PENSAI — Decision Governance Engine

Monolith v3.2.1

Dichiarazione di Conformità al Regolamento (UE) 2024/1689 — AI Act

Versione documento: **1.0** | Classificazione: **Pubblica** | Data: **Aprile 2025**

Il presente documento è una dichiarazione tecnica di conformità a carattere orientativo. Non costituisce valutazione giuridica vincolante. Per certificazione ufficiale o notifica a un organismo notificato è richiesta analisi legale specifica per il singolo contesto applicativo.

1. Executive Summary

PENSAI è un framework deterministico di supporto decisionale multi-criterio (Decision Governance Engine) progettato con conformità AI Act integrata. Il sistema non è un sistema autonomo di intelligenza artificiale: produce raccomandazioni strutturate, esplicabili e tracciabili, sottoposte obbligatoriamente a supervisione umana prima di qualunque effetto giuridico o operativo.

L'architettura implementa nativamente i requisiti degli Articoli 9, 10, 11, 12, 13, 14, 15 e 61 del Regolamento (UE) 2024/1689. La conformità è verificabile a livello di codice sorgente e di output di sistema, senza dipendenza da dichiarazioni esterne.

Dimensione di Conformità	Stato
Pratiche vietate (Artt. 5–6)	**CONFORME — Assenti per design**
Supervisione umana (Art. 14)	**CONFORME — Hard enforcement**
Tracciabilità e audit (Art. 12)	**CONFORME — Trace ID per step**
Trasparenza ed esplicabilità (Art. 13)	**CONFORME — Breakdown criterio per criterio**
Gestione del rischio (Art. 9)	**CONFORME — Classificazione dinamica**
Integrità crittografica (Art. 15)	**CONFORME — SHA-256 su input/output**
Governance dei dati (Art. 10)	**PARZIALE — Dipende dall'operatore**
Monitoraggio post-mercato (Art. 61)	**CONFORME — Log continuo con alert**

2. Inquadramento Normativo

2.1 Classificazione del sistema

PENSAI rientra nella categoria dei sistemi di supporto decisionale deterministici. Non apprende dai dati, non modifica il proprio comportamento autonomamente, non prende decisioni senza supervisione umana. Non rientra nelle categorie di sistemi AI con pratiche vietate ai sensi dell'Articolo 5 del Regolamento.

La classificazione di rischio effettiva dipende dal contesto di deployment dell'operatore. PENSAI include al proprio interno un modulo di classificazione dinamica del rischio (AIRiskManager) che mappa ogni decisione sulla tassonomia dell'AI Act: Minimal Risk, Limited Risk (Annex IV), High Risk (Annex III).

2.2 Ruolo operativo

PENSAI opera esclusivamente come "decision-support layer": produce una raccomandazione strutturata (VIA_UNICA) corredata da score, confidence, breakdown per criterio e uncertainty map. Nessun output acquista effetto giuridico, economico o operativo senza esplicita supervisione e override da parte di un operatore umano autorizzato.

2.3 Pratiche vietate — Articolo 5

Il sistema non implementa, non supporta e non consente le seguenti pratiche, proibite in modo assoluto dal Regolamento:

- Manipolazione subliminale o sfruttamento di vulnerabilità cognitive
- Social scoring basato su comportamento individuale
- Identificazione biometrica remota in tempo reale
- Profilazione o inferenza di caratteristiche sensibili (etnia, opinioni politiche, salute)

L'architettura del sistema è priva di qualunque componente che possa implementare tali pratiche. La conformità è strutturale (compliance by design), non procedurale.

3. Architettura del Sistema

3.1 Pipeline decisionale

PENSAI esegue una pipeline deterministica e immutabile articolata nei seguenti stadi sequenziali, ciascuno registrato in audit trail con timestamp e trace ID univoco:

Stadio	Descrizione
SECURITY_CHECK	Validazione input: injection detection (regex + Aho-Corasick), flooding check (max 20 opzioni), score manipulation detection (variance analysis).
INTEGRITY_CHECK	Verifica crittografica hash SHA-256 della rubrica di criteri. Rilevamento di qualunque modifica non autorizzata ai pesi o alle soglie.
FILTER	Eliminazione delle opzioni che violano hard constraints dichiarati. Tracciamento del numero di opzioni rimaste.
SCORE	Scoring ponderato multi-criterio con: fattore di affidabilità (provenance.reliability), penalità incertezza (UncertaintyMap), criteri critici non compensabili, inversione opzionale per criteri di costo.
DECISION	Determinazione VIA_UNICA (opzione consigliata), NO_CASE (nessuna opzione soddisfa le soglie) o HUMAN_OVERRIDE (supervisore umano ha modificato la

Stadio	Descrizione
	raccomandazione).
RISK EVAL	Classificazione dinamica del rischio (MINIMAL / LIMITED / HIGH). Decisioni HIGH richiedono human_override obbligatorio: il sistema genera ValidationException se non fornito.
SEALED	Apposizione sigillo crittografico SHA-256 sull'output finale. Il sigillo può essere verificato da terze parti in qualunque momento.
COMPLIANCE	Esecuzione moduli AI Act: risk management (Art. 9), technical documentation (Art. 11), regulatory log (Art. 12), post-market monitoring (Art. 61).

3.2 Principi architetturali

- Immutabilità by default: tutti gli oggetti di dominio sono dichiarati final e readonly. Nessuna trasformazione di stato è possibile in-place.
- Type safety: il sistema è scritto con strict_types=1. Tutti i valori sono validati alla costruzione degli oggetti di dominio.
- Architettura esagonale: separazione netta tra domain layer (logica pura), application layer (casi d'uso) e infrastructure layer (sicurezza, monitoring, logging).
- Determinismo: a parità di input, l'output è sempre identico. Il tie-breaking utilizza l'ID dell'opzione come comparatore stabile.
- Sicurezza by construction: la validazione avviene al confine del sistema, prima di qualunque elaborazione.

4. Mapping degli Articoli AI Act

4.1 Articolo 9 — Gestione del Rischio

Il modulo AIRiskManager implementa un sistema di gestione del rischio su tutto il ciclo di vita della decisione, conforme all'Articolo 9. La classificazione avviene dinamicamente per ogni esecuzione, non staticamente a livello di prodotto.

Livello di Rischio	Trigger	Misure attivate
MINIMAL	Priority LOW, nessuna deriva	Audit trail, sigillo crittografico
LIMITED	Priority NORMAL, o deriva rilevata da MINIMAL	+ Obblighi di trasparenza
HIGH	Priority HIGH/CRITICAL, o deriva da stato precedente	+ Human override obbligatorio, monitoring continuo PER_DECISION, alert immediato

La trend analysis confronta il livello di rischio della decisione corrente con il risultato precedente, rilevando andamenti crescenti o decrescenti. Il feedback loop per decisioni HIGH prevede review ogni ora con trigger su SCORE_DRIFT e RANKING_INSTABILITY.

4.2 Articolo 10 — Governance dei Dati

Il sistema supporta la governance dei dati attraverso un meccanismo di provenienza (provenance) associato a ciascuna opzione. Ogni opzione può dichiarare la propria sorgente dati e un indice di affidabilità (reliability, valore 0.0–1.0) che influenza direttamente il calcolo dello score.

- Se la provenienza è assente, il sistema applica un reliability di default (0.8) e registra un evento PROVENANCE_MISSING nell'audit trail.
- Se la provenienza è presente ma malformata (source o reliability assenti), viene sollevata una ValidationException che blocca il processo.
- Il sistema non accede direttamente a dataset di addestramento — non essendo un sistema ML, non dispone di parametri appresi.

NOTA OPERATORE: La piena conformità all'Articolo 10 richiede che l'operatore garantisca la qualità, la rappresentatività e la correttezza dei dati di input forniti a PENSAI. Il sistema fornisce gli strumenti tecnici (provenance tracking, reliability scoring) ma non può sostituire una policy di governance dei dati a livello organizzativo.

4.3 Articolo 11 — Documentazione Tecnica

Il modulo AITechDoc genera automaticamente documentazione tecnica strutturata ad ogni esecuzione del sistema. La documentazione è inclusa nell'output e copre:

- Identità e versione del sistema (PENSAI_MONOLITH_3.2.1)
- Scopo dichiarato: "Deterministic multi-criteria decision analysis with governance controls"
- Architettura: pattern esagonale, strati Domain / Application / Infrastructure
- Specifica degli input: rubrica di criteri con pesi in basis points (somma = 10.000), opzioni con score per criterio [0–1], hard constraints, livello di priorità
- Specifica degli output: VIA_UNICA con score e confidence, ranking completo con breakdown per criterio, drift report
- Algoritmi: scoring ponderato con reliability adjustment e uncertainty penalty, Kendall Tau per stabilità del ranking, confidence multi-fattore
- Limitazioni dichiarate: non autonomo, qualità output limitata dalla qualità input, nessuna capacità di apprendimento, max 20 opzioni
- Feature di compliance: human override, audit trail, explainability, integrity verification, drift monitoring, risk classification

4.4 Articolo 12 — Tenuta dei Registri

Il modulo AIRegulatoryLog registra ogni decisione in un log regolatorio persistente in formato JSONL (append-only, file-locked). Ogni record include:

Campo	Contenuto
record_type / timestamp	Tipo "AI_DECISION", timestamp UTC ISO 8601
decision_id / version	Identificativo univoco della decisione e versione della rubrica
submitted_by	Identità del richiedente (fornita dall'operatore)
decision_type / selected_option	Tipo di esito (VIA_UNICA / NO_CASE / HUMAN_OVERRIDE) e opzione selezionata
score / confidence	Punteggio finale e indice di confidenza della raccomandazione
risk_level / human_override	Classificazione AI Act e presenza di override umano
integrity_verification	Hash SHA-256 rubrica e output, stato di verifica del sigillo
audit_trail_summary	Lista degli stadi eseguiti e durata totale di elaborazione in millisecondi

4.5 Articolo 13 — Trasparenza ed Esplicabilità

Ogni output di PENSAI include un breakdown dettagliato per criterio che espone, per ciascuna opzione valutata:

- raw_score: punteggio grezzo fornito in input
- effective_score: punteggio dopo eventuale inversione (criteri di costo)
- weighted_score: contributo al punteggio totale in base al peso del criterio
- reliability: fattore di affidabilità della fonte dati
- uncertainty_penalty: penalità applicata per incertezza dichiarata
- critical_pass: indicatore booleano se il criterio critico è soddisfatto

Questa struttura garantisce che chiunque — operatore, supervisore, autorità di controllo — possa ricostruire passo per passo come si è formata la raccomandazione. Il livello di trasparenza supera i requisiti minimi dell'Articolo 13.

4.6 Articolo 14 — Supervisione Umana

> *La supervisione umana è implementata come hard enforcement a livello architetturale, non come raccomandazione procedurale. Non è possibile ottenere output valido da PENSAI per decisioni ad alto rischio senza fornire esplicita autorizzazione umana.*

Il meccanismo di supervisione umana funziona come segue:

- Ogni decisione classificata HIGH RISK richiede il campo human_override nell'input, contenente: authorized_by (identità del supervisore), role (ruolo autorizzato, es. "admin" o "supervisor"), reason (motivazione testuale), e opzionalmente una firma digitale.
- Se human_override è assente per una decisione HIGH RISK, il sistema solleva una ValidationException con messaggio esplicito: "HIGH RISK decision requires mandatory human oversight."
- L'override è tracciato in un OverrideChain immutabile che conserva l'intera catena di decisioni umane, inclusa la decisione originale del sistema.
- Il supervisore può confermare la raccomandazione del sistema o sostituirla con una scelta diversa, entrambe le opzioni sono registrate nel regulatory log.

4.7 Articolo 15 — Accuratezza, Robustezza e Sicurezza Informatica

Accuratezza e Robustezza

- Confidence multi-fattore: l'indice di confidenza è calcolato considerando separazione tra score del primo e secondo classificato, qualità della provenienza, varianza dei punteggi e penalità da incertezza.
- Criteri non compensabili: i criteri marcati come critical con soglie absolute o percentile non possono essere "compensati" da punteggi alti in altri criteri. L'opzione viene esclusa se non soddisfa la soglia.
- Drift detection: il sistema calcola il Kendall Tau tra il ranking della decisione corrente e quello precedente. Un valore < 0.7 indica instabilità e attiva richiesta di revisione umana.
- Gestione incertezza (UncertaintyMap): ogni opzione può dichiarare variabili di incertezza con scenari best/expected/worst. Un worst case "FATAL" esclude l'opzione dalla selezione.

Sicurezza Informatica

- Injection detection: il sistema scansiona ricorsivamente tutti i campi di input (max 10.000 nodi, max 50 livelli di profondità) per rilevare pattern di prompt injection, jailbreak e bypass.

- Score manipulation detection: la varianza dei punteggi di ogni opzione è analizzata. Punteggi uniformemente alti con varianza < 0.005 sono rilevati come manipolazione e bloccati.
- Option flooding: il numero massimo di opzioni accettate per singola richiesta è 20. Richieste con più opzioni sollevano SecurityException.
- Integrità crittografica: l'output è sigillato con SHA-256 su payload normalizzato e ordinato. Il sigillo include hash separato dell'input e hash della rubrica.

4.8 Articolo 61 — Monitoraggio Post-Mercato

Il modulo AIPostMonitor implementa il monitoraggio post-mercato continuo richiesto dall'Articolo 61. Per ogni decisione viene registrato un monitoring entry che include:

- Performance indicators: score, confidence, numero di opzioni valutate, tempo di esecuzione in millisecondi
- Quality indicators: stato di drift, necessità di revisione, indice di stabilità del ranking (Kendall Tau)
- Incident indicators: presenza di decisione HIGH RISK, applicazione di human override, incidenti di sicurezza

Per decisioni HIGH RISK o che richiedono revisione, il sistema attiva automaticamente un alert ad alta severità ("High-risk AI decision detected") distribuito tramite event dispatcher e registrato nel sistema di logging.

5. Gap Analysis e Condizioni di Conformità

La seguente analisi identifica le aree in cui la conformità di PENSAI è completa a livello architetturale e quelle che richiedono interventi da parte dell'operatore per il completamento.

Area	Stato	Dettaglio / Azione richiesta
Supervisione umana	**COLMATO**	Hard enforcement nativo: impossibile bypassare per decisioni HIGH RISK.
Tracciabilità	**COLMATO**	Audit trail con trace ID univoco per ogni stadio della pipeline, timestamp UTC.
Integrità output	**COLMATO**	SHA-256 su input, output e rubrica. Verifica da terzi possibile senza accesso al sistema.
Esplicabilità	**COLMATO**	Breakdown criterio per criterio con tutti i fattori contributivi esposti nell'output.
Gestione rischio	**COLMATO**	Classificazione dinamica per decisione, trend analysis, monitoring continuo.
Governance dei dati	**PARZIALE**	AZIONE OPERATORE: definire policy di qualità e rappresentatività dei dati di input. Il sistema offre provenance tracking e reliability scoring come strumenti.
Validazione formale	**PARZIALE**	AZIONE OPERATORE: condurre validazione su dataset reali per il dominio applicativo specifico prima del go-live in contesti ad alto rischio.
Classificazione legale	**PARZIALE**	AZIONE OPERATORE: formalizzare la classificazione AI Act per il caso d'uso specifico (credito, selezione personale, PA, ecc.) con supporto legale.
Notifica organismo notificato	**N/A**	Richiesta solo per sistemi ad alto rischio in Annex III con uso in contesti specifici. Dipende dal deployment dell'operatore.

6. Strategia di Compliance per l'Operatore

6.1 Livello minimo (obbligatorio per deployment)

- Definire formalmente il ruolo di PENSAI nel processo decisionale: strumento di supporto vs. sistema decisionale autonomo (deve essere sempre la prima opzione).
- Implementare un sistema di Identity Management per il campo submitted_by e per i ruoli di supervisione (human_override.authorized_by, human_override.role).
- Garantire che i log regolatori (ai_act_regulatory.jsonl) siano archiviati, protetti e accessibili per almeno 10 anni per sistemi ad alto rischio.
- Rendere disponibile all'utente finale, in forma comprensibile, l'informazione che sta interagendo con un sistema di supporto AI e la spiegazione della raccomandazione ricevuta.

6.2 Livello avanzato (raccomandato per domini sensibili)

- Implementare un dataset validation layer upstream a PENSAI per garantire la qualità dei dati di input prima dell'elaborazione.
- Condurre audit periodici (almeno annuali) della rubrica di criteri e dei pesi per verificarne la correttezza e l'assenza di bias sistematici.
- Integrare il monitoraggio post-mercato (post_market.jsonl) con un sistema centralizzato di osservabilità (es. OpenTelemetry-compatible, già supportato dall'architettura).
- Formalizzare la documentazione tecnica auto-generata in un Technical File conforme all'Allegato IV dell'AI Act per i contesti ad alto rischio.

7. Conclusioni e Posizionamento Strategico

PENSAI è compatibile per architettura con il Regolamento (UE) 2024/1689 — AI Act. Le caratteristiche di conformità sono integrate nel design del sistema, non aggiunte a posteriori, e sono verificabili direttamente dall'output di ogni singola esecuzione.

Il posizionamento formale del prodotto è quello di Decision Governance Layer AI Act-ready: un livello di governance decisionale che può essere integrato in processi esistenti garantendo tracciabilità, esplicabilità, supervisione umana e monitoraggio continuo in conformità con il Regolamento europeo.

La conformità completa, in particolare per deployment in domini ad alto rischio ai sensi dell'Allegato III (credito, selezione del personale, pubblica amministrazione), richiede l'integrazione delle misure a carico dell'operatore descritte nella Sezione 5 e nella Sezione 6.

Dichiarazione Finale

PENSAI non è un sistema autonomo di intelligenza artificiale. È un framework deterministico di supporto decisionale coerente con i principi di IA antropocentrica e di controllo umano previsti dal Regolamento (UE) 2024/1689.