

HLEG Trustworthy checklist

HLEG, High-Level Expert Group on AI: una **lista di controllo per la valutazione di una AI affidabile**.

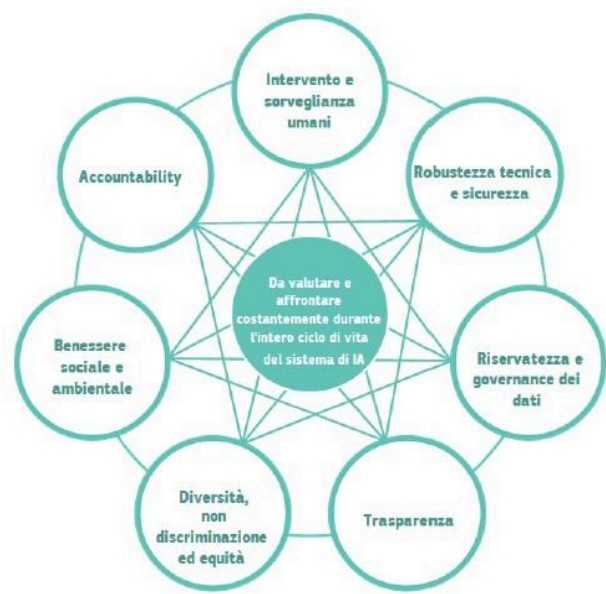


Figura 2: Interrelazione dei sette requisiti. Sono tutti di pari importanza, si avvalorano vicendevolmente e dovrebbero essere attuati e valutati durante l'intero ciclo di vita di un sistema di IA.

Intervento e sorveglianza umani

Diritti fondamentali

Domanda	Sì	No	N/A	Note
Nei casi d'uso che possono avere potenziali effetti negativi sui diritti fondamentali, è stata effettuata una valutazione d'impatto su tali diritti? Sono stati individuati e documentati i possibili compromessi realizzati tra i diversi principi e diritti?				

Il sistema di IA interagisce con il processo decisionale degli utenti finali umani (ad esempio azioni raccomandate o decisioni da prendere, presentazione di opzioni)?				
In questi casi, sussiste il rischio che il sistema di IA influenzi l'autonomia umana interferendo in modo non intenzionale con il processo decisionale dell'utente finale?				
È stato valutato se il sistema di IA debba comunicare agli utenti che una decisione, un contenuto, un consiglio o un risultato sono frutto di una decisione algoritmica?				
Nel caso in cui il sistema di IA disponga di un chat bot o di un sistema conversazionale, gli utenti finali umani sono messi a conoscenza del fatto che stanno interagendo con un agente non umano?				

Intervento umano

Domanda	Sì	No	N/A	Note,
Nel caso in cui il sistema di IA sia implementato in processi lavorativi, è stato valutato se la distribuzione delle mansioni tra il sistema di IA e i lavoratori umani consente interazioni significative e un controllo e una sorveglianza umani adeguati?				
Il sistema di IA migliora o aumenta le capacità umane?				
Sono state adottate misure di salvaguardia per prevenire l'eccessiva fiducia nel sistema di IA o l'eccessiva dipendenza da esso nei processi di lavoro?				

Sorveglianza umana

Domanda	Sì	No	N/A	Note
È stato valutato quale sarebbe il livello appropriato di controllo umano per il sistema di IA e il caso d'uso in questione?				
È possibile eventualmente descrivere il livello di controllo o di coinvolgimento umano? Chi è l'attore del "controllo umano" e quali sono i momenti o gli strumenti dell'intervento umano?				
Sono stati messi in atto meccanismi e misure per garantire tali possibili controllo o sorveglianza umani o per garantire che le decisioni siano prese sotto la responsabilità generale di esseri umani?				
Sono state adottate misure per rendere possibili verifiche e risolvere le problematiche relative alla gestione dell'autonomia dell'IA?				
Nel caso di un sistema di IA o di un caso d'uso autonomo o basato sull'autoapprendimento, sono stati messi in atto meccanismi più specifici di controllo e supervisione?				
Che tipo di meccanismi di rilevamento e di risposta sono stati stabiliti per valutare l'eventualità di esiti avversi?				
È stato previsto un "pulsante di arresto" o una procedura per interrompere in sicurezza un'operazione, se necessario? Questa procedura interrompe l'intero processo o una sua parte o delega il controllo a un essere umano?				

2. Robustezza tecnica e sicurezza

Resilienza agli attacchi e sicurezza

Domanda	Sì	No	N/A	Note
Sono state valutate le possibili forme di attacco a cui il sistema di IA potrebbe essere vulnerabile?				
In particolare, sono stati presi in considerazione i diversi tipi di vulnerabilità e la loro natura, ad esempio l'inquinamento dei dati, le infrastrutture fisiche, gli attacchi informatici?				
Sono state messe in atto misure o sistemi per garantire l'integrità e la resilienza del sistema di IA contro possibili attacchi?				
È stato valutato il comportamento del sistema in situazioni e ambienti imprevisti?				
È stato valutato se e in che misura il sistema possa avere un duplice uso? In caso affermativo, sono state adottate misure preventive adeguate (ad esempio, non pubblicare le ricerche o non distribuire il sistema)?				

Piano di emergenza e sicurezza generale

Domanda	Sì	No	N/A	Note
È stato accertato che il sistema abbia un piano di emergenza adeguato nell'eventualità di adversarial attack o di altre situazioni impreviste (ad esempio, procedure di commutazione tecnica o la richiesta di intervento di un operatore umano prima di procedere)?				
È stato preso in considerazione il livello di rischio posto dal sistema di IA in questo caso d'uso specifico?				

Sono stati messo in atto processi per misurare e valutare i rischi e la sicurezza?				
Sono state fornite le informazioni necessarie in caso di rischio per l'integrità fisica umana?				
È stata presa in considerazione la sottoscrizione di un'assicurazione contro eventuali danni provocati dal sistema di IA?				
Sono stati identificati i rischi potenziali per la sicurezza posti da (altri) prevedibili usi della tecnologia, compreso l'uso improprio accidentale o malevolo? Esiste un piano per mitigare o gestire questi rischi?				
È stato valutato se vi sono probabilità che il sistema di IA possa arrecare danni o nuocere agli utenti o a terze parti? In caso affermativo, si è provveduto a valutare la probabilità, i possibili danni, i soggetti colpiti e la gravità?				
Nel caso in cui vi sia il rischio che il sistema di IA provochi danni, sono state prese in considerazione le norme in materia di responsabilità e di protezione dei consumatori e come se ne è tenuto conto?				
È stato preso in considerazione il possibile impatto ambientale o il rischio per la salute animale?				
Nell'analisi dei rischi si è valutato se i problemi di sicurezza o di rete (ad esempio i rischi di sicurezza informatica) comportino rischi per la sicurezza o arrechino danni a causa di comportamenti non intenzionali del sistema di IA?				

Sono stati valutati i probabili effetti di un guasto del sistema di IA che comporti risultati errati, l'indisponibilità del sistema o risultati socialmente inaccettabili (ad es. pratiche discriminatorie)?				
Nell'ambito degli scenari delineati sopra, sono stati definiti livelli minimi e regole in base ai quali attivare piani alternativi/di emergenza?				
Sono stati definiti e testati i piani di emergenza?				

Precisione

Domanda	Sì	No	N/A	Note
È stato valutato quale livello e quale definizione di precisione sarebbero necessari nel contesto del sistema di IA e del caso d'uso?				
È stato valutato come misurare e assicurare la precisione?				
Sono state messe in atto misure per garantire che i dati utilizzati siano completi e aggiornati?				
Sono state messe in atto misure per valutare la necessità di dati aggiuntivi, ad esempio per migliorare la precisione o per eliminare le distorsioni?				
È stato valutato il danno che si verificherebbe se il sistema di IA facesse previsioni imprecise?				
Sono state previste modalità per misurare se il sistema effettua un numero inaccettabile di previsioni imprecise?				

Se il sistema effettua previsioni imprecise, è stata prevista una serie di misure per risolvere il problema?				
--	--	--	--	--

Affidabilità e riproducibilità

Domanda	Sì	No	N/A	Note
È stata messa in atto una strategia per monitorare e testare se il sistema di IA rispetta gli obiettivi, le finalità e le applicazioni previste?				
È stato verificato se è necessario prendere in considerazione contesti specifici o condizioni particolari per garantire la riproducibilità?				
Sono stati messi in atto processi o metodi di verifica per misurare e garantire diversi aspetti di affidabilità e riproducibilità?				
Sono stati messi in atto processi per descrivere quando, in date situazioni, un sistema di IA commette errori?				
Questi processi di prova e verifica dell'affidabilità dei sistemi di IA sono stati chiaramente documentati e resi operativi?				
Sono stati messi in atto meccanismi o metodi di comunicazione per garantire agli utenti (finali) che il sistema di IA è affidabile?				

3. Riservatezza e governance dei dati

Rispetto della riservatezza e protezione dei dati

Domanda	Sì	No	N/A	Note
Sono stati stabiliti meccanismi che, secondo i casi d'uso, consentano a terzi di segnalare problemi di riservatezza o di protezione dei dati riguardanti i processi del sistema di IA per la raccolta dati (sia per l'addestramento che per il funzionamento) e il loro trattamento?				
Sono stati valutati il tipo e la portata dei dati all'interno dei set (ad esempio se contengono dati personali)?				
Sono state prese in considerazione modalità per sviluppare il sistema di IA o addestrare il modello senza utilizzare dati personali o potenzialmente sensibili o utilizzandoli il meno possibile?				
Sono stati previsti meccanismi di notifica e controllo dei dati personali a seconda dei casi d'uso (come il consenso valido e la possibilità di revoca, se del caso)?				
Sono state adottate misure per aumentare la riservatezza, come la crittografia, l'anonimizzazione e l'aggregazione?				
Nel caso in cui esista un responsabile della protezione dei dati personali, è stato coinvolto nel processo fin dall'inizio?				

Qualità e integrità dei dati

Domanda	Sì	No	N/A	Note
---------	----	----	-----	------

Per quanto riguarda la gestione ordinaria e la governance dei dati, il sistema è conforme alle eventuali norme pertinenti (ad esempio ISO, IEEE) o ai protocolli più diffusi?				
Sono stati stabiliti meccanismi di sorveglianza relativi alla raccolta, all'archiviazione, all'elaborazione e all'utilizzo dei dati?				
È stato valutato in che misura si ha il controllo della qualità delle fonti di dati esterne utilizzate?				
Sono stati messi in atto processi per garantire la qualità e l'integrità dei dati? Sono stati presi in considerazione altri processi? Come viene verificato che i set di dati non vengano compromessi o violati?				

Accesso ai dati

Domanda	Sì	No	N/A	Note
Quali protocolli, processi e procedure sono stati seguiti per gestire e garantire una corretta governance dei dati?				
È stato valutato chi può accedere ai dati degli utenti e in quali circostanze?				
È stato accertato che queste persone siano qualificate, siano autorizzate ad accedere ai dati e che abbiano le competenze necessarie per comprendere la politica di protezione dei dati nei dettagli?				

È stato previsto un meccanismo di sorveglianza che registri sia chi ha avuto accesso ai dati sia i tempi, i luoghi, le modalità e le finalità di accesso?				
---	--	--	--	--

4. Trasparenza

Tracciabilità

Domanda	Sì	No	N/A	Note
Sono state messe in atto misure che garantiscono la tracciabilità?				
Nel caso di un sistema di IA basato su regole, è stato documentato il metodo di programmazione o il modo in cui il modello è stato costruito?				
Nel caso di un sistema di IA basato sull'apprendimento, è stato documentato il metodo di addestramento dell'algoritmo, quali dati di input sono stati raccolti e selezionati e come ciò è avvenuto?				
Nel caso di un sistema di IA basato su regole, sono stati documentati gli scenari o i casi utilizzati per testare e convalidare?				
Nel caso di un modello basato sull'apprendimento, sono state documentate le informazioni sui dati utilizzati per testare e convalidare?				

Sono stati documentati i risultati dell'algoritmo o le decisioni adottate in base ad esso, come pure altre possibili decisioni che potrebbero derivare da casi diversi (ad esempio per altri sottogruppi di utenti)?				
--	--	--	--	--

Spiegabilità

Domanda	Sì	No	N/A	Note
È stato valutato fino a che punto le decisioni prese dal sistema di IA e i loro risultati sono comprensibili?				
È stato accertato che la spiegazione dei motivi per cui un sistema ha adottato una certa decisione che ha prodotto determinati risultati possa essere resa comprensibile per gli utenti che la desiderano?				
È stato valutato in quale misura le decisioni del sistema influenzano i processi decisionali dell'organizzazione?				
È stato valutato il motivo per cui un particolare sistema è stato distribuito in un dato settore?				
È stato valutato il modello di business relativo al sistema (ad esempio in quale modo crea valore per l'organizzazione)?				
Il sistema di IA è stato progettato tenendo presente fin dall'inizio l'interpretabilità?				
Sono state effettuate ricerche sul modello più semplice e più interpretabile possibile per l'applicazione in questione e si è provato ad usarlo?				

È stata valutata la possibilità di analizzare i dati di addestramento e di test? È possibile modificarli e aggiornarli nel corso del tempo?				
È stato valutato se vi è la possibilità di esaminare l'interpretabilità dopo aver addestrato e sviluppato il modello o se è possibile accedere al flusso di lavoro interno del modello?				

Comunicazione

Domanda	Sì	No	N/A	Note
È stato comunicato agli utenti (finali), tramite una clausola di esclusione della responsabilità o altri mezzi, che stanno interagendo con un sistema di IA e non con un altro essere umano? Il sistema di IA è stato etichettato come tale?				
Sono stati messi in atto meccanismi per informare gli utenti in merito alle ragioni e ai criteri che determinano i risultati del sistema di IA?				
Tali informazioni sono comunicate in modo chiaro e comprensibile agli utenti destinatari?				
Sono stati stabiliti processi che prendono in considerazione i riscontri degli utenti e li utilizzano per adattare il sistema?				
Sono stati comunicati anche i rischi potenziali o percepiti, come le distorsioni (bias)?				
A seconda del caso d'uso, sono state prese in considerazione anche la comunicazione e la trasparenza nei confronti di terzi o del pubblico in generale?				

È stato precisato qual è lo scopo del sistema di IA e chi o cosa può trarre vantaggio dal prodotto/servizio?				
Gli scenari d'uso del prodotto sono descritti dettagliatamente e comunicati chiaramente, eventualmente utilizzando anche forme di comunicazione alternative, per garantire che l'informazione sia comprensibile e adeguata per gli utenti destinatari?				
A seconda del caso d'uso, si è riflettuto sulla psicologia umana e su potenziali limiti, come il rischio di confusione, il pregiudizio di conferma (confirmation bias) o la fatica cognitiva?				
Nel caso dello sviluppo, le caratteristiche, i limiti e le potenziali carenze del sistema di IA sono stati chiaramente comunicati a chiunque lo distribuisca in un prodotto o servizio?				
Nel caso della distribuzione, le caratteristiche, i limiti e le potenziali carenze del sistema di IA sono stati chiaramente comunicati all'utente finale o al consumatore?				

5. Diversità, non discriminazione ed equità

Evitare distorsioni inique

Domanda	Sì	No	N/A	Note
È stata prevista una strategia o una serie di procedure per evitare di creare o rafforzare distorsioni inique (unfair bias) nel sistema di IA, sia				

per quanto riguarda l'uso dei dati di input che per la progettazione dell'algoritmo?				
Sono state valutate e riconosciute le possibili limitazioni derivanti dalla composizione dei set di dati utilizzati?				
Sono state prese in considerazione la diversità e la rappresentatività degli utenti nei dati? Si è provveduto a effettuare test per popolazioni specifiche o casi d'uso problematici?				
Sono state effettuate ricerche relative agli strumenti tecnici disponibili per migliorare la comprensione dei dati, del modello e delle prestazioni? Tali strumenti sono utilizzati?				
Sono stati messi in atto processi per testare e monitorare eventuali distorsioni durante le fasi di sviluppo, distribuzione e utilizzo del sistema?				
A seconda del caso d'uso, è stato previsto un meccanismo che permetta a terzi di segnalare problemi di distorsione, discriminazione o scarsa prestazione del sistema di IA?				
Sono stati presi in considerazione misure e metodi di comunicazione chiari in merito a come e a chi esporre tali problematiche?				
Sono stati presi in considerazione, oltre agli utenti (finali), anche terze parti che potrebbero essere indirettamente interessate dal sistema di IA?				
È stato valutato se esiste la possibilità che si verifichi una variabilità delle decisioni nelle stesse condizioni?				

In caso affermativo, sono state valutate le possibili cause della variabilità?				
In caso di variabilità, è stato stabilito un meccanismo di misurazione o di valutazione del potenziale impatto di tale variabilità sui diritti fondamentali?				
È stata prevista un'adeguata definizione operativa di "equità" da applicare alla progettazione di sistemi di IA?				
La definizione di equità utilizzata è di uso comune? Sono state prese in considerazione altre definizioni prima di scegliere quella utilizzata?				
Sono stati previsti un'analisi quantitativa o parametri per misurare e testare la definizione di equità utilizzata?				
Sono stati stabiliti meccanismi per garantire l'equità dei sistemi di IA? Sono stati presi in considerazione altri possibili meccanismi?				

Accessibilità e progettazione universale

Domanda	Sì	No	N/A	Note
È stato accertato che il sistema di IA si adegui a una gamma estesa di preferenze e abilità individuali?				
È stato valutato se il sistema di IA è utilizzabile da persone con bisogni speciali, disabilità o a rischio di esclusione? In che modo questo aspetto è stato inserito nella progettazione del sistema e come viene verificato?				

È stato accertato che le informazioni sul sistema di IA siano accessibili anche agli utenti di tecnologie assistive?				
Questa categoria di persone è stata consultata durante la fase di sviluppo del sistema di IA?				
Si è tenuto conto degli effetti del sistema di IA sugli utenti potenziali?				
Il team coinvolto nella realizzazione del sistema di IA è rappresentativo degli utenti destinatari? È rappresentativo della popolazione in senso lato, considerando anche altri gruppi che potrebbero essere marginalmente interessati?				
È stato valutato se vi sono persone o gruppi che potrebbero essere interessati in maniera sproporzionata da conseguenze negative?				
Sono pervenuti riscontri da altri team o gruppi con esperienze e background diversi?				

Partecipazione dei portatori di interessi

Domanda	Sì	No	N/A	Note
È stato preso in considerazione un meccanismo per far partecipare allo sviluppo e all'utilizzo del sistema di IA diversi portatori di interessi?				
L'introduzione del sistema di IA nell'organizzazione è stata preparata informando e coinvolgendo in anticipo i lavoratori interessati e i loro rappresentanti?				

6. Benessere sociale e ambientale

IA sostenibile e rispettosa dell'ambiente

Domanda	Sì	No	N/A	Note
Sono stati messi in atto meccanismi per misurare l'impatto ambientale dello sviluppo, della distribuzione e dell'utilizzo del sistema di IA (ad esempio, la quantità e il tipo di energia utilizzata dal centro dati, ecc.)?				
Sono state previste misure per ridurre l'impatto ambientale del ciclo di vita del sistema di IA?				

Impatto sociale

Domanda	Sì	No	N/A	Note
È stato valutato se il sistema di IA incoraggia gli esseri umani a sviluppare attaccamento ed empatia verso il sistema?				
È stato accertato che il sistema di IA segnali chiaramente che la sua interazione sociale è simulata e che non ha capacità né di "capire" né di "provare sentimenti"?				

È stato accertato che l'impatto sociale del sistema di IA sia ben compreso? Ad esempio, è stato valutato se esiste un rischio di perdita di posti di lavoro o di dequalificazione della forza lavoro? Quali misure sono state adottate per contrastare tali rischi?				
---	--	--	--	--

Società e democrazia

Domanda	Sì	No	N/A	Note
È stato valutato l'impatto più generale sulla società dell'uso del sistema di IA al di là del singolo utente (finale), come ad esempio i portatori di interessi indirettamente interessati?				

7. Accountability

Verificabilità

Domanda	Sì	No	N/A	Note
Sono stati messi in atto meccanismi che facilitano la verificabilità del sistema da parte di soggetti interni e/o indipendenti, come ad esempio la tracciabilità e la registrazione dei processi e dei risultati del sistema di IA?				

Riduzione al minimo degli effetti negativi e loro segnalazione

Domanda	Sì	No	N/A	Note

È stata effettuata una valutazione del rischio o dell'impatto del sistema di IA che tenga conto dei diversi portatori di interessi direttamente o indirettamente interessati?				
Sono stati messi in atto sistemi di formazione e istruzione per sviluppare pratiche di accountability?				
Quali lavoratori o settori del team sono coinvolti nelle pratiche di accountability? Tale processo va oltre la fase di sviluppo?				
Nei corsi di formazione si insegna anche il possibile quadro giuridico applicabile al sistema di IA?				
È stata presa in considerazione l'istituzione di un "comitato di revisione etica dell'IA" o un meccanismo analogo per discutere di accountability e pratiche etiche in generale, comprese le "zone grigie" potenzialmente poco chiare?				
Oltre alle iniziative interne o ai quadri per la supervisione etica e dell'accountability, esiste qualche tipo di orientamento esterno o sono stati istituiti anche processi di verifica?				
Sono stati messi in atto processi che permettano a terzi (ad esempio fornitori, consumatori, distributori/venditori) o ai lavoratori di segnalare potenziali vulnerabilità, rischi o distorsioni nel sistema/applicazione dell'IA?				

Documentazione dei compromessi

Domanda	Sì	No	N/A	Note , rif.
---------	----	----	-----	----------------

È stato stabilito un meccanismo per individuare gli interessi e i valori pertinenti chiamati in causa dal sistema di IA e i potenziali compromessi tra di essi?				
Con quale processo si adottano decisioni in merito a tali compromessi? È stata garantita la documentazione della decisione adottata in merito ai compromessi?				

Capacità di ricorso

Domanda	Sì	No	N/A	Note
È stata stabilita una serie adeguata di meccanismi di ricorso in caso di danni o effetti negativi?				
Sono stati messi in atto meccanismi per fornire informazioni sia agli utenti (finali) sia a terzi sulle possibilità di ricorso?				

Per ogni domanda è possibile contrassegnare una sola delle tre opzioni (Sì, No, N/A) e utilizzare la colonna "Note, riferimento, documento" per indicare evidenze, riferimenti normativi, documenti di supporto o osservazioni specifiche relative all'implementazione.

Il Gruppo di lavoro suggerisce di adottare la lista “per la valutazione dell'affidabilità dell'IA nelle fasi di sviluppo, distribuzione o utilizzo dell'IA” e di adattarla allo specifico caso d'uso del sistema.

E' bene tenere presente che la lista di controllo non può essere esaustiva: bisogna “garantire che un'IA affidabile non sia una questione di caselle da spuntare, ma un processo continuo di individuazione dei requisiti, valutazione delle soluzioni e miglioramento dei risultati durante l'intero ciclo di vita del sistema di IA e di coinvolgimento dei portatori di interessi in tale processo”.

Legenda:

- Sì (Applied): il requisito è stato implementato e sono disponibili evidenze documentali
- No (Non eseguito): il requisito non è stato ancora implementato
- N/A (Not Applicable): il requisito non è applicabile al caso d'uso specifico del sistema di IA