatore Avv. Mauro Solinas, Avvocato cassazionista, Ordine di Oristano

AI Literacy e Arbitrato: consapevolezza tecnologica e responsabilità decisionale alla prova della decisione algoritmica.
Relatore: Dott.ssa Antonella Fancello, Docente di "Amministrazione Digitale - E-Government" presso l'Università di Sassari

Dibattito sulle seguenti problematiche con la partecipazione dei relatori

- Quali strumenti IA gli arbitri possono iniziare a usare da domani
- Dove allocare risorse per massimo impatto
- Errori comuni da evitare nell'adozione dell'IA
- Come prepararsi per l'evoluzione tecnologica dei prossimi 2-3 anni

Indice generale

Il procedimento arbitrale come infrastruttura fiduciaria 5

Il valore informativo dell'arbitrato e la crisi del perimetro 5

Vecchie minacce, nuova scala 6

IA come strumento, amplificatore e bersaglio 7

Prompt injection e trasformazione della prova in istruzione 8

RAG poisoning, contaminazione del *corpus* e integrità della base informativa 9

IA "agentica", esfiltrazione e rischio di autonomia eccessiva 9

Allucinazioni, *automation bias* e *deskilling* del decisore 10

Deepfake, autenticità della prova e *liar's dividend* 11

Dall'obbligo di sicurezza alla non-delega 12

NIS2, Cyber Resilience Act e standard tecnici come linguaggio comune per il giurista 13

Verso un governo procedurale dell'IA 14

Security by design nel procedimento arbitrale 14

Resilienza e incident response con dimensione procedurale 16

La procedural order come dispositivo di sicurezza 17

Conclusioni 18

Cybersicurezza, intelligenza artificiale e arbitrato: dal threat landscape alla resilienza della decisione arbitrale

Relazione dell'Avv. Francesco Paolo Micozzi

Francesco Paolo Micozzi, Avvocato cassazionista e docente di informatica giuridica. Membro del comitato scientifico, per il Dipartimento di Giurisprudenza dell'Università degli Studi di Perugia, della School on Artificial Intelligence in health for people well-being dell'Istituto di fisica applicata "Nello Carrara" del Consiglio Nazionale delle Ricerche. Si occupa di diritto penale, con particolare predilezione per il diritto dell'informatica, delle nuove tecnologie, intelligenza artificiale, cybersecurity, data protection / privacy e diritto d'autore. Guida il modulo Jean Monnet "Cybersecurity Issues and Breaches in European Rules: a practical approach" (CIBER), e componente dell'Academic Staff del Centro d'Eccellenza Jean Monnet "Building the Age of a Lawful and sustainable Data-Use" (BALDUS). Docente per la Scuola Superiore della Magistratura sui temi correlati alle indagini nell'ambito dei cybercrime. Autore di monografie e saggi in tema di diritto penale dell'informatica, sicurezza informatica, protezione dei dati personali e diritto d'autore (ultima pubblicazione: "Sicurezza informatica. Obblighi e responsabilità dopo il recepimento della NIS2 e la L. 90/2024", Ed. Wolters Kluver, 2024. Data Protection Officer / Responsabile per la protezione dei dati personali per enti pubblici e società. Tra le altre materie di competenza, oltre ai temi dell'informatica giuridica, privacy e data protection anche anticorruzione, amministrazione digitale, diritto d'autore e sicurezza informatica. Componente del pool di esperti a supporto dell'EDPB (European Data Protection Board – Comitato Europeo per la Protezione dei Dati Personali) lato tecnico su "new technologies and information security" e lato giuridico sulle nuove tecnologie. Fa parte del gruppo di lavoro della Fondazione Italiana per l'Innovazione Forense (FIIF) presso il Consiglio Nazionale Forense. Componente della Commissione per la valutazione degli avvocati specialisti nella materia "Diritto dell'informazione, della comunicazione digitale e della protezione dei dati personali".

Il procedimento arbitrale come infrastruttura fiduciaria

La trasformazione digitale dell'arbitrato non consiste soltanto nel passaggio dal fascicolo cartaceo al deposito telematico o dalla riunione in presenza all'udienza da remoto. Essa produce una diversa geografia del procedimento. Il fascicolo non è più custodito in un luogo identificabile, ma circola tra piattaforme di deposito, sistemi cloud, caselle di posta elettronica, strumenti di videoconferenza, applicazioni di traduzione e, sempre più spesso, sistemi di intelligenza artificiale in grado di interrogare, sintetizzare o riorganizzare il materiale probatorio. Da questa prospettiva, l'arbitrato non è soltanto una tecnica alternativa di risoluzione delle controversie, ma un'infrastruttura fiduciaria nella quale l'affidabilità della decisione dipende anche dall'affidabilità dei flussi informativi che la precedono.

La tesi di fondo che qui si sostiene è che, nell'arbitrato contemporaneo, la cybersicurezza protegga qualcosa di più ampio dei sistemi informatici. Protegge la possibilità stessa di decidere in modo attendibile. Una violazione della confidenzialità può esporre segreti, strategie difensive e informazioni negoziali. Una violazione dell'integrità può alterare documenti, metadati, traduzioni, estrazioni o ricostruzioni cronologiche. Una violazione della disponibilità può impedire depositi, udienze e deliberazioni. Una violazione dell'autenticità può rendere incerto chi abbia prodotto una dichiarazione, sottoscritto un documento o generato un contenuto. Una carenza di tracciabilità può impedire di ricostruire come un output di IA abbia influenzato il procedimento.

Questa ricostruzione impone di superare due semplificazioni opposte. La prima consiste nel ritenere che l'IA sia incompatibile con l'arbitrato e debba essere esclusa in via generale. La seconda, speculare, consiste nel considerare l'IA come un mero acceleratore neutrale, da governare con clausole generiche di riservatezza o con una dichiarazione di "*human in the loop*". Entrambe le posizioni sarebbero inconferenti. L'IA può essere utile – e, in alcuni contesti, può ridurre tempi, costi e asimmetrie informative – tuttavia, il suo uso è accettabile solo se inserito in un ambiente capace di distinguere i casi d'uso, controllare i dati, presidiare le fonti, limitare i privilegi degli agenti IA, documentare gli usi materialmente rilevanti e conservare una supervisione umana effettiva (e non di mera forma o passaggio procedurale acritico).

Il valore informativo dell'arbitrato e la crisi del perimetro

L'arbitrato è spesso descritto come procedimento flessibile, specializzato e riservato. La riservatezza, tuttavia, non ha la stessa consistenza in ogni ordinamento e non discende sempre automaticamente dalla convenzione arbitrale. Può derivare dalla legge applicabile, dal regolamento dell'istituzione, da un ordine del tribunale, da obblighi professionali o dalla natura delle informazioni trattate. La distinzione è essenziale sul piano giuridico, ma non riduce il problema di sicurezza poiché, ad esempio, le parti si aspettano comunque che il fascicolo resti sottratto a soggetti estranei e che segreti commerciali, dati personali, informazioni tecniche e strategie difensive non siano accessibili, manipolabili o ricostruibili da terzi.

Un procedimento arbitrale può riunire contratti, corrispondenza interna, dati finanziari, documentazione tecnica, informazioni sulla *supply chain*, relazioni peritali, dati sanitari, proprietà intellettuale, proposte transattive e testimonianze. Per un concorrente, un attore statuale, un gruppo criminale o un altro soggetto interessato, conoscere la struttura di un prezzo, una vulnerabilità di

prodotto, una strategia industriale o una soglia negoziale può essere più utile che interferire direttamente con l'esito del procedimento.

Il problema è accentuato dalla frammentazione strutturale dell'ambiente arbitrale. A differenza di un sistema informativo aziendale chiuso, il procedimento coinvolge istituzione arbitrale, segreteria, arbitri, eventuali segretari del tribunale, parti, avvocati, consulenti tecnici, interpreti, traduttori, fornitori di piattaforme, cloud provider, servizi di videoconferenza e strumenti di trascrizione. Il fascicolo può essere duplicato in decine di caselle email, cartelle sincronizzate, computer portatili, telefoni, sistemi di backup e archivi di studio. In questa configurazione, il tradizionale concetto di perimetro diventa instabile e la sicurezza effettiva del procedimento tende a equipararsi al livello del partecipante meno protetto.

La triade classica della sicurezza informatica, *confidentiality, integrity e availability* (confidenzialità, integrità e disponibilità), resta essenziale ma non più sufficiente. Nel procedimento arbitrale devono essere aggiunte almeno autenticità, tracciabilità e accountability. L'autenticità consente di stabilire chi o che cosa abbia generato un documento, un output o una informazione. La tracciabilità permette di ricostruire accessi, trasformazioni, versioni e interazioni con strumenti automatizzati. L'accountability attribuisce decisioni, controlli e responsabilità a soggetti identificabili. Senza questi tre attributi, la sicurezza del fascicolo non basta a garantire l'affidabilità della decisione.

Vecchie minacce, nuova scala

L'analisi delle minacce non deve partire dall'eccezionalità tecnologica, ma dalla continuità operativa degli attacchi. I dati ENISA più recenti mostrano che phishing, sfruttamento di vulnerabilità, ransomware, DDoS e supply chain continuano a costituire assi portanti del rischio europeo. Il rapporto ENISA Threat Landscape 2025, fondato su 4.875 incidenti relativi al periodo luglio 2024 - giugno 2025, segnala la persistenza di vettori iniziali tradizionali e la rapidità con cui vulnerabilità pubbliche vengono convertite in strumenti di intrusione. La lezione per l'arbitrato è immediata: non si deve immaginare che il rischio principale sia sempre un attacco sofisticato al modello di IA. Spesso l'accesso iniziale passa da credenziali sottratte o lasciate esposte, email credibili, servizi remoti mal configurati, dispositivi non aggiornati o fornitori compromessi[1].

Il quadro nazionale conferma la stessa direzione. Nel marzo 2026 l'Agenzia per la cybersicurezza nazionale ha registrato un incremento rilevante di eventi e incidenti, sottolineando il ruolo delle credenziali valide precedentemente sottratte e dei servizi di accesso remoto non adeguatamente configurati. Per un'istituzione arbitrale, il dato non rileva perché misuri specificamente gli arbitrati, ma perché evidenzia la dipendenza da servizi comuni. Un incidente che colpisce un provider di identità, una piattaforma cloud, un sistema di videoconferenza o un fornitore di repository può rendere indisponibile un procedimento anche senza una compromissione diretta della camera arbitrale[2].

---

[1] ENISA, ENISA Threat Landscape 2025, v. 1.2, 9 gennaio 2026, disponibile in https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025. Il report analizza 4.875 incidenti nel periodo 1 luglio 2024-30 giugno 2025.

[2] Agenzia per la cybersicurezza nazionale, Operational Summary ACN - Marzo 2026: NIS2 aumenta la visibilità degli incidenti cyber, 21 aprile 2026, https://www.acn.gov.it/portale/web/guest/-/operational-summary-acn-marzo-2026-nis2-aumenta-la-visibilita-degli-incidenti-cyber

L'IA generativa non cancella le minacce esistenti, le rende più credibili, scalabili e personalizzabili. Un attaccante può produrre messaggi linguisticamente corretti, adattati al lessico della procedura, coerenti con il dominio dell'istituzione e riferiti a nomi, date o numeri di procedimento effettivi. Può generare un falso messaggio della segreteria, una richiesta apparentemente urgente dell'arbitro presidente, un invito a una videoconferenza con pagina di autenticazione contraffatta, un audio sintetico che imita una voce nota o un documento tecnico contenente istruzioni invisibili a un lettore umano ma interpretabili da un assistente di IA. La formazione non può più limitarsi a suggerire di diffidare delle email scritte male. Deve insegnare a verificare il canale, riconoscere le richieste anomale e usare autenticazione resistente al phishing.

IA come strumento, amplificatore e bersaglio

Per governare il rischio è utile distinguere tre funzioni dell'intelligenza artificiale nel procedimento arbitrale. In primo luogo, **l'IA è uno strumento di lavoro**. Può supportare OCR, classificazione, ricerca semantica, traduzione, sintesi, estrazione di entità, confronto di versioni, costruzione di cronologie, trascrizione e assistenza redazionale. Questi usi non sono però equivalenti. Una trascrizione automatica verificata da un professionista non ha lo stesso impatto di un sistema che seleziona le prove rilevanti per il tribunale. Un correttore linguistico non coincide con un sistema che suggerisce la motivazione del lodo. La valutazione deve riguardare il caso d'uso concreto, non l'etichetta commerciale dello strumento.

In secondo luogo, **l'IA è un amplificatore dell'attaccante**. Può abbassare il costo della ricognizione, aumentare la qualità dell'attacco di ingegneria sociale, tradurre correttamente messaggi malevoli, produrre varianti persuasive, generare codice o assistere l'impersonificazione vocale e testuale. Il vantaggio non consiste nell'attribuire al sistema capacità autonome fantascientifiche ma, comunque, l'automazione riduce il costo marginale dell'attacco e consente di adattare rapidamente contenuti e tono alla vittima, per rendere l'attacco maggiormente efficace e dirompente nei suoi effetti.

In terzo luogo, **l'IA è essa stessa bersaglio** e componente vulnerabile. Il NIST "Generative AI profile"[3] evidenzia rischi di confabulation (allucinazioni), cioè risposte errate ma presentate come affidabili; rischi per la protezione dei dati personali; produzione o circolazione di informazioni inesatte, manipolate o non verificabili; esposizione ad attacchi informatici o accessi non autorizzati; errori derivanti da un uso improprio dell'IA da parte degli operatori; e problemi legati ai diversi fornitori, piattaforme e servizi che compongono l'intero ciclo di funzionamento del sistema. L'OWASP Top 10 for LLM Applications 2025[4] organizza le principali vulnerabilità delle applicazioni basate su modelli linguistici intorno a *prompt injection* (inserimento di istruzioni malevole o manipolative nel testo fornito al modello, per indurlo a comportamenti non previsti o non autorizzati), divulgazione di informazioni sensibili (rischio che il sistema riveli dati personali, informazioni riservate, segreti aziendali o credenziali), *supply chain* (criticità derivanti da componenti, modelli, dataset, librerie, plugin o servizi di terzi impiegati nello sviluppo o nell'esercizio del sistema), *data and model poisoning*

[3] AUTIO, C. , SCHWARTZ, R. , DUNIETZ, J. , JAIN, S. , STANLEY, M. , TABASSI, E. , HALL, P. AND ROBERTS, K., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.AI.600-1, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958388

[4] OWASP GenAI Security Project, 2025 Top 10 Risk & Mitigations for LLMs and Gen AI Apps, https://genai.owasp.org/llm-top-10/.

(alterazione o contaminazione dei dati o del modello, tale da produrre risultati distorti, vulnerabili o manipolabili), *improper output handling* (mancato controllo, filtraggio o validazione delle risposte generate prima che siano usate da altri sistemi o mostrate agli utenti), *excessive agency* (attribuzione al sistema di IA di eccessiva autonomia, permessi o capacità operative, con rischio di azioni indesiderate o dannose), *system prompt leakage* (rivelazione delle istruzioni interne che regolano il comportamento del modello), debolezze di *vector* ed *embedding* (vulnerabilità nei meccanismi con cui il sistema rappresenta, ricerca e recupera informazioni, ad esempio nei sistemi RAG), *misinformation* (produzione o diffusione di informazioni false, fuorvianti o non verificabili) e consumo non limitato (uso eccessivo o incontrollato delle risorse del modello, con possibili effetti di indisponibilità del servizio, aumento dei costi o abuso del sistema). In un arbitrato, questa tassonomia consente di comprendere che un assistente documentale non è una semplice interfaccia di ricerca. È un sistema composto da modello, prompt, basi documentali, database vettoriale, connettori, autorizzazioni, log e strumenti di azione.

Prompt injection e trasformazione della prova in istruzione

La tecnica del *prompt injection* merita un'attenzione particolare perché incide sul rapporto tra prova e interpretazione automatica della stessa. Nei sistemi LLM integrati con applicazioni esterne, il confine tra dati e istruzioni può diventare poroso, permeabile. Istruzioni malevole possono essere incorporate in contenuti destinati a essere percepiti e interpretati (e quindi eseguiti) dall'applicazione, compromettendo il comportamento del sistema senza che l'utente invii direttamente un comando ostile.

L'arbitrato costituisce un ambiente particolarmente esposto a questa vulnerabilità. Il sistema di IA legge spesso documenti prodotti dalla controparte o da terzi quali contratti, email, relazioni tecniche, tabelle, pagine web, PDF, immagini con testo incorporato, metadati e commenti. Un documento depositato può contenere testo bianco su sfondo bianco, istruzioni nei metadati, contenuti fuori pagina, commenti invisibili nella visualizzazione ordinaria o riferimenti a risorse esterne. Se l'assistente documentale dispone di strumenti ulteriori, come inviare messaggi, salvare file, interrogare altri archivi o aprire link, il documento non è più soltanto oggetto di lettura ma può diventare interprete e attore di comandi specifici (spesso in modo totalmente oscuro all'utilizzatore).

Il passaggio concettuale è rilevante anche sul piano giuridico. La prova digitale non è più soltanto un contenuto da interpretare, ma può tentare di dirigere l'interprete automatico (leggasi il sistema IA) impiegato nella procedura. Ne deriva che tutto il materiale del fascicolo deve essere trattato come potenziale input da cui diffidare. Le istruzioni di sistema devono restare separate dal contenuto dei documenti; i connettori devono essere limitati; le azioni esterne devono richiedere approvazione umana; il corpus deve essere segregato per procedimento; i log devono consentire di ricostruire fonti e recuperi, senza trasformarsi in un nuovo archivio incontrollato di dati riservati. Poiché non esiste una difesa assoluta, la progettazione deve applicare il principio del "*least agency*", ossia concedere all'agente soltanto le capacità indispensabili, per il tempo necessario e con verifiche umane sui passaggi irreversibili.

RAG poisoning, contaminazione del *corpus* e integrità della base informativa

La seconda vulnerabilità sulla quale vogliamo concentrare la nostra attenzione riguarda il rapporto tra modelli generativi e basi documentali. Il *retrieval augmented generation* (RAG) viene spesso presentato come tecnica capace di ridurre le allucinazioni, poiché collega la risposta del modello a documenti recuperati da un corpus informativo. In realtà, il RAG non elimina il rischio, ma ne sposta una parte sulla qualità, integrità e governabilità della base documentale, nonché sulle modalità con cui il sistema seleziona i contenuti rilevanti e costruisce il prompt aumentato da sottoporre al modello generativo. Se il corpus, ossia l'insieme dei documenti acquisiti (*ingestion*), suddivisi in *chunk*, rappresentati tramite *embedding* (nel DB vettoriale), indicizzati nel sistema di retrieval e successivamente recuperati per comporre il contesto della risposta, è contaminato; se i metadati sono ingannevoli; se la deduplicazione è opaca; o se il sistema privilegia passaggi manipolati, la risposta può risultare linguisticamente coerente, ma sostanzialmente deformata.

La letteratura recente ha mostrato che anche l'iniezione di pochi testi malevoli in una base RAG può orientare risposte verso esiti scelti dall'attaccante. Ulteriori studi hanno evidenziato backdoor e vulnerabilità del recupero, con effetti sia sulla disponibilità sia sull'orientamento semantico dell'output. Nell'arbitrato il rischio può assumere forme pratiche come duplicati creati per dominare la ricerca semantica, versioni alterate, documenti con metadati falsi, appunti interni mescolati alle prove, fonti giuridiche non verificate, traduzioni incorporate come se fossero originali, o modifiche del repository effettuate da utenti con privilegi eccessivi[5].

La risposta non può consistere nella fiducia cieca nelle citazioni generate. Deve essere costruita sul ciclo di vita del *corpus*, anima del RAG. Occorrono fonti autorizzate, identità abilitate alla scrittura, "versionamento", registrazione degli *hash*, conservazione degli originali, deduplicazione documentata, separazione tra prove, fonti giuridiche e materiali di lavoro, citazioni puntuali verso il passaggio recuperato e possibilità di ricostruire quali documenti abbiano contribuito all'output. Nel contesto arbitrale, la governance della base RAG è una forma di catena di custodia.

IA "agentica", esfiltrazione e rischio di autonomia eccessiva

Per IA agentica si fa comunemente riferimento ai sistemi di intelligenza artificiale capaci non solo di generare contenuti testuali o audiovisivi, ma anche di perseguire un obiettivo attraverso una sequenza di operazioni, pianificando passaggi intermedi, utilizzando strumenti esterni (come, ad esempio, software specifici disponibili online o già presenti sul sistema informatico dell'utilizzatore) e, in alcuni casi, attivando azioni nell'ambiente digitale. L'elemento caratterizzante dell'IA Agentica non è una, quindi, "volontà" autonoma in senso umano, ma la possibilità che il sistema IA trasformi un'istruzione in una catena di attività operative come, ad esempio, consultare banche dati, interrogare repository, aprire documenti, chiamare API, compilare moduli, generare file, inviare comunicazioni o modificare (ma anche cancellare) contenuti.

Questa dimensione agentica rende più esigente il controllo. Un chatbot che produce testo può sbagliare, ma un agente che usa strumenti esterni può eseguire azioni. Questa capacità rende la prompt injection

[5] J. XUE, M. ZHENG, Y. HU, F. LIU, X. CHEN, Q. LOU, 'BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models', arXiv:2406.00083, 2024.

più grave, perché l'istruzione malevola può non limitarsi a influenzare una risposta testuale, ma spingersi al punto da innescare un flusso di azioni. Studi recenti sugli agenti mostrano che prompt injection semplici possono causare leakage di dati personali osservati durante l'esecuzione di compiti, soprattutto quando l'attività stessa richiede estrazione o autorizzazione di dati[6].

In un procedimento arbitrale, un agente che legge il fascicolo non dovrebbe poter inviare comunicazioni; un agente che prepara bozze non dovrebbe poter depositare atti; un agente che consulta fonti esterne non dovrebbe avere accesso indiscriminato a documenti riservati; un agente che produce una cronologia non dovrebbe poter modificare il repository originario. La separazione dei compiti non è una raffinatezza architetturale, ma una garanzia procedurale. Riduce la probabilità che una singola compromissione produca danni completi su confidenzialità, integrità e disponibilità.

L'*excessive agency* è quindi il punto in cui la sicurezza informatica incontra la teoria del mandato arbitrale. Più l'agente può fare, più aumenta la distanza tra l'intenzione umana e l'effetto procedimentale. Il controllo umano deve essere collocato prima dell'azione, non solo dopo la produzione dell'output. L'approvazione *ex post* di un deposito già inviato, di una comunicazione già trasmessa o di un file già cancellato non costituisce supervisione effettiva.

Allucinazioni, *automation bias* e *deskilling* del decisore

La confabulazione (o la tendenza a produrre allucinazioni) non è un difetto accidentale dei modelli generativi, ma un rischio connesso al loro funzionamento probabilistico. La nozione di "allucinazione", applicata ai sistemi di IA generativa, infatti, non deve essere intesa nell'accezione clinica o psicopatologica che il termine assume con riferimento alla dimensione umana, né come indice di un malfunzionamento tecnico del sistema IA. Nei modelli generativi, infatti, l'output è prodotto attraverso meccanismi statistico-probabilistici di previsione e composizione del linguaggio in cui il sistema seleziona sequenze testuali che risultano plausibili rispetto al contesto, ai dati di addestramento e alle istruzioni ricevute. Il problema nasce quando tale plausibilità linguistica non corrisponde alla realtà fattuale, alla fonte richiamata o al materiale effettivamente disponibile nel procedimento. In altri termini, l'IA può generare una risposta formalmente coerente e persuasiva, ma sostanzialmente falsa, inesatta o non verificabile. Volendo semplificare ulteriormente, il sistema non "comprende" input e output nel senso umano del termine, né verifica autonomamente la verità delle proprie affermazioni secondo criteri fattuali, documentali o giuridici. Input e output sono trasformati in rappresentazioni matematiche ed elaborati attraverso un percorso statistico-probabilistico. In questo senso ampio, ogni output generativo è un contenuto sintetico, prodotto dal modello e non semplicemente "ritrovato" nella realtà. Quindi, in senso ampio, potremmo dire che tutti gli output dei sistemi IA generativi sono, formalmente "allucinazioni".

In senso tecnico, tuttavia, l'allucinazione si manifesta quando la plausibilità statistico-linguistica dell'output viene scambiata per attendibilità fattuale, documentale o argomentativa, senza che vi sia una corrispondenza verificabile con la fonte, con il record o con la realtà richiamata. Nel dominio giuridico può manifestarsi come citazione inesistente, attribuzione errata di un principio a una decisione reale, sintesi che omette un'eccezione decisiva, ricostruzione cronologica che confonde

[6] M. ALIZADEH, Z. SAMEI, D. STETSENKO, F. GILARDI, *Simple Prompt Injection Attacks Can Leak Personal Data Observed by LLM Agents During Task Execution*, arXiv:2506.01055, 2025.

soggetti o date, traduzione che altera una qualificazione tecnica o bozza di motivazione che introduce fatti non presenti nel record. Il controllo umano non può ridursi alla rilettura stilistica. Deve essere un controllo capace di tornare alla fonte, verificare la pertinenza della citazione e ricostruire il percorso argomentativo.

La psicologia della decisione aggiunge una cautela ulteriore. L'*automation bias* descrive la tendenza a sovra-affidarsi all'output di un sistema automatizzato, soprattutto quando il risultato appare coerente, rapido e formalmente autorevole. Studi empirici indicano che i *bias* incorporati nei suggerimenti algoritmici possono propagarsi nelle decisioni umane anche oltre il momento dell'esposizione diretta. Una revisione sistematica recente ha mostrato come l'automation bias si manifesti in contesti di collaborazione uomo-IA e dipenda da fattori cognitivi, organizzativi e di design dell'interfaccia[7].

Il tema più sottile è il **deskilling**. Uno studio multicentrico osservazionale pubblicato nel 2025 su The Lancet Gastroenterology & Hepatology ha rilevato una diminuzione della performance non assistita di endoscopisti dopo esposizione ordinaria a un sistema di IA, pur con le cautele dovute al disegno osservazionale e al dominio medico specifico. Non si tratta di trasferire meccanicamente il risultato agli arbitri, ma di coglierne la lezione: mantenere un essere umano "nel circuito" non significa necessariamente conservarne la capacità di giudizio. Se l'arbitro o il consulente si abitua a leggere solo sintesi, accettare selezioni automatiche delle prove o correggere motivazioni già strutturate dal modello, la supervisione può diventare formale proprio quando sembra garantita[8].

La risposta deve essere organizzativa, e il decisore deve disporre di tempo, accesso alle fonti, competenze e libertà effettiva di ignorare l'output. Nelle questioni più importanti è opportuno che la ricostruzione autonoma preceda la consultazione dell'IA; che i passaggi essenziali siano verificati direttamente sui documenti; che le deliberazioni collegiali espongano il percorso umano; che siano conservate procedure operative per proseguire senza IA; che la formazione non insegni soltanto a usare lo strumento, ma anche a riconoscerne omissioni ed errori. Un indicatore credibile di *oversight* non è quante volte una persona abbia cliccato "approva", ma se quella persona sarebbe stata in grado di motivare il dissenso e completare il compito senza lo strumento.

Deepfake, autenticità della prova e *liar's dividend*

Audio, video e immagini sintetiche abbassano il costo della falsificazione e la disponibilità degli strumenti per ottenerla e, nello stesso tempo, rendono più plausibile contestare la prova autentica. Il rischio non consiste soltanto nella produzione di un contenuto falso, ma anche nel cosiddetto *liar's dividend*, che consiste nell'opportunità, per una parte, di sostenere che una prova genuina sia, in realtà, artificiale. In arbitrato, dove la prova tecnica e documentale può avere un peso decisivo, questa dinamica incide sul contraddittorio e sull'onere di allegazione e contestazione.

I detector automatici di elementi sintetici non possono essere trattati come prova conclusiva. Possono generare falsi positivi e falsi negativi, dipendono dalla qualità del file e dalla compressione, e diventano

[7] G. ROMEO et al., *Exploring automation bias in human–AI collaboration: a systematic literature review*, AI & Society, 2025, DOI 10.1007/s00146-025-02422-7.
[8] K. BUDZYŃ et al., *Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study*, The Lancet Gastroenterology & Hepatology, 2025, 10(10), pp. 896-903, DOI 10.1016/S2468-1253(25)00133-5; correzione DOI 10.1016/S2468-1253(25)00294-8.

rapidamente obsoleti. La valutazione deve combinare provenienza, file originale, modalità di acquisizione, metadati, hash, marche temporali, testimonianze, riscontri esterni, analisi forense e, quando disponibili, credenziali di provenienza o firme del dispositivo. L'AI Act introduce obblighi di trasparenza per determinati contenuti sintetici e la Commissione ha pubblicato nel giugno 2026 un codice di buone pratiche[9] su marcatura ed etichettatura, ma tali strumenti non sostituiscono l'accertamento dell'autenticità nel caso concreto.

Dall'obbligo di sicurezza alla non-delega

Il GDPR costituisce il primo presidio normativo, perché molti trattamenti dell'arbitrato riguardano dati personali, anche particolari o giudiziari. Gli articoli 5, 25 e 32 impongono minimizzazione, protezione fin dalla progettazione e misure tecniche e organizzative adeguate al rischio; gli articoli 28 e 44-49 incidono su fornitori e trasferimenti; gli articoli 33 e 34 governano le violazioni; l'articolo 35 richiede una valutazione d'impatto quando il trattamento può presentare un rischio elevato. L'uso di IA aggiunge un trattamento e spesso un trasferimento o un coinvolgimento di subfornitori. Occorre quindi verificare ruolo del provider, finalità, retention, riuso dei dati, localizzazione, cancellazione, logging e possibilità di audit[10].

L'Opinion 28/2024 dell'EDPB è particolarmente utile perché ricorda che un modello addestrato su dati personali non può essere considerato anonimo in ogni caso e che l'eventuale anonimato deve essere valutato in concreto. L'EDPB chiarisce inoltre che il legittimo interesse richiede una verifica effettiva di necessità e bilanciamento, e che un trattamento illecito nello sviluppo può incidere sulla liceità del successivo impiego del modello, salvo che questo sia stato debitamente anonimizzato. Per l'arbitrato, ciò significa che la dichiarazione commerciale secondo cui un servizio sarebbe "sicuro" o "non addestrato sui dati del cliente" non esaurisce la *due diligence*.

L'AI Act introduce una disciplina distinta ma convergente. L'Allegato III, punto 8, lettera a), considera ad alto rischio i sistemi destinati a essere utilizzati da o per conto di un'autorità giudiziaria per assisterla nella ricerca e interpretazione dei fatti e del diritto e nell'applicazione del diritto a un insieme concreto di fatti, nonché i sistemi usati in modo analogo nella risoluzione alternativa delle controversie. La norma non rende automaticamente ad alto rischio qualunque software usato in arbitrato. Un calendario, una videoconferenza, un OCR o una traduzione verificata non coincidono con un sistema che seleziona prove, propone la ricostruzione dei fatti o orienta la motivazione. La classificazione dipende da finalità, deployment, ruolo del soggetto che usa il sistema, influenza sull'esito e possibilità di rientrare nelle eccezioni dell'articolo 6, paragrafo 3.

Al momento, la classificazione ad alto rischio deve inoltre essere letta alla luce delle bozze di linee guida della Commissione, pubblicate il 19 maggio 2026 e sottoposte a consultazione, che mirano a chiarire l'applicazione dell'articolo 6 e dell'Allegato III. Deve essere letto con cautela anche il calendario applicativo, posto che l'AI omnibus e l'accordo politico del 7 maggio 2026 hanno indicato un rinvio per alcune regole sui sistemi ad alto rischio, anche se ciò dovrebbe indurre a qualificare il

[9] Commissione europea, Code of Practice on Transparency of AI-Generated Content, pubblicato il 10 giugno 2026, destinato a sostenere l'adempimento degli obblighi di trasparenza dell'art. 50 AI Act- https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content.

[10] European Data Protection Board, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 17 dicembre 2024.

dato come evoluzione ancora da tradurre nel testo definitivo e non come ragione per sospendere la governance. Restano comunque attuali l'obbligo di AI literacy applicabile dal 2 febbraio 2025, gli obblighi di trasparenza dell'articolo 50, la disciplina privacy e i doveri professionali.

La legge italiana n. 132/2025 rafforza l'impostazione antropocentrica e, per le professioni intellettuali, con l'articolo 13, limita l'uso dell'IA ad attività strumentali e di supporto, preservando la prevalenza del lavoro intellettuale oggetto della prestazione e richiedendo un'informazione chiara al cliente sui sistemi utilizzati. Per l'attività giudiziaria, l'articolo 15 riserva al magistrato l'interpretazione e applicazione della legge, la valutazione dei fatti e delle prove e l'adozione dei provvedimenti. Quest'ultima disposizione non si applica testualmente all'arbitro come se fosse una norma del procedimento arbitrale, ma esprime un principio coerente con il mandato personale del decisore: l'IA può assistere, non sostituire il giudizio.

Il codice di procedura civile conferma la centralità di contraddittorio, imputabilità dell'attività istruttoria e responsabilità del lodo. L'articolo 816-bis fonda il potere delle parti e degli arbitri di regolare il procedimento, imponendo ragionevoli ed equivalenti possibilità di difesa. L'articolo 816-ter consente assistenza tecnica e deleghe istruttorie, ma non dissolve l'imputazione a soggetti identificabili. Gli articoli 822 e 823 collocano la decisione nella partecipazione degli arbitri e nella sottoscrizione del lodo. L'uso dell'IA non costituisce di per sé motivo autonomo di nullità; può però contribuire a produrre difetti già previsti dall'articolo 829, come violazione del contraddittorio, inosservanza di forme imposte dalle parti, motivazione apparente, contraddittorietà del dispositivo o omessa decisione. La governance dell'IA è dunque anche presidio di validità e tenuta del lodo.

NIS2, Cyber Resilience Act e standard tecnici come linguaggio comune per il giurista

Il d.lgs. n. 138/2024, di recepimento della direttiva NIS2, impone ai soggetti essenziali e importanti misure tecniche, operative e organizzative adeguate e proporzionate, comprendenti analisi del rischio, gestione degli incidenti, continuità operativa, backup, sicurezza della supply chain, sviluppo sicuro, verifica dell'efficacia, igiene informatica, formazione, crittografia, controllo degli accessi e MFA. Non è corretto affermare in astratto che ogni camera arbitrale rientri nel perimetro NIS2; occorre verificare soggetto giuridico, registrazione, servizi, sistemi e infrastrutture effettivamente utilizzate. Tuttavia, anche quando non vi sia applicabilità diretta, NIS2 produce effetti indiretti attraverso contratti, procurement e aspettative di sicurezza lungo la supply chain.

Il Cyber Resilience Act rileva soprattutto nella due diligence sui prodotti digitali. Una camera arbitrale che acquista una piattaforma non diventa normalmente produttore, ma potrà chiedere aggiornamenti, gestione delle vulnerabilità, supporto, documentazione, sicurezza by design e condizioni di manutenzione coerenti con i requisiti europei. Il punto pratico è che la sicurezza del procedimento dipende anche da ciò che il fornitore può dimostrare prima dell'incidente, non soltanto da ciò che promette dopo.

Gli standard tecnici non devono diventare una collezione di sigle. NIST CSF 2.0, Zero Trust Architecture, ISO/IEC 27001, ISO/IEC 42001 e ISO/IEC 23894 offrono un linguaggio comune per definire funzioni, responsabilità, controlli, evidenze e miglioramento continuo. La loro utilità, in arbitrato, non consiste nel trasformare ogni istituzione in una grande organizzazione certificata, ma nel

consentire una traduzione proporzionata dei principi: inventario, governo, protezione, rilevazione, risposta, ripristino, gestione dei fornitori e misurazione dell'efficacia[11].

Verso un governo procedurale dell'IA

La soft law arbitrale ha anticipato molti nodi oggi centrali. Il "*Protocol on cybersecurity in international arbitration*" del 2022[12] propone un approccio basato su rischio e ragionevolezza, calibrato su valore, dimensione, risorse, capacità e profilo del caso, raccomandando di affrontare la sicurezza il prima possibile e normalmente non oltre il primo case management meeting. La Roadmap ICCA-IBA [13] del 2022 integra protezione dei dati e procedimento arbitrale, mostrando come parti, avvocati, arbitri e istituzioni non possano essere ricondotti automaticamente a una catena titolare-responsabile semplice.

Le linee guida SVAMC del 2024 hanno dato forma al principio di non delega, affermando che l'arbitro non deve delegare ad alcuno strumento di IA il proprio mandato personale, in particolare il processo decisionale, e che non deve fondarsi su informazioni esterne al record senza garantire informazione e contraddittorio. Il CIArb, nella versione aggiornata della Guideline del settembre 2025, ha consolidato un approccio che combina rischi, raccomandazioni generali, poteri del tribunale, uso dell'IA da parte degli arbitri e modelli di accordo o *procedural order*. La SCC Guide del 2024 e la VIAC Note del 2025 rafforzano la stessa direzione, ossia riservatezza, qualità dell'output, verifica umana, integrità, disclosure proporzionata e disciplina nella case management conference.

L'esperienza dell'AAA-ICDR "AI Arbitrator", disponibile dal novembre 2025 per alcuni casi *construction documents-only*, mostra che l'istituzionalizzazione dell'IA nell'arbitrato non è più ipotesi teorica. La qualificazione commerciale dello strumento non deve però confondere la struttura giuridica, posto che il modello assiste, analizza e prepara, ma un arbitro umano resta coinvolto nella revisione e nell'emissione del lodo. Proprio perché queste sperimentazioni sono iniziate, diventa essenziale pretendere validazione, tracciabilità, "spiegabilità" operativa, possibilità di correzione e controllo effettivo dell'influenza dell'output sul ragionamento umano.

Security by design nel procedimento arbitrale

Security by design non significa aggiungere una misura tecnica alla fine dell'organizzazione del procedimento. Significa progettare simultaneamente tecnologia, regole procedurali e responsabilità. Il punto di partenza deve essere il caso d'uso (quale attività il sistema migliora, quali dati riceve, quale soggetto lo usa, quale decisione può influenzare, quali fornitori intervengono, quali azioni può compiere, chi controlla l'output, quali log sono conservati, cosa accade se il servizio è indisponibile o compromesso, come vengono esportati e cancellati dati e log). La domanda corretta non è se l'IA sia

[11] ISO/IEC 27001:2022, Information security management systems; ISO/IEC 42001:2023, Artificial intelligence management system; ISO/IEC 23894:2023, Artificial intelligence risk management. Le norme ISO sono richiamate come riferimenti tecnico-organizzativi e non come presupposto necessario di certificazione.

[12] https://cdn.arbitration-icca.org/s3fs-public/document/media_document/ICCA-reports-no-6-icca-nyc-bar-cpr-protocol-cybersecurity-international-arbitration-2022-edition.pdf

[13] ICCA-IBA, Roadmap to Data Protection in International Arbitration, 2022 https://www.arbitration-icca.org/icca-reports-no-7-icca-iba-roadmap-data-protection-international-arbitration

"ammessa", ma a quali condizioni un determinato impiego sia compatibile con riservatezza, contraddittorio, integrità della prova e non delega del giudizio.

La prima area è l'**identità**. Gli account condivisi sono incompatibili con accountability (attribuzione delle attività a soggetti identificabili e responsabili) e tracciabilità (possibilità di ricostruire chi ha fatto cosa, quando e con quali strumenti). Ogni partecipante deve usare credenziali nominative (utenze personali non condivise), MFA resistente al phishing (autenticazione a più fattori progettata per resistere anche a tentativi di sottrazione delle credenziali), ruoli separati (profili di accesso distinti in base alle funzioni svolte), minimo privilegio (attribuzione dei soli permessi strettamente necessari), revoca immediata al termine dell'incarico o in caso di compromissione (disattivazione tempestiva degli accessi quando il soggetto non è più autorizzato o l'account è a rischio), access review (verifica periodica degli utenti abilitati e dei relativi permessi) e monitoraggio di accessi anomali (controllo di accessi insoliti per orario, luogo, dispositivo o comportamento). Lo zero trust (modello di sicurezza fondato sull'assenza di fiducia implicita e sulla verifica continua di identità, dispositivo, contesto e autorizzazioni) non deve essere inteso come prodotto, ma come rifiuto della fiducia automatica fondata sulla sola appartenenza alla rete locale. Anche una struttura piccola può applicarne la logica partendo da inventario (elenco aggiornato di utenti, dispositivi, sistemi e dati), segregazione dei fascicoli (separazione degli ambienti e dei documenti relativi ai diversi procedimenti), autenticazione forte (meccanismi di accesso più robusti della sola password) e controllo dei privilegi amministrativi (limitazione e verifica degli account con poteri elevati sui sistemi).

La seconda area è il **dato**. Il fascicolo dovrebbe vivere in un repository segregato per procedimento (archivio digitale separato per ciascun arbitrato), con cifratura in transito e a riposo (protezione crittografica dei dati sia durante la trasmissione sia quando sono memorizzati), gestione sicura delle chiavi (controllo rigoroso delle chiavi crittografiche necessarie per accedere ai dati), backup cifrati (copie di sicurezza protette da crittografia), log protetti (registrazioni degli accessi e delle operazioni preservate da alterazioni o cancellazioni non autorizzate) e regole di retention (tempi e modalità di conservazione e cancellazione dei dati). L'email dovrebbe essere canale di notifica e non archivio principale, perché la dispersione degli allegati nelle caselle di posta riduce il controllo effettivo sul fascicolo. I link devono scadere (non devono restare validi indefinitamente), gli allegati sensibili devono transitare su spazi controllati (ambienti governati dall'istituzione o dai soggetti autorizzati) e le comunicazioni urgenti devono avere un canale alternativo concordato (ad esempio telefono, PEC o altro canale previamente validato). La cifratura, da sola, non basta: non protegge dall'account compromesso (quando l'attaccante usa credenziali valide) né da un'applicazione autorizzata che esfiltra dati (quando un software formalmente abilitato trasferisce informazioni verso l'esterno). Deve essere integrata con autorizzazioni (regole su chi può accedere a cosa), monitoraggio (controllo degli eventi rilevanti) e limitazione delle esportazioni (restrizioni al download, alla copia o al trasferimento dei documenti).

La terza area è l'**IA generativa e agentic**a. I documenti del fascicolo devono essere trattati come contenuti non fidati (input potenzialmente manipolati o contenenti istruzioni ostili); le istruzioni di sistema devono essere separate e protette (le regole interne che guidano il comportamento del modello non devono essere confuse con i documenti da analizzare né esposte agli utenti); i connettori devono essere allowlisted (devono poter operare solo verso strumenti, archivi o servizi espressamente autorizzati); le azioni esterne devono richiedere conferma umana (nessun invio, deposito, modifica o cancellazione dovrebbe avvenire senza approvazione preventiva di un soggetto responsabile); il corpus RAG deve essere versionato e scrivibile solo da soggetti autorizzati (la base documentale usata per

recuperare informazioni e costruire il contesto della risposta deve avere versioni controllate e accessi in scrittura limitati); ogni output rilevante deve citare fonti verificabili (deve indicare i documenti o i passaggi da cui trae le informazioni); modello, versione, prompt critici e impostazioni devono essere tracciati (deve essere possibile ricostruire quale sistema sia stato usato, con quali istruzioni e con quale configurazione); i test devono includere prompt injection (tentativi di manipolare il modello mediante istruzioni inserite nei contenuti), esfiltrazione (tentativi di estrarre dati riservati), poisoning (contaminazione dei dati o del corpus per orientare le risposte) e bypass (tentativi di aggirare regole, filtri o limitazioni del sistema). L'ambiente deve essere separato dagli altri procedimenti e i log devono essere governati come dati sensibili, non accumulati senza finalità e termine.

La quarta area è il **fornitore**. La *due diligence* (verifica preventiva dell'affidabilità tecnica, organizzativa e giuridica del fornitore) non può limitarsi a un questionario generico o a una certificazione. Devono essere esaminati architettura (struttura tecnica del servizio), isolamento dei tenant (separazione tra i dati e gli ambienti dei diversi clienti), uso dei dati per training o miglioramento (eventuale impiego dei dati del procedimento per addestrare, affinare o migliorare il modello o il servizio), localizzazione (luogo in cui dati e sistemi sono collocati), trasferimenti (passaggi di dati verso Paesi terzi o altri soggetti), subfornitori (soggetti ulteriori coinvolti nell'erogazione del servizio), autenticazione (modalità di verifica dell'identità degli utenti), amministratori privilegiati (soggetti con poteri elevati sui sistemi), cifratura (protezione crittografica dei dati), logging (registrazione degli accessi e delle operazioni), retention (tempi di conservazione dei dati), vulnerability disclosure (modalità con cui il fornitore riceve, gestisce e comunica le vulnerabilità), patching (aggiornamento e correzione delle vulnerabilità), incident notification (comunicazione tempestiva degli incidenti di sicurezza), business continuity (capacità di mantenere le funzioni essenziali anche in caso di crisi), disaster recovery (ripristino dei sistemi dopo eventi gravi), portabilità (possibilità di esportare dati e documenti in formato utilizzabile), exit strategy (modalità ordinate di cessazione del rapporto e migrazione verso altro fornitore), cancellazione verificabile (prova dell'effettiva eliminazione dei dati), cambi di versione del modello (modifiche del sistema IA che possono incidere sui risultati) e diritto di sospendere funzioni rischiose. Una certificazione può ridurre l'incertezza, ma non sostituisce la valutazione del caso d'uso concreto.

Resilienza e incident response con dimensione procedurale

La sicurezza assoluta non esiste. Per questo l'arbitrato digitale deve essere progettato anche per il giorno in cui il controllo fallisce. Resilienza significa mantenere o ripristinare funzioni essenziali preservando diritti, prove e contraddittorio. Non coincide con il semplice ritorno online della piattaforma. Un procedimento può riprendere tecnicamente e tuttavia restare compromesso se non sono stati preservati log, file originali, hash, timeline, comunicazioni e possibilità di verifica.

Un piano di *incident response* arbitrale (cioè una procedura organizzata per gestire un incidente di sicurezza che incida sul procedimento arbitrale) deve coordinare rilevazione (individuazione dell'evento anomalo o della violazione), *triage* (prima valutazione della gravità, dell'urgenza e dell'estensione dell'incidente), preservazione (conservazione delle prove digitali, dei log, dei file originali e degli elementi utili alla ricostruzione), contenimento (misure per impedire che l'incidente si estenda ulteriormente), continuità (garanzia della prosecuzione delle funzioni essenziali del procedimento), valutazione giuridica (verifica degli obblighi normativi, contrattuali e procedurali conseguenti all'incidente), comunicazione (informazione ai soggetti che devono essere coinvolti o

avvisati), eradicazione (rimozione della causa tecnica dell'incidente, ad esempio malware, account compromessi o configurazioni vulnerabili), ripristino (ritorno a condizioni operative sicure) e *lessons learned* (analisi successiva dell'incidente per correggere procedure, controlli e responsabilità). Deve inoltre coordinare più orologi, ossia più termini e scansioni temporali rilevanti: il GDPR prevede, se ricorrono i presupposti, la notifica all'autorità di controllo entro 72 ore dalla conoscenza della violazione dei dati personali; NIS2 può imporre pre-notifica, notifica e relazione finale secondo tempi propri; il contratto con il fornitore può prevedere termini ulteriori di segnalazione, cooperazione o rimedio; il tribunale arbitrale deve decidere se sospendere termini, ripetere un deposito, cambiare canale di comunicazione o disporre misure di preservazione. L'incidente non è solo un evento tecnico, ma un fatto procedurale, perché può incidere su riservatezza, integrità della prova, contraddittorio, termini e validità degli atti.

La *tabletop exercise* (esercitazione di simulazione svolta "a tavolino", senza attendere un incidente reale) è uno strumento particolarmente efficace per rivelare le zone grigie, cioè le aree in cui non è chiaro chi debba decidere, con quali poteri, in quali tempi e sulla base di quali informazioni. Simulare un account di arbitro compromesso (uso illecito o non autorizzato delle credenziali di un arbitro), un *ransomware* sul repository (software malevolo che cifra o blocca l'archivio digitale del procedimento), una trascrizione pubblicata per errore, un *deepfake* prodotto come prova (contenuto audio, video o immagine artificialmente generato o alterato), una *prompt injection* che accede a documenti di altro procedimento (istruzione malevola inserita in un documento o input per indurre il sistema di IA a violare le regole di accesso), o l'indisponibilità del provider nel giorno del deposito (mancato funzionamento del fornitore della piattaforma o del servizio digitale nel momento in cui un atto deve essere depositato), consente di scoprire chi decide, con quali informazioni, in quali tempi e con quali poteri. Il risultato utile non è il punteggio dell'esercizio, ma l'elenco delle decisioni che nessuno sapeva di dover prendere.

La procedural order come dispositivo di sicurezza

La prima riunione di organizzazione del procedimento arbitrale non dovrebbe riguardare soltanto calendario, lingua, termini e modalità di scambio degli atti. Dovrebbe disciplinare anche il livello di riservatezza dei dati trattati, la piattaforma digitale autorizzata, l'identificazione degli utenti, l'uso dell'autenticazione a più fattori, i permessi di accesso, i formati dei documenti, le firme digitali, i controlli sull'integrità dei file, gli strumenti di intelligenza artificiale consentiti o vietati, il trattamento delle istruzioni date ai sistemi di IA e delle risposte da essi generate, i casi in cui l'uso dell'IA deve essere dichiarato, la verifica delle fonti, l'autenticità di audio, video e immagini, il ruolo dei fornitori, le udienze da remoto, le registrazioni, la gestione degli incidenti, i canali alternativi di comunicazione, la conservazione dei documenti e la loro cancellazione finale. In questo modo l'ordinanza procedurale non è soltanto uno strumento di organizzazione del processo arbitrale, ma diventa anche uno strumento di sicurezza e di governo dell'uso dell'intelligenza artificiale.

La **comunicazione dell'uso dell'IA** deve essere concreta e utile, non meramente formale. Non sarebbe efficiente imporre alle parti di dichiarare qualsiasi funzione automatizzata, anche minima, perché ciò potrebbe appesantire inutilmente il procedimento e, in alcuni casi, incidere sulla riservatezza, sul segreto professionale o sulla strategia difensiva. Tuttavia, l'assenza totale di informazione sull'uso dell'IA può compromettere il contraddittorio e la fiducia nel procedimento. La comunicazione dovrebbe quindi essere richiesta quando l'IA genera o modifica una prova, incide in modo significativo

sulla selezione o valutazione dei documenti, introduce informazioni esterne al fascicolo, contribuisce alla formazione di un'opinione peritale, produce una parte rilevante della motivazione, oppure quando la conoscenza del suo impiego è necessaria per consentire alla controparte di contestare, verificare o riprodurre il risultato. La comunicazione deve inoltre essere effettuata quando è prevista dalla legge, dal regolamento arbitrale, da un ordine del tribunale arbitrale o da un obbligo professionale.

Il principio di non delega deve essere tradotto in regole operative. L'arbitro può usare strumenti di supporto, anche basati su IA, ma deve conservare una conoscenza effettiva del fascicolo, valutare personalmente le prove, controllare direttamente le fonti giuridiche, formare autonomamente il proprio convincimento, assumere la responsabilità della motivazione e partecipare realmente alla deliberazione. Una bozza generata da un sistema di IA non è una semplice tastiera evoluta: il sistema seleziona informazioni, le ordina e le formula in un determinato modo. Quanto più il testo prodotto dall'IA entra nel lodo, tanto maggiore deve essere il controllo umano e tanto più difficile diventa sostenere che l'IA sia stata utilizzata come mero strumento materiale di supporto.

Conclusioni

La cybersicurezza dell'arbitrato non è un tema secondario, né una questione riservata agli specialisti informatici. È una componente della qualità della giustizia arbitrale. Serve a proteggere la riservatezza delle parti, l'equilibrio del contraddittorio, l'affidabilità delle prove e la possibilità che la decisione finale sia davvero il risultato di un giudizio umano responsabile.

L'intelligenza artificiale rende questa esigenza ancora più evidente. Il rischio non riguarda soltanto la perdita o la violazione dei dati, ma anche il modo in cui quei dati vengono letti, selezionati, sintetizzati e interpretati. Un fascicolo può essere integro, ma essere analizzato da un sistema non affidabile; una prova può essere autentica, ma essere oscurata da contenuti artificiali o fuorvianti; una decisione può essere formalmente firmata da arbitri umani, ma risultare sostanzialmente influenzata da strumenti non controllati.

La risposta non deve essere la diffidenza assoluta verso la tecnologia. Deve essere, piuttosto, una governance consapevole e integrata. Il diritto stabilisce obblighi, garanzie e limiti. La tecnologia aiuta a prevenire, individuare e contenere i rischi. L'organizzazione assegna responsabilità, tempi e competenze. Le regole procedurali rendono le misure chiare, verificabili e opponibili alle parti. La competenza umana conserva ciò che nessun sistema automatizzato può sostituire: giudizio, responsabilità, capacità critica e possibilità di dissenso.

Progettare la sicurezza fin dall'inizio, nell'arbitrato, significa tenere insieme questi elementi. Non è un appesantimento del procedimento, ma una condizione per renderlo più solido, più affidabile e più adatto al contesto digitale. L'obiettivo non è rinunciare all'intelligenza artificiale, ma usarla senza consegnarle ciò che deve restare umano: la valutazione delle prove, la costruzione del ragionamento e la responsabilità della decisione.

Resta un rischio meno visibile di una violazione informatica, ma forse ancora più profondo e subdolo che dipende dall'uso continuativo dell'IA e consiste in un indebolimento progressivo delle capacità necessarie per controllarla. Per questo l'arbitrato potrà trarre reale vantaggio dall'intelligenza artificiale

solo se saprà preservare ciò che le parti cercano quando scelgono un arbitro: non una macchina più veloce, ma un giudizio competente, indipendente, responsabile e affidabile.