T

Elio Grande

Il loop espanso. Per un ecosistema dell'intelligenza artificiale

1. *Introduzione*

Quest'articolo desidera, sotto forma di idea, piantare il seme di un possibile ecosistema dell'intelligenza artificiale – o per meglio dire, un ecosistema abitato *anche* da sistemi artificiali intelligenti – a partire però da una concreta prova di esistenza: il metodo di design "a specchio" di sistemi di intelligenza artificiale collaborativi battezzato *Endless Tuning* (Fabris *et al.*, 2024), già implementato in un protocollo e testato con alcuni esperti di dominio (Grande, 2025)[1]. Un tale ecosistema, di radice ermeneutica, ne costituisce dunque per un verso l'espansione realizzabile; per un altro, a testimonianza di quanto tangibilmente aporetico sia l'inizio (Fabris, 2002), ne costituisce già la parziale realizzazione: già, infatti, esistevano i dataset impiegati per gli esperimenti, *già* i moduli di programmazione, e così via lungo l'abbozzo di un circolo.

Alla strutturazione di questo ecosistema dovrebbero a dire il vero contribuire, tra loro cooperando, competenze di varia natura. Agli specialisti del task da eseguire (medico, operatore bancario, consulente etc.), dovrebbero essere affiancati lo sviluppatore e il manutentore di sistemi di intelligenza artificiale (magari, tra l'altro, la stessa persona). Al manager, invece, il compito

[1] È possibile testare il protocollo online in tre versioni, rispettivamente riconoscimento dello stile di un'opera d'arte, concessione di prestiti bancari e riconoscimento della presenza o meno di polmonite in una radiografia toracica. I codici sono disponibili su GitHub alla pagina https://github.com/eliogrande/TheEndlessTuning, e possono essere eseguiti tramite un *Jupyter Notebook* disponibile su *Google Colab* alla pagina https://colab.research.google.com/drive/1m1mDWTlE5egzT_oSR18hHLlcwwrX401N?authuser=1#scrollTo=3pMgbEeXeB5X. Si raccomanda, per la riuscita dell'esecuzione, l'opzione di *runtime* che fa uso di processore grafico.

 DOI: 10.4454/tqnjph66

di preparare il campo organizzando un'infrastruttura, secondo però le indicazioni del giurista che dovrebbe perlomeno accertarsi della compliance con le prescrizioni del Regolamento Europeo sull'Intelligenza Artificiale (European Parliament *et al.*, 2024), del *General Data Protection Regulation* (European Union, 2016) etc. *Dulcis in fundo*, l'eticista: un ruolo apparentemente marginale, a prediligere l'approccio deontologico; eppure, un tessitore degno dell'Eros del mito. Mano ad ago e filo, dunque: dell'infrastruttura summenzionata, a mo' di trama, l'etica dell'intelligenza artificiale costituisce inevitabilmente l'ordito – forse, certo, per varie ragioni invisibile – giacché ne *orienta* il disegno. Dapprima in piccolo, nella relazione tra lo strumento/agente e l'operatore; poi, espandendosi, fungendo da baricentro del contesto sociale di quella medesima relazione. Secondo quest'ordine, perciò, procederemo: l'etica nell'*interface design*, l'esperienza dell'applicazione concreta di *Endless Tuning*, la descrizione dei suoi possibili risvolti sociali.

2. *Il modo in cui una relazione si manifesta:* interface design

Non difettano i codici etici cui riferirsi, a partire da prodotti istituzionali come le *Ethics Guidelines for Trustworthy AI* (European Commission *et al.*, 2019). Anzitutto, però, c'è la difficoltà di trascrivere efficacemente principi morali in uno script (Coeckelbergh, 2019) perché, un modello non potendo riconoscere da sé valori adattandosi ad ogni situazione, rimangono facilmente aggirabili. Per esempio, l'imposizione di *constraints*, come il LLM che non rivela come mettere in atto un attacco *DoS*, non copre il rischio di *jailbreaking*. Poi, il carattere di "*very soft law*" non conferisce all'etica dell'intelligenza artificiale forza vincolante. Facilmente la si può ridurre a mero commentario (Floridi, 2023; Bietti, 2020). Fece scalpore il disappunto di Mark Coeckelbergh e Thomas Metzinger quando, alla stesura del White Paper da parte della Commissione Europea (European Commission, 2020), molte delle raccomandazioni etiche elaborate dal HLEG erano state ignorate o ridotte a semplici dichiarazioni di principio[2].

Quando "cosa" e "prodotto" coincidono, non diviene invece il filosofo un designer, la cui arte risieda però *nell'interrogare il prodotto stesso*; nel questionare, *mentre la disegna*, l'essenza di quella "cosa"? Secondo Floridi (2017),

[2] L'articolo, uscito sul *Tagesspiel* il 14/04/2020, è consultabile in inglese alla pagina https://background.tagesspiegel.de/digitalisierung-und-ki/briefing/europe-needs-more-guts-when-it-comes-to-ai-ethics? (ultimo accesso 11/10/2025).

condivisibilmente, nell'infosfera l'etica che salvaguardi la libertà dell'agire diviene dunque un design pro-etico: con una similitudine, non un guard-rail entro cui limitare il proprio agire bensì un incentivo all'agire consapevole. Oggetto di design, d'altra parte, è sempre un'interfaccia – concetto che insieme a quello di protocollo può essere esteso anche oltre gli stretti confini di un dispositivo (Floridi, 2017). Ma cos'è un'interfaccia? Questo termine, scrive Zannoni (2024), implica certamente comunanza, collegamento e relazione; tuttavia «è necessario riflettere sul fatto che l'interfaccia non esiste ed è solo un'entità astratta a cui ci riferiamo quando ci relazioniamo tra diversi sistemi reali o virtuali» (Zannoni, 2024: 7). Spesso all'interfaccia è associato il termine *affordance*, coniato da James Jerome Gibson (1966) col significato di indicare ciò che le cose più o meno direttamente ci invitano a fare. Val la pena, per giungere al punto, riportare le sue proprie parole: «When the constant properties of constant objects are perceived (the shape, size, color […]), the observer can go on to detect their affordances. I have coined this word as a substitute for values, a term which carries an old burden of philosophical meaning» (Gibson, 1966: 285). *Dunque, qual è la natura di tale valore, precondizione di un'interfaccia?* Una sola sembra la risposta: *l'utilizzabilità*. "Design", in altre parole, sarebbe sinonimo di "design di oggetti d'uso" costantemente "sotto-mano" (Heidegger, 2024): essi *si offrono* all'uso, venire utilizzati costituisce la loro propria essenza.

Questo, tuttavia, non è affatto ovvio: l'utilizzabilità non pare infatti l'unico modo d'essere, né l'unica condizione di possibilità di design di un'interfaccia. Più sottilmente, con (Fabris, 2016; Buber, 1970), sosteniamo che l'utilizzabilità, producendo oggetti e "cose", costituisca un'implicita e concreta scelta, l'attuazione di una tra varie possibilità che *sono, costituiscono* una relazione, così facendo già azzoppando tale relazione, poiché se ne occlude l'orizzonte di possibilità. Non a caso, in materia di intelligenza artificiale proprio a questo punto il termine *reliability* (che indica la robustezza intesa come resistenza a perturbazioni in ingresso, eppure presenta, sempre non per caso, anche una sfumatura di significato esistenziale derivata dal latino *religamen*, da cui *religio*) inizia a far spazio a quello di *trustworthiness*. Da accordo che coinvolge ad accordo che delega – salvo poi cercare di rimediare, contraddittoriamente, chiedendo l'esplicabilità di algoritmi ormai incredibilmente performanti.

Facendo ricerca all'interno dello Spoke 1.6 del grande progetto nazionale italiano FAIR, intitolato *Legal and Ethical Design of Trustworthy AI Systems*, abbiamo dunque scelto come principio guida di design quello della relazione stessa (*anche*) con lo strumento. Tale relazione è intesa come

il continuo coinvolgimento nello sforzo di estrarre informazione rilevante (aggettivo qui migliore che "utile"). Il principio di fondo è un principio-contenitore se vogliamo, o meta-principio, desunto dall'etica della vulnerabilità e della cura. Questo trasforma il design in un problema di *governo* dell'intelligenza artificiale, più che di uso o controllo. Sotto questo profilo, un'interfaccia è dunque intesa come *il modo in cui una relazione si manifesta* – comportando però ossimoricamente che si dia un margine – passi il termine – di s-progetto, di non-design. Opzione che pare invero giustificata: altrove (Fabris *et al.*, 2024; Grande, 2025) abbiamo riportato ragionevoli indizi che i dispositivi appartenenti alla famiglia dell'intelligenza artificiale siano in senso lato detentori di un margine di *inutilizzabilità*. Per esempio, inconsulte ed "innaturali", poiché *assai precisamente errate*, sono le reazioni pressoché di un modello alla presenza di *adversarial examples*, cioè dati in ingresso affetti da perturbazioni minimali mirate a ribaltare il risultato di un'inferenza (scoperti in Szegedy *et al.*, 2013); per un *survey*, cfr. Wiyatno *et al.*, 2019). Oppure, la ragione che vincola l'ingresso all'uscita di un modello nei cicli di training ed inferenza di modelli opachi come *Deep Neural Networks*, *Support Vector Machines* e *Random Forests* rimane parzialmente oscura anche dinanzi a modelli trasparenti, di cui cioè sono noti i parametri. O ancora, il proliferare di modelli ad elevata complessità, relativamente a numero di parametri o FLOPs necessari all'addestramento ha reso noto l'emergere di alcune abilità (Wei *et al.*, 2022) come il *few-shot prompting* che, sebbene da taluni negate (Schaeffer *et al.*, 2023), destano tuttora meraviglia.

3. *Proof of concept*

Endless Tuning è un metodo di design (destinato originariamente ai processi decisionali, sebbene confidiamo possa essere esteso anche all'intelligenza artificiale generativa) che risponde a due bisogni: da una parte quello di compensare le tentazioni del cosiddetto *automation bias* (Skitka *et al.*, 1999), ovvero la tendenza a delegare, con fiducia cieca, decisioni alle macchine; dall'altra quello di tracciare le responsabilità in caso di danno, fornendo la possibilità di dirimere la questione piuttosto che cercando a priori il "colpevole". Per far ciò, si adoperano due regole. La prima è che, trovandosi l'uno allo specchio dell'altro, il sistema di intelligenza artificiale e l'operatore dovrebbero riuscire ad imparare l'uno dall'altro e ciascuno a suo modo, ovvero l'operatore riflettendo a partire dai suggerimenti del sistema ed il sistema essendo soggetto a vincoli di apprendimento per i casi futuri.

Sintonizzandosi l'uno e l'altro sulla verità del fenomeno in osservazione, potenzialmente senza una fine. La seconda è che qualunque interazione con il sistema e qualunque dato prodotto in maniera non facilmente riproducibile siano salvati in una memoria, per poi essere eventualmente consultati in sede di giudizio. Forse che *Endless Tuning* sia perciò incentrato sulla responsabilità civile piuttosto che quella morale? In verità, fermo restando che non esiste un "design del senso di colpa", scopo precipuo di questo progetto è facilitare delle relazioni. In quest'ottica, sebbene all'operatore sia destinato un ruolo rilevante e gli sia richiesta proprio un'assunzione di responsabilità nel prendere una decisione, tale responsabilità non è pienamente configurabile a priori, bensì si costruisce durante il flusso decisionale stesso, e si ri-costruisce a posteriori. Questo, coerentemente sia con un approccio all'etica della cura e della vulnerabilità – una duplice *different voice* (Gilligan, 1993): quella di chi sarà eventualmente soggetto alle decisioni e quella dell'intelligenza artificiale stessa – sia con il criterio di responsabilità oggettiva che pare oggi il più adeguato a dirimere tali questioni (dal Verme, 2024).

Ne abbiamo istanziato un protocollo applicativo che prevede, in ordine: una prima opinione da parte dell'operatore; alcuni suggerimenti sull'informazione processata *e dunque indirettamente* relativi allo stato di cose sotto osservazione, senza che l'operatore conosca ancora il risultato della predizione (utilizzo inverso, *ante-hoc*, di *eXplainable AI*, non lontano dal paradigma di *evaluative AI* (Miller, 2023); confronti di similarità con altri casi); rilevazione dell'effettiva predizione, modulata però sulla *confidence* (misura di certezza) di tutti i risultati possibili; riaddestramento correttivo del modello, previo lo stoccaggio di alcuni dati (*fine-tuning set*) in un "salvadanaio". Ad ogni stadio l'opinione dell'operatore può essere aggiornata, ogni interazione è tracciata in un log con un *timestamp*, ed elementi come mappe e parametri inizializzati in modo randomico vengono salvati. Questo protocollo è stato a sua volta realizzato in tre prototipi, ovvero tre analoghe applicazioni *Streamlit*[3] (basate però su modelli di intelligenza artificiale differenti)[4] dedicate rispettivamente al riconoscimento dello stile di un'opera d'arte, al riconoscimento della presenza o meno di polmonite in una radiografia toracica, ed

[3] Cfr. http://streamlit.io/.

[4] Per i dettagli tecnici, cfr. Grande (2025). Solo per completezza d'informazione, comunque, specifichiamo qui che si tratta di due *ResNet* (He *et al.*, 2016), rispettivamente a 34 e 18 strati, la prima addestrata su alcune classi del *WikiArt Dataset* (Steubk, 2022) e la seconda sul *Chest X-Ray Images (Pneumonia) Dataset* (Mooney, 2018); il terzo modello è un albero decisionale addestrato sul *Loan Approval Classification Dataset* (Wei Lo, 2024).

alla valutazione di candidati per la concessione o meno di prestiti bancari. Da quest'ultimo modello, per esempio, dinanzi alle caratteristiche del candidato (come età, reddito etc.), venivano direttamente estratte delle regole in linguaggio naturale che facevano evincere come alcune *features* si distinguessero da certi valori standard (non sempre uguali). Questo (come analoghi passaggi negli altri prototipi) aveva lo scopo di compensare possibili bias di ancoraggio (cfr. Kahneman, 2011) da parte del decisore, forzando quest'ultimo a rivalutare il caso con maggiore attenzione (*cognitive forcing* è infatti il nome del genere di strategia adottata: cfr. Buçinca *et al.*, 2021).

I prototipi sono stati testati attraverso delle interviste ad altrettanti esperti di dominio (un direttore di banca, un docente di estetica ed un primario di radiologia), con risultati complessivamente positivi. Sebbene infatti alcuni difetti siano stati concretamente riscontrati, per esempio riguardo l'espressività di alcune mappe di salienza dei pixel, gli intervistati hanno dichiarato di non essersi sentiti in alcun modo sostituiti dal sistema di intelligenza artificiale[5] – il quale ha prodotto in tutti e tre i casi predizioni corrette. Inoltre, eccetto il medico, favorevole piuttosto all'opzione di rigide linee guida per stabilire a priori le responsabilità in caso di danno, gli esperti hanno condiviso l'impostazione di tracciamento delle attività. *Endless Tuning* rientra dunque nella categoria dell'*hybrid decision-making*, un paradigma di collaborazione tra essere umano ed intelligenza artificiale che sempre più si fa avanti (Punzi *et al.*, 2023), nel tentativo di intercettare il miglior compromesso tra l'accuratezza ed i diritti fondamentali della persona, la privacy ed il benessere collettivo, l'ergonomia e la sostenibilità.

4. *Il loop esteso*

Se da un lato *Endless Tuning* è un design collaborativo, dall'altro la sua doppia ciclicità è a tutti gli effetti un'ermeneutica, dove a turno e tuttavia quasi simultaneamente l'operatore ed il sistema giocano a vicenda il ruolo di interprete ed interpretato. Un'ermeneutica dell'intelligenza artificiale coerente con la più generale versione digitale descritta da (Romele, 2019), che proponeva il concetto di e*magination* – un analogo dell'immaginazione

[5] Preferiamo qui adoperare la parola "sistema", piuttosto che "modello", in quanto il modello stesso è avvolto in un'interfaccia complessa, che include un algoritmo di spiegazione ed un compressore di dati atto a misurare la distanza geometrica tra il caso sotto osservazione ed altri presi dal dataset di apprendimento.

produttiva ravvisabile nella capacità, da parte di un algoritmo, di analizzare grandi moli di dati. Quel *surplus* di informazione (latore di senso) estraibile dalla e nell'interazione, bilanciando (pure in modo asimmetrico) il potere di iniziativa degli agenti in campo, allontana il baricentro sia da un approccio *human-in-the-loop* che da uno *machine-in-the-loop*: il loop stesso ne diviene, piuttosto, il centro. Dal suo nucleo esso *inevitabilmente* si espande, coinvolgendo, come giustamente notato da (Romele, 2019), altre figure. Per questo un ecosistema si profila, dove i loop divengono concentrici. Sul lato sinistro della Figura 1, il cui loop interno è lo schema base di *Endless Tuning*, dietro il sistema di intelligenza artificiale si trovano i dati di addestramento e il complesso di figure come il progettista dell'interfaccia, lo sviluppatore del modello e del software, e l'annotatore dei dati. Sul lato destro, invece, ci sono gli operatori: l'individuo, per esempio, può ben essere colui che prende effettivamente la decisione; gli altri, anch'essi operatori, intervengono successivamente insieme al primo come annotatori di nuovi dati da fornire al modello stesso durante la fase di aggiornamento.

Figura 1. *Endless Tuning* considerato come un sistema socio-tecnologico esteso. L'icona della matita è stata presa da Flaticon[6].

Se i dataset annotati sul portale *Kaggle*, le architetture disponibili nei framework *PyTorch* o *Scikit-Learn*, una discreta potenza computazionale (solo in un caso siamo dovuti ricorrere ad un server), degli esperti e soprattutto l'impulso alla condivisione (l'accesso a quasi tutte le risorse, infatti, è stato gratuito) non fossero stati già una realtà, questo *proof of concept* non sarebbe stato sperimentabile. Questo doppio circolo ermeneutico digitale, parzialmente, esiste già: in principio è davvero la relazione. Se dunque per

[6] https://www.flaticon.com/free-icons/pencil.

un verso si tratta di espanderla, per un altro si tratta semplicemente di riconoscerla, mettendo insieme i pezzi. Se da un lato le regole del metodo rimangono le medesime, applicate su una scala più grande (per esempio, dal *fine-tuning* alla sostituzione del modello medesimo; dal singolo decisore all'equipe; dalla cura dell'individuo soggetto alla decisione a quella della comunità di appartenenza), dall'altro lato il loop espanso dev'essere cucito sulle esigenze di un territorio, dunque in un contesto applicativo mutevole. Difficile, quindi, fornire un protocollo. Piuttosto, un esempio può far più di mille parole. Può valere pertanto la pena di lasciar emergere i caratteri, le funzionalità, le conseguenze etiche e giuridiche di un tale ecosistema descrivendone icasticamente un'istanza fittizia – immaginata, per altro, a partire da uno dei nostri tre casi di studio.

5. *Un esempio di ecosistema*

Un'ondata di freddo travolge l'area periferica industriale di una città, cogliendo impreparato il signor Mario Rossi, un uomo sulla cinquantina i cui polmoni concedono presto spazio ad un'infiammazione. Timoroso che la situazione possa peggiorare, Mario decide di andare a fare una radiografia presso un centro diagnostico vicino alla sua abitazione. Si posiziona dunque dinanzi al macchinario radiogeno, che coglie l'istantanea dell'interno del suo torace.

Il radiologo ha a disposizione un modello di intelligenza artificiale – una rete neurale convoluzionale – con cui, al momento di valutare la diagnosi sulla presenza o meno e sulle caratteristiche di una polmonite, avvia un dialogo di confronto. Mentre, infatti, quest'ultimo fa la sua ipotesi, l'interfaccia del software gli propone via via, inducendolo a riflettere, suggerimenti ed indizi fino a presentargli esplicitamente il risultato della propria predizione. Una certa fiducia, dunque, è concessa all'operatore, il quale è tuttavia costretto a prendere una decisione oppure a rimandare, con le dovute annotazioni, per fare ulteriori accertamenti, mantenendo un ampio margine di libertà di decisione. Si tratta di un software – il modello, un algoritmo di spiegazione, un sistema di gestione dei dati e l'interfaccia utente che avvolge il tutto – che l'Azienda Sanitaria Provinciale ha acquistato da un'azienda francese, conforme ai requisiti dell'*AI Act*. Un caso medico, infatti, è considerato in vetta alla piramide del rischio. Nondimeno, gli sviluppatori hanno potuto godere di un vantaggio produttivo, a partire dal fatto che l'iniziativa decisionale rimane nelle mani dell'operatore. Infatti, secondo il

recital 53 dell'*AI Act*, esistono condizioni che potrebbero mitigare il rischio direttamente relativo al sistema. Tra queste, quella che il task eseguito dal sistema di intelligenza artificiale sia inteso a migliorare il risultato di una precedente attività umana destinata al medesimo compito.

Il contratto con lo sviluppatore prevede che il modello, pre-addestrato su dati medici coerenti con il task ma ancora un po' troppo generici, venga trasferito su un server centrale di proprietà dell'ASP, ed il software ivi eseguito. Alla medesima ASP, sempre da contratto, il compito di assumere un manutentore delle procedure di *fine-tuning* e di gestione dei dati (Queste, tuttavia, sono scelte meramente dovute al fatto che l'azienda sia straniera. La volontà della dirigenza, infatti, è di circoscrivere il più possibile l'uso dei dati al territorio). Un'iniziativa, invece, da parte della dirigenza dell'ASP, è consistita nell'assumere un gruppo di eticisti dell'intelligenza artificiale allo scopo di formare gli operatori, che purtroppo (nel nostro esempio fittizio) difettano di *AI literacy*.

Fatta la diagnosi, la radiografia del signor Rossi viene anonimizzata. Se ne eliminano alcuni metadati, tra cui nome e cognome, e la diagnosi stessa insieme all'immagine (il cui annotatore è dunque il radiologo) vengono inserite in un dataset sanitario locale, di dimensioni in costante incremento, cui gli omologhi sistemi diagnostici delle strutture sanitarie vicine (nonché del medesimo centro cui Mario si è rivolto) attingeranno, tra qualche mese, per aggiornare le proprie basi di conoscenza relative alla salute sul territorio. In altre parole, ergonomizzandosi. Ogni giorno, infatti, ciascuno dei centri diagnostici della provincia esegue circa venti radiografie, per un totale di circa duecento immagini giornaliere. Dopo qualche mese, il "salvadanaio" grezzo di dati è pronto. A dire il vero, anche nel caso della radiografia al torace di Mario Rossi, proprio in virtù della peculiarità delle condizioni di salute medie in quella zona industriale, alcuni aggiornamenti erano già stati effettuati rispetto al modello pre-addestrato dall'azienda straniera sviluppatrice del software. L'aggiornamento parziale del modello (che in certo modo pertiene al paradigma del *Continual Learning* (Lesort *et al.*, 2020)) ha una duplice finalità: da un lato, l'adeguamento a possibili variazioni nelle distribuzioni locali di dati, dovute a caratteri genotipici o fattori ambientali; dall'altro, l'autocorrezione secondo le indicazioni degli esperti.

L'anonimizzazione ha lo scopo di alleggerire le difficoltà nel trattamento dei dati personali. Unicamente a Mario Rossi spetta, infatti, la scelta se condividerli o meno. Tuttavia, conoscendo e confidando nel personale sanitario, decide di accettare. Il fatto che per essere processati i dati non debbano migrare su un server remoto, magari oltreoceano, comporta da un lato una ras-

sicurazione ulteriore sotto il profilo della privacy, dall'altro potrebbe essere d'ausilio nel limitare gli effetti di azioni malevole, come il *data poisoning*. Sembra inoltre soddisfatto il "diritto alla spiegazione" stabilito dal GDPR al recital 71 ed all'articolo 22. Per ciò che infatti concerne l'operatore, per così dire, nulla di nuovo. Per quel che invece può concernere il modello, che non ha propriamente voce in capitolo, l'utilizzo di tecniche di *eXplainable AI* può essere più che sufficiente nel far conoscere gli indizi che possono aver spinto il decisore umano. Un limite è invece la disponibilità di dati. Qualora fossero più rari, per esempio a causa di un minor numero di pazienti, si aprirebbero tre possibilità. La prima è di acquistarli, la seconda è di generarli (sotto supervisione), la terza è richiederne ad altri enti la condivisione. Pratica, quest'ultima, auspicata nel *Data Act* europeo (European Union, 2023), su cui per altro si basa anche l'interoperabilità tra centri diagnostici all'interno dell'ASP stessa. Un board di supervisori (i direttori dei reparti di radiologia degli ospedali di zona insieme ai docenti competenti in materia dell'Università della provincia) periodicamente rivede le etichettature dei dati e dà il via libera ad un nuovo aggiornamento dei parametri del dispositivo. Aggiornamento che, date la ristrettezza del dominio di applicazione e le garanzie fornite dalla (necessaria e cruciale) competenza dell'operatore, non richiede ingenti risorse computazionali.

Salvo che per informazioni riservate e per l'inevitabile opacità dei modelli, l'intero ciclo si svolge in trasparenza e i processi decisionali, insieme alle operazioni di manutenzione, vengono registrati in un libro mastro che consenta di tenerne traccia. Pertanto, laddove qualcosa andasse storto e Mario Rossi ne avesse un domani a lamentarsi dinanzi alla giustizia, sarà possibile rivedere l'intero processo decisionale al rallentatore e dirimere la questione di chi e come debba risponderne. Da una parte, infatti, è implementata una forma di *accountability* computazionale (Comandé, 2018); dall'altra diverrà possibile, per ciascun soggetto coinvolto, dimostrare se e come si è agito per evitare il peggio o, con un'espressione giuridica, invertire l'onere della prova. Una risposta al cosiddetto *problem of many hands* (Nissenbaum, 1996), dovuto alla complessità della *supply chain*. Una grande fetta di responsabilità sarà senz'altro detenuta dal radiologo cui si è sottoposto Mario Rossi, ma è possibile che i suggerimenti forniti dal modello fossero ingannevoli, magari per una leggerezza del *UI designer*; oppure, ancora, il manutentore del sistema non aveva bloccato il riaddestramento quando erano emersi indizi di *overfitting*, e così via.

6. *Conclusione*

Questo è, naturalmente, soltanto uno schizzo di un ecosistema dell'intelligenza artificiale orientato all'etica delle relazioni. Ogni configurazione concreta va attentamente ponderata in base al contesto e alle competenze in gioco. Vi sono certamente dei rischi: oltre alla summenzionata necessità di ottenere dati in quantità sufficiente, c'è la stretta dipendenza dal tipo di task. Un compito, infatti, che (come per esempio quello linguistico) richiede grandi modelli, diventa difficile da portare a termine su piccole realtà – almeno, fin quando le speranze del *Continual Learning* non saranno realizzate. V'è poi il rischio che un circolo virtuoso diventi vizioso e porti ad un costante peggioramento di prestazioni. Non troviamo, tuttavia, altra risposta a quest'ultima possibile obiezione se non che affrontare questo rischio con impegno *è* l'etica dell'intelligenza artificiale. Questo piccolo esperimento di immaginazione mostra come il loop di *Endless Tuning*, considerato come piccolo sistema socio-tecnologico, possa espandersi implicando la partecipazione attiva di tutti i soggetti coinvolti. L'etica diventa qui principio operativo, traducendo la vulnerabilità in responsabilità condivisa e rendendosi finanche capace di mediare tra culture diverse. Ogni regione del mondo porta con sé la sua *fairness*: riaddestrare (o raffinare) significa, implicitamente, anche trovare un compromesso. La vulnerabilità appartiene a tutti i soggetti in gioco: vulnerabile, nel nostro esempio, è il paziente come il radiologo, come per altro l'intero complesso se, in seguito ad epidemie che fanno saltare le distribuzioni di dati, diviene necessario ripartire da zero. Questa, però, è una malizia della natura. Il nostro compito è quello di gestire il rischio, esistenziale e in ciò sempre incalcolabile, all'interno di una cornice temporale sempre limitata (poiché ogni modello, implementando una funzione matematica, assume la stabilità della realtà che descrive). In tal senso, progettare l'intelligenza artificiale significa co-costruire spazi di decisione consapevole. Espandere il loop, infine, equivale a espandere la responsabilità stessa del pensare e dell'agire.

Riferimenti bibliografici

Bietti, E. (2020). *From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy*, in «Proceedings of the 2020 conference on fairness, accountability, and transparency», pp. 210-219.

Buber, M. (1970). *I and thou. A new translation with a prologue "I and you" and notes by W. Kaufmann*, Charles Scribner's Sons, New York.

Buçinca, Z., Malaya, M.B., Gajos, K.Z. (2021). *To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making*, in «Proceedings of the ACM on Human-computer Interaction», 5 (CSCW1), pp. 1-21.

Coeckelbergh, M. (2019). *Artificial intelligence: Some ethical issues and regulatory challenges*, in «Technology and regulation», pp. 31-34.

Comandé, G. (2018). *Responsabilità ed accountability nell'era dell'Intelligenza Artificiale*, in «Giurisprudenza e Autorità Indipendenti nell'epoca del diritto liquido», La Tribuna SRL, pp. 1001-1013.

dal Verme, T.D.M.C. (2024). *Intelligenza artificiale e responsabilità civile: uno studio sui criteri di imputazione*, Editoriale Scientifica, Napoli.

European Commission (2020). *White Paper on Artificial Intelligence - A European Approach to Excellence and Trust*, COM (2020) 65 final, Brussels.

European Commission (2019). *Directorate-General for Communications Networks, Content and Technology & High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*, Publications Office, Luxembourg.

European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828*, in «Official Journal of the European Union», L 219, 12 July 2024.

European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, in «Official Journal of the European Union», L 119, pp. 1-88.

European Union (2023). *Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data (Data Act). Official Journal of the European Union*, L 2023/2854.

Fabris, A. (2002). *Paradossi del senso*, Morcelliana, Brescia.

Fabris, A. (2016). *RelAzione. Una filosofia performativa*, Morcelliana, Brescia.

Fabris, A., Dadà, S., Grande, E. (2024). *Towards a Relational Ethics in AI. The Problem of Agency, The Search for Common Principles, the Pairing of Human and Artificial Agents,* in A. Fabris, S. Belardinelli, *Digital Environments and Human Relations: Ethical Perspectives on AI Issues*, Springer Nature Switzerland, Cham, pp. 9-42.

Floridi, L. (2017). *La quarta rivoluzione: come l'infosfera sta trasformando il mondo*, Raffaello Cortina Editore, Milano.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford.

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems* (revised ed.), George Allen & Unwin, London.

Grande, E. (2025). *The Endless Tuning. An Artificial Intelligence Design To Avoid Human Replacement and Trace Back Responsibilities*, arXiv preprint arXiv:2507.14909.

He, K., Zhang, X., Ren, S., Sun, J. (2016). *Deep Residual Learning for Image Recognition*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Heidegger, M. (2024). *Essere e tempo*, vol. 23, Minerva Heritage Press, Roma.

Kahneman, D. (2011). *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., Díaz-Rodríguez, N. (2020). *Continual Learning for Robotics: Definition, Framework, Learning Strategies, Opportunities and Challenges*, in *Information Fusion*, Elsevier, Amsterdam, pp. 52-68.

Miller, T. (2023). *Explainable AI Is Dead, Long Live Explainable AI!*, arXiv:2302.12389v3 [cs.AI].

Mooney, P. (2018). *Chest X-Ray Images (Pneumonia)*, https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, CC BY 4.0 License, accessed 16 June 2025.

Nissenbaum, H. (1996). *Accountability in a Computerized Society*, in «Science and Engineering Ethics», 2 (1), pp. 25-42.

Punzi, C., Setzu, M., Pellungrini, R., Giannotti, F., Pedreschi, D. (2023). *Towards Synergistic Human-AI Collaboration in Hybrid Decision-Making Systems*, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=3UL0kVUAAAAJ&sortby=pubdate&citation_for_view=3UL0kVUAAAAJ:hMod-77fHWUC.

Romele, A. (2019). *Digital Hermeneutics: Philosophical Investigations in New Media and Technologies*, Routledge, London.

Schaeffer, R., Miranda, B., Koyejo, S. (2023). *Are Emergent Abilities of Large Language Models a Mirage?*, in *Advances in Neural Information Processing Systems* 36, pp. 55565-55581.

Skitka, L.S., Mosier, K.L., Burdick, M. (1999). *Does Automation Bias Decision Making?*, in «International Journal of Human-Computer Studies» 51, pp. 991-1006.

Steubk (2002). *WikiArt. A Collection of WikiArt Images* (www.wikiart.org), https://www.kaggle.com/datasets/steubk/wikiart, CC0 Public Domain License, accessed 16 June 2025.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J. , Erhan, D., Goodfellow, I., Fergus, R. (2013). *Intriguing Properties of Neural Networks*, arXiv:1312.6199v4 [cs.CV].

Wei Lo, T. (2024). *Loan Approval Classification Dataset*, https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data?resource=download, Apache License 2.0, accessed 8 June 2025.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W. (2022). *Emergent Abilities of Large Language Models*, in *Transactions on Machine Learning Research*ì, arXiv:2206.07682v2 [cs.CL].

Wiyatno, R.R., Xu, A., Dia, O., De Berker, A. (2019). *Adversarial Examples in Modern Machine Learning: A Review*, arXiv preprint arXiv:1911.05268.

Zannoni, M. (2024). *Il design delle interfacce*, Quodlibet, Macerata.

Engish title: The expanded loop. For an ecosystem of artificial intelligence

Abstract

This article aims to plant the seed of a possible ecosystem of artificial intelligence, rooted in hermeneutics, beginning from a concrete proof of existence: the ethical design method Endless Tuning *and its implemented prototypes. The ethics of artificial intelligence, by recognizing the importance of vulnerability, becomes the key to the relationships within an infrastructure that includes various figures – manager, developer, ethicist, domain expert, etc. – thus orienting its design firstly on a small scale, in the relationship between the tool/agent and the operator; then expanding, acting as the center of gravity of the social context of that same relationship. According to this order, therefore, we shall proceed: ethics in interface design, the experience of the concrete application of* Endless Tuning*, and the description of its possible social implications. The latter will present a fictional yet realistic example – a case of medical image diagnostics – as both a possible extension and a recognition of a circular ecosystem that already partially exists.*

Keywords: ethics of artificial intelligence; hermeneutics; interface design; ecosystem; relation.

Elio Grande
University of Pisa
elio.grande@phd.unipi.it