Stato, Chiese
e pluralismo confessionale

Graziano Mioli
(avvocato della Rota Romana)

Algor-etica e *moral-algo* in difesa dell'umano nell'era delle IA *

Algor-ethics *and* moral-algo *in defense of humanity in the age of AI* *

ABSTRACT: This essay proceeds from the awareness - expressed by the Catholic Magisterium and shared by Digital Humanism - that humanity stands at an epochal crossroads before the advance of artificial intelligence. After surveying the contemporary debate, from Transhumanism and Posthumanism to Christian Digital Humanism, and outlining the key tenets of the Vatican document *Antiqua et nova*, the essay examines *algor-ethics* as a machine-side guardrail designed to embed ethical principles into algorithm development. It argues, however, that this protective measure alone is insufficient. The anthropomorphism of LLM-based conversational systems - identified in the technical literature as a source of emotional dependence, manipulation, empathy erosion, and dehumanisation - demands a complementary intervention on the human-user side. The paper therefore introduces *moral-algo*, a framework of ethical and pastoral guidance for human interaction with AI. It includes an experimental protocol termed *Rules of Prompt Pragmatics*, grounded in the linguistic de-anthropomorphisation of both system outputs and user inputs. The convergence of these two guardrails aims to safeguard human dignity and integrity in the age of AI.

ABSTRACT: Il presente contributo muove dalla consapevolezza, espressa dal Magistero della Chiesa cattolica e condivisa dall'Umanesimo digitale, che l'umanità si trova a un bivio epocale dinanzi allo sviluppo dell'intelligenza artificiale. Dopo aver ricostruito il quadro del dibattito contemporaneo - dal Transumanesimo al Postumanesimo, fino all'Umanesimo digitale cristiano - e aver sintetizzato i contenuti della Nota *Antiqua et nova*, si analizza il ruolo dell'*algor-etica*, intesa come guardrail lato macchina-agente volto a integrare

---

* Contributo sottoposto a valutazione dei pari - Peer-reviewed paper.

Il contributo sviluppa temi presentati in versione ridotta nella Prolusione all'Inaugurazione dell'Anno Giudiziario 2026 del Tribunale Ecclesiastico Interdiocesano Flaminio, a Bologna, il 12 febbraio 2026. La consultazione di pubblicazioni e documenti risulta aggiornata al 14 marzo 2026. Le citazioni tratte da opere, articoli e documenti in lingua inglese sono mantenute nell'originale.

Submitted: 17.03.2026 – Approved: 23.04.2026 – Published: 18.05.2026
DOI: https://doi.org/10.54103/1971-8543/31613

principi etici nella progettazione degli algoritmi. Si dimostra tuttavia che tale presidio, per quanto necessario, non è sufficiente: l'antropomorfismo dei sistemi conversazionali basati su LLM - documentato dalla letteratura tecnico-scientifica come fonte di rischi quali dipendenza emotiva, manipolazione, erosione dell'empatia e deumanizzazione - richiede un intervento complementare sul versante dell'uomo-utente. Viene quindi proposta la *moral-algo*, un quadro di orientamenti etico-pastorali per l'interazione umana con l'IA, che include un protocollo sperimentale denominato *Regole della Pragmatica del Prompt*, fondato sulla deantropomorfizzazione linguistica sia degli output del sistema sia degli input dell'utente. Il connubio dei due guardrail mira a preservare la dignità e l'integrità dell'essere umano nell'era delle IA.

PAROLE-CHIAVE: Artificial Intelligence, Algor-ethics, Moral-algo, Anthropomorphism, Dehumanization. Intelligenza artificiale, Algor-etica, Moral-algo, Antropomorfismo, Deumanizzazione.

1 - IA e Magistero nel quadro del dibattito contemporaneo

> "L'umanità si trova a un bivio dinanzi all'immenso potenziale generato dalla rivoluzione digitale guidata dall'Intelligenza Artificiale. L'impatto di questa rivoluzione è di vasta portata, trasformando campi come l'educazione, il lavoro, l'arte, l'assistenza sanitaria, l'amministrazione, l'ambito militare e la comunicazione"[1].

Queste parole di Papa Leone XIV esprimono meglio di qualunque altra la consapevolezza della Chiesa tanto del momento epocale in corso - una transizione che può implicare un mutamento antropologico radicale in direzioni sconosciute -, quanto dell'urgenza di scelte decisive: imboccare la strada sbagliata, davanti a quel bivio, potrebbe condurre a derive dalle conseguenze inimmaginabili.

Per questo la Chiesa avverte il desiderio "di prendere parte a questi dibattiti che riguardano direttamente il presente e il futuro della nostra famiglia umana", contribuendovi in modo "sereno e informato",

[1] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV, a firma del Cardinale Segretario di Stato Pietro Parolin, in occasione dell'AI for Good Summit*, 10 luglio 2025, p. 1 (testo integrale nel sito della Santa Sede *www.vatican.va*).

«sottolineando anzitutto la necessità di valutare le ramificazioni dell'intelligenza artificiale alla luce dello "sviluppo integrale della persona e della società" (Nota Antiqua et nova, n. 6)»[2].

D'altro canto, come è stato osservato, si tratta semplicemente dell'esercizio di quel

> "diritto nativo della Chiesa cattolica di annunciare sempre e dovunque i princìpi morali anche circa l'ordine sociale, e così pure pronunciare il giudizio su qualsiasi realtà umana, in quanto lo esigono i diritti fondamentali della persona umana o la salvezza delle anime (can. 747 § 2 CIC)"[3].

Infatti, di fronte

> "ai cambiamenti culturali, nel corso della storia, la Chiesa non è mai rimasta passiva; ha sempre cercato di illuminare in ogni tempo con la luce e la speranza di Cristo, di discernere il bene dal male, quanto di buono nasceva da quanto aveva bisogno di essere cambiato, trasformato, purificato"[4].

A maggior ragione quando il giudizio riguarda la risposta alle domande esistenziali sulle quali l'uomo si interroga da sempre e che oggi tornano ineludibili in una nuova forma:

> "cosa è dunque l'uomo, qual è la sua specificità e quale sarà il futuro di questa nostra specie chiamata *homo sapiens* nell'era delle intelligenze artificiali? Come possiamo rimanere pienamente umani e orientare verso il bene il cambiamento culturale in atto?"[5].

In premessa va osservato che la riflessione cattolica - ma anche di altre tradizioni religiose[6] - si inserisce in un dibattito filosofico e antropologico particolarmente vivo, che vede coinvolti filosofi, pensatori e futurologi. Nel corso di questo dibattito si sono delineate

---

[2] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti alla Seconda Conferenza Annuale su Intelligenza Artificiale, Etica e Governance d'impresa*, 19-20 giugno 2025, p. 2 (testo integrale in *www.vatican.va*).

[3] **R. SANTORO**, *Chiesa cattolica e intelligenza artificiale: dalla* Rome Call for AI Ethics *alle* Linee guida vaticane, in *Stato, Chiese e pluralismo confessionale*, Rivista telematica (*https://riviste.unimi.it/index.php/statoechiese/article/view/28764*), n. 4 del 2025, p. 40.

[4] **LEONE XIV**, *Saluto del Santo Padre Leone XIV agli Influencer e Missionari Digitali*, 29 luglio 2025, p. 3 (testo integrale in *www.vatican.va*).

[5] **FRANCESCO**, *Messaggio per la LVIII Giornata Mondiale delle Comunicazioni Sociali. Intelligenza artificiale e sapienza del cuore: per una comunicazione pienamente umana*, 24 gennaio 2024, p. 1 (testo integrale in *www.vatican.va*).

[6] Per un'analisi dell'approccio al tema da parte delle confessioni religiose musulmana, buddhista e taoista, si veda **R. SANTORO**, *Chiesa cattolica,* cit., pp. 54-57.

fondamentalmente due visioni: una che mantiene l'uomo al centro, un'altra che lo espelle da tale posizione, articolandosi in diverse correnti.

Partendo da queste ultime, si segnala innanzitutto il *Transumanesimo*, un movimento filosofico e culturale che promuove l'uso di nanotecnologia, biotecnologia e IA per superare i limiti umani fondamentali - invecchiamento e mortalità -, mediante l'accrescimento cognitivo e fisico, nella prospettiva di una fase evolutiva radicale[7].

[7] Per i principi di questa corrente di pensiero, si veda il suo Manifesto *The Transhumanist Declaration* (in *https://www.humanityplus.org/the-transhumanist-declaration*) nonché la *Carta dei Principi dei Transumanisti Italiani* (in *https://transumanisti.net/carta-dei-principi/*); per un approfondimento cfr. **N. BOSTROM**, *A History of Transhumanist Thought* (sul sito ufficiale dell'autore, *https://nick bostrom.com/papers/a-history-of-transhumanist-thought/*). Durante una conferenza al MIT l'8 ottobre 2025, Ray Kurzweil, informatico, inventore, imprenditore e futurista americano nonché uno degli esponenti chiave del transumanesimo tecnologico, ha dichiarato: "As we move forward, the lines between humans and technology will blur, until we are […] one and the same. In the 2030s, robots the size of molecules will go into our brains, noninvasively, through the capillaries, and will connect our brains directly to the cloud [...]. By 2045, once we have fully merged with AI, our intelligence will no longer be constrained […] it will expand a millionfold. This is what we call the singularity" (notizia e dichiarazioni in *https://news.mit.edu/2025/ray-kurzwei-reinforces-his-optimism-tech-progress-1010*).

Una declinazione ancora più estrema della visione transumanista è quella di *Iniziative 2045* del russo Dmitry Itskov, che mira a raggiungere un'immortalità cibernetica: in essa si vagheggia addirittura il trasferimento della coscienza umana a supporti non biologici attraverso la creazione di avatar robotici e olografici (si veda il sito del progetto *www.2045.com*). È difficile, leggendo queste 'ideazioni', non ricordare le parole di Papa Francesco a proposito di ideologie che sottendono l'ossessione di "accrescere oltre ogni immaginazione il potere dell'uomo" e del fatto che "l'intelligenza artificiale e i recenti sviluppi tecnologici si basano sull'idea di un essere umano senza limiti, le cui capacità e possibilità si potrebbero estendere all'infinito grazie alla tecnologia": **FRANCESCO**, *Esortazione apostolica* Laudate Deum *a tutte le persone di buona volontà sulla crisi climatica*, 4 ottobre 2023, p. 6, il cui testo integrale è edito nel sito ufficiale della Santa Sede (*www.vatican.va*). In altri termini, sempre per citare Papa Francesco, si tratta di "una prometeica presunzione di autosufficienza" (**ID.**, *Messaggio per la LVII Giornata Mondiale della Pace. Intelligenza artificiale e pace*, 1° gennaio 2024, p. 6; testo integrale in *www.vatican.va*), della "tentazione originaria di diventare *come* Dio *senza* Dio" (**ID.**, *Messaggio per la LVIII Giornata*, cit., p. 3; sul tema si veda **G. DEL MISSIER**, *Transumanesimo e intelligenza artificiale: aspetti etici e antropologici*, in *Apuleia Theologica*, II (2025), pp. 67-81).

Quindi il *Postumanesimo*, una corrente accademica che de-centra l'essere umano come misura universale, criticando l'antropocentrismo e le dicotomie cartesiane (mente/corpo, natura/cultura), ed è orientata a rifondare etica e ontologia in una prospettiva non-antropocentrica, focalizzata sulle relazioni ecologiche, le alterità (animali, macchine) e le 'soggettività nomadi e ibride'[8].

Infine il *Metaumanesimo*, una proposta filosofica che intende superare la contrapposizione tra la visione tecno-ottimista del Transumanesimo e quella critico-decostruttiva del Postumanesimo, fondandosi su un naturalismo non-antropocentrico, riconoscendo l'essere umano come entità non ontologicamente distinta dagli altri viventi e criticando la superiorità razionale dell'*anthropos*[9].

Un'aperta e serrata critica tanto al *Transumanesimo* quanto al *Postumanesimo* si rinviene nel Documento della Commissione Teologica Internazionale *Quo vadis, humanitas? Pensare l'antropologia cristiana di fronte ad alcuni scenari sul futuro dell'umano*, la cui pubblicazione è stata autorizzata il 9 febbraio 2026 dal Prefetto del Dicastero per la Dottrina Della Fede[10]: il testo rappresenta una riflessione sistematica sull'antropologia cristiana, sulla dignità e sull'identità dell'essere umano in un'epoca di cambiamenti tecnologici radicali. Al «sogno di "diventare come dei" (cf. *Gen* 3,4) di certo *transumanesimo* o *postumanesimo*»[11] viene contrapposto, in una

> "proposta teologica e pastorale riguardo alla vita umana intesa come *vocazione*"[12], «un "antropocentrismo situato", cioè una visione

[8] Per un approfondimento di questa corrente di pensiero, si vedano **R. BRAIDOTTI**, *The Posthuman*, Polity Press, Cambridge, 2013, e **F. FERRANDO**, *Philosophical Posthumanism*, Bloomsbury Academic, London, 2019. Una critica alle posizioni transumaniste e postumaniste si trova nella nota n. 9 del documento del **DICASTERO PER LA DOTTRINA DELLA FEDE - DICASTERO PER LA CULTURA E L'EDUCAZIONE**, *ANTIQUA ET NOVA.* Nota *sul rapporto tra intelligenza artificiale e intelligenza umana*, 14 gennaio 2025 (il cui testo integrale è edito nel sito ufficiale della Santa Sede *www.vatican.va*; considerati i reiterati riferimenti a questo testo all'interno del presente contributo, d'ora in poi, nelle note, per brevità, **AN**, con l'indicazione del numero del paragrafo citato). In essa vi si legge che tali posizioni «si basano su una percezione fondamentalmente negativa della corporeità, la quale è vista più come un ostacolo che come parte integrante dell'identità umana, chiamata anch'essa a partecipare della piena realizzazione della persona. Una tale visione negativa è in contrasto con una corretta comprensione della dignità umana. Pur sostenendo i genuini progressi scientifici, la Chiesa afferma che tale dignità si fonda sulla "persona come unità inscindibile" di corpo e anima, per cui essa "inerisce anche al suo corpo, il quale partecipa a suo modo all'essere immagine di Dio della persona umana"».

> del mondo che da un lato sostiene il valore peculiare e centrale dell'essere umano in mezzo al meraviglioso concerto di tutti gli esseri e dall'altro riconosce che la vita umana è incomprensibile e insostenibile senza le altre creature»[13].

Effettivamente, l'*Umanesimo digitale*, a differenza delle summenzionate correnti di pensiero, rimarca la centralità dell'umano nella tecnologia. Il suo obiettivo è la "salvaguardia e miglioramento delle condizioni di vita degli esseri umani attraverso l'impiego delle possibilità tecnologiche" e il loro controllo a livello culturale, sociale e politico; tale forma di nuovo umanesimo mira quindi "a configurare la digitalizzazione in modo tale che essa contribuisca all'umanizzazione del mondo", rammentando sempre che "la forma umana di esistenza non è

[9] Per i principi di questa corrente di pensiero, si veda *A* METAHUMANIST *MANIFESTO by Jaime del Val and Stefan Lorenz Sorgner* (disponibile in *https://metabody.eu/wp-content/uploads/2016/02/A-METAHUMANIST-MANIFESTO.pdf*). Per un approfondimento, cfr. da ultimo **S.L. SORGNER**, *On Transhumanism: The Moral Debate on Human Enhancement*, Penn State University Press, Philadelphia, 2020, nonché **M. BALISTRERI**, *Transhumanism According to Stefan Lorenz Sorgner: Why the Posthuman Project Requires Responsibility and Empathy*, in *Deliberatio. Studies in contemporary philosophical challenges*, I (2021), pp. 57-66.

[10] Il testo integrale nel sito della Santa Sede (*www.vatican.va*); d'ora in poi, nelle note, per brevità, **QV**, con l'indicazione del numero del paragrafo citato. Per una prima analisi e commento si vedano **I. PIRO**, *Quo vadis, humanitas?*, in *L'Osservatore Romano*, 4 marzo 2026 (disponibile in *https://www.osservatoreromano.va/it/news/2026-03/quo-052/quo-vadis-humanitas.html*), nonché **G. TRIDENTE**, *Cosa ci dice sull'IA (e sull'uomo) l'ultimo documento della Commissione Teologica Internazionale*, in *Anima digitale*, Substack (*https://giovannitridente.substack.com/p/cosa-ci-dice-sullia-e-sulluomo-lultimo*), 9 marzo 2026. Il Documento sviluppa la precedente critica alle posizioni transumaniste e postumaniste che, in modo più sintetico, era contenuta nella nota n. 9 del documento del **DICASTERO PER LA DOTTRINA DELLA FEDE - DICASTERO PER LA CULTURA E L'EDUCAZIONE**, *ANTIQUA ET NOVA. Nota sul rapporto*, cit. In essa vi si legge che tali posizioni «si basano su una percezione fondamentalmente negativa della corporeità, la quale è vista più come un ostacolo che come parte integrante dell'identità umana, chiamata anch'essa a partecipare della piena realizzazione della persona. Una tale visione negativa è in contrasto con una corretta comprensione della dignità umana. Pur sostenendo i genuini progressi scientifici, la Chiesa afferma che tale dignità si fonda sulla "persona come unità inscindibile" di corpo e anima, per cui essa "inerisce anche al suo corpo, il quale partecipa a suo modo all'essere immagine di Dio della persona umana"».

[11] **QV**, n. 24.

[12] **QV**, n. 18; infatti la "concezione della vita come vocazione è la prospettiva in cui si può/si deve collocare il processo decisivo e complesso dell'identità a livello personale e sociale" (*ivi*, n. 160).

[13] **QV**, n. 19.

una appendice dello sviluppo tecnico"[14]. Esso trova compiuta espressione nel *Vienna Manifesto on Digital Humanism* del maggio 2019[15], redatto da un gruppo internazionale di intellettuali, accademici e professionisti, che - nella consapevolezza dell'urgenza di prendere decisioni - chiama a riflettere e agire sullo sviluppo tecnologico contemporaneo e futuro, ponendo al centro i valori e i bisogni umani e proclamando una serie di principi fondamentali: democrazia e inclusione, privacy e libertà, regole e trasparenza, competitività contro monopoli, decisioni umane, interdisciplinarità scientifica, responsabilità

[14] Si vedano **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale. Un'etica per l'epoca dell'intelligenza artificiale*, FrancoAngeli, Milano, 2020, p. 15. Gli Autori sintetizzano nelle righe conclusive del libro: "L'umanesimo digitale non assume una posizione difensiva né intende frenare il progresso tecnologico nell'epoca dell'Intelligenza Artificiale. Vuole piuttosto favorire il progresso umano, utilizzando le opportunità digitali per rendere le nostre vite più ricche, più efficienti e più sostenibili. Non coltiva il sogno di una forma del tutto nuova di esistenza umana come fanno i transumanisti, rimane scettico nei confronti di aspettative utopistiche, ma è ottimista per quanto riguarda la capacità degli esseri umani di riuscire a plasmare le potenzialità digitali". Da ultimo, vedasi anche **H. WERTHNER**, *Digital Humanism. On Digitalization and Artificial Intelligence*, Springer, Cham, 2025.

[15] Il Manifesto, che è già stato sottoscritto da oltre 1000 leader mondiali, è disponibile in *https://caiml.org/dighum/dighum-manifesto/*. Va ricordato che un precedente squisitamente italico si può rinvenire in **M. CHIARIATTI et al.,** *Manifesto per l'IA. L'intelligenza artificiale non è nemica: dobbiamo imparare a lavorarci insieme*, in *Il Sole 24 ore*, 23 maggio 2022 (*https://www.ilsole24ore.com/art/l-intelligenza-artificiale-non-e-nemica-dobbiamo-imparare-lavorarci-insieme-AEAbB8YB*).

universitaria, dialogo fra ricercatori e società, etica professionale, nuovi *curricula* educativi, istruzione precoce[16].

Per una visione di un Umanesimo digitale cristiano rimane invece insuperabile l'apporto del teologo morale Giannino Piana[17]. Egli sottolinea come di fronte al pericolo del transumanesimo, che produce uno "svuotamento dell'umano" assumendo "così i tratti di un antiumanesimo",

> "diviene allora importante procedere alla delineazione dei presupposti di un nuovo umanesimo, che recuperi la lezione delle grandi tradizioni del passato, aggiornandole con attenzione ai nuovi scenari aperti dagli sviluppi della scienza e della tecnica"[18].

Tre, a suo avviso, sono i lineamenti imprescindibili "più qualificanti di una antropologia umanista". Il primo è "il recupero della dimensione misterica della persona, della sua unicità e irripetibilità, e perciò dell'impossibilità di una sua totale oggettivazione". Il secondo è la

> «adesione a una visione solidale dell'umano, che implica il superamento di un'interpretazione individualistica dell'uomo per fare propria un'interpretazione personalistica, capace di integrare in se stessa individualità e relazione. In quanto persona, l'uomo è infatti soggetto unico e soggetto relazionale: la relazione cioè, lungi

[16] È doveroso menzionare anche l'*Onlife Manifesto* rilasciato l'8 febbraio 2013, che si trova in **L. FLORIDI**, *The Onlife Manifesto. Being Human in a Hyperconnected Era*, Springer, London, 2015, pp. 7-13. Il titolo del Manifesto è spiegato nell'Introduzione del volume: «We decided to adopt the neologism "onlife" that I had coined in the past in order to refer to the new experience of a hyperconnected reality within which it is no longer sensible to ask whether one may be online or offline". Frutto del lavoro di un gruppo di studiosi presieduti da Luciano Floridi, docente di *Practice of Cognitive Science* presso la Yale University, esso parte da una critica ai quadri concettuali dualistici e moderni (ad esempio, tra umano/macchina ed entità/interazioni) che sono inadatti per comprendere l'era iperconnessa e le sue trasformazioni. Lo scopo è il "*re-engineering* concettuale" per aggiornare tali quadri, permettendo così ai *policymaker* di affrontare le sfide etiche e politiche in modo costruttivo e con maggiore fiducia nel futuro. Le sue proposte centrali mirano a stabilizzare una concezione politica del Sé relazionale (libero, ma intrinsecamente sociale), sviluppare una alfabetizzazione digitale critica che riconosca l'intreccio con la tecnologia, e proteggere le capacità attentive umane come risorsa critica e limitata. Il Manifesto promuove il riconoscimento che la libertà è legata alla pluralità e alla capacità di dare inizio a qualcosa di nuovo in modo imprevisto, e non alla sovranità o al controllo totale.

[17] In particolare con il suo libro **G. PIANA**, *Umanesimo per l'era digitale. Antropologia, etica, spiritualità*, Interlinea, Novara, 2022.

[18] **G. PIANA**, *Umanesimo*, cit., p. 38 s.

> dall'essere qualcosa di esterno o di accidentale, rientra a tutti gli effetti nella definizione della sua natura e diviene fattore costitutivo della sua identità[19]. […] Infine, il terzo lineamento rinvia alla necessità dell'apertura a una prospettiva trascendente, dalla quale viene all'uomo una costante tensione in avanti che lo spinge ad andare costantemente "oltre"»[20].

Sul contributo della Chiesa Cattolica a questo dibattito[21] - in attesa di un'enciclica che tratti organicamente l'argomento[22], e a prescindere

[19] Il tema ritorna esplicito in **QV**, n. 119, che poi ricorda al n. 85: "Tali relazioni, costitutive dell'identità personale, sono strutturate nella famiglia, in un popolo, con le sue tradizioni, e nell'appartenenza più vasta all'umano comune" (cfr. anche *ivi*, n. 137). A esplorare il contenuto di queste relazioni sono quindi dedicati i successivi paragrafi nn. 86-98.

[20] **G. PIANA**, *Umanesimo*, cit., p. 39 s. Leggiamo sempre in **QV**, n. 62: "A livello dell'individuo è l'anima immortale che dà forma, che unifica cioè e organizza la materia in un corpo vivente, conferendo all'essere umano una trascendenza che il *post-* e il *transumanesimo* non possono né raggiungere né superare".

[21] Per un'accurata ricostruzione della dottrina della Chiesa, fino ad aprile 2022, tramite l'analisi del pensiero dei Pontefici, della riflessione accademica a opera di istituzioni vaticane durante le rispettive assemblee plenarie e degli approfondimenti di una parte della stampa cattolica di riferimento attraverso le loro pubblicazioni periodiche, si rimanda al lavoro di **G. TRIDENTE**, *ANIMA DIGITALE. La Chiesa alla prova dell'Intelligenza Artificiale*, Tau Editrice, Todi, 2022. Stante l'esaustività dell'analisi pregressa, nel presente contributo ci concentreremo prevalentemente, anche se non esclusivamente, su materiali ulteriori e/o successivi. Inoltre per una disamina aggiuntiva sino al giugno 2024, si veda **R. SANTORO, P. PALUMBO, F. GRAVINO**, *Diritto canonico digitale*, Editoriale Scientifica, Napoli, 2024, pp. 69-95, nonché, sino a gennaio 2025, **R. SANTORO**, *Chiesa cattolica*, cit., pp. 37-62. A *Il discernimento del Magistero su sviluppo e tecnologia* sono poi interamente dedicati i paragrafi di **QV**, nn. 25-30. Invece, per una riflessione teologica, filosofica ed educativa sul 'mondo del digitale', con il contributo di diversi autori provenienti da vari ambiti scientifici, al fine di esplorare i versanti relativi alla conoscenza e alla formazione degli strumenti digitali per coloro che sono impegnati in ruoli educativi, con un'attenzione alle implicazioni per la religiosità e la fede, si veda **AA. VV.**, *Umanesimo digitale. Educarsi al digitale per educare*, a cura di L. TONELLO, Triveneto Theology Press, Padova, 2023.

[22] Durante il suo primo incontro con i Cardinali subito dopo l'elezione, nello spiegare le ragioni della scelta del nome, Papa Leone XIV ha dichiarato: "Papa Leone XIII, con la storica Enciclica *Rerum Novarum* affrontò la questione sociale nel contesto della prima grande Rivoluzione industriale. Oggi la Chiesa offre a tutti il suo patrimonio di Dottrina sociale per rispondere a un'altra rivoluzione industriale e agli sviluppi dell'intelligenza artificiale, che comportano nuove sfide per la difesa della dignità umana, della giustizia e del lavoro": la circostanza è riferita nell'articolo di **A. SPADARO**, *La sfida dell'IA da Francesco a Leone*, in *L'Osservatore Romano*, 13 agosto 2025 (*https://www.osservatoreromano.va/it/news/2025-08/quo-187/la-sfida-dell-ia-da-francesco-a-leone.html*); si veda anche **QV**, n. 4.

dalla copiosa produzione di Papa Francesco (una trentina di interventi durante il suo pontificato), dai più recenti discorsi di Papa Leone XIV (il penultimo dei quali di particolare rilievo[23]), e dalle riflessioni del menzionato documento *Quo vadis, humanitas*? -, il testo di maggiore sistematicità[24] è finora la citata Nota *Antiqua et nova* sul rapporto tra intelligenza artificiale e intelligenza umana[25], che già nelle prime righe dell'Introduzione dichiara l'intento di rispondere, per l'appunto, alla chiamata alla riflessione[26]:

> "Con antica e nuova sapienza (cf. *Mt* 13,52) siamo chiamati a considerare le odierne sfide e opportunità poste dal sapere scientifico e tecnologico, in particolare dal recente sviluppo dell'intelligenza artificiale (IA)"[27].

---

[23] Oltre ai primi tre sopracitati, fra quelli specificamente in materia digitale e di intelligenza artificiale vanno annoverati in ordine di tempo: **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti del Builders AI Forum,* 6-7 novembre 2025, pp. 1-2; **ID.**, *Messaggio del Santo Padre Leone XIV ai partecipanti al Congresso Internazionale della Pontificia Accademia per la Vita: "AI and Medicine: The Challenge of Human Dignity"*, 10-12 novembre 2025, pp. 1-3; **ID.**, *Discorso del Santo Padre Leone XIV ai partecipanti alla Conferenza "Artificial Intelligence and Care for our Common Home" organizzata da Fondazione Centesimus Annus Pro Pontifice e Strategic Alliance of Catholic Research University*, 5 dicembre 2025, pp. 1-3; **ID.**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata Mondiale delle Comunicazioni Sociali. Custodire voci e volti umani*, 24 gennaio 2026, pp. 1-7; **ID.**, *Message of Pope Leo XIV, signed by the Cardinal Secretary of State, Pietro Parolin, on the occasion of the International Day of Mathematics,* 14 marzo 2026 (i testi integrali nel sito della Santa Sede *www.vatican.va*).

[24] In esso sono confluite, subendo un considerevole ampliamento e approfondimento, le riflessioni contenute in tre discorsi particolarmente significativi tenuti da Papa Francesco nel 2024 in tema di intelligenza artificiale: **FRANCESCO**, *Messaggio per la LVII Giornata*, cit., pp. 1-11; **ID.**, *Messaggio per la LVIII Giornata*, cit., pp. 1-7; **ID.**, *Discorso del Santo Padre Francesco alla Sessione del G7 sull'Intelligenza Artificiale. Uno strumento affascinante e tremendo*, 14 giugno 2024, pp. 1-6 (i testi integrali nel sito della Santa Sede *www.vatican.va*).

[25] Per un commento si veda **F. PATSCH**, «*Antiqua et nova». L'intelligenza artificiale al servizio della dignità umana e del bene comune*, in *La civiltà cattolica*, CLXXVI (2025), pp. 207-218.

[26] La **AN**, alla nota 98, indica anche un altro testo come contributo a questo dibattito dal punto di vista della Chiesa Cattolica: "Per un'ulteriore discussione circa le questioni etiche sollevate dall'IA a partire da una prospettiva cristiana cattolica, si veda Gruppo di Ricerca sull'AI del Centro per la Cultura Digitale del Dicastero per la Cultura e l'Educazione, *Encountering Artificial Intelligence: Ethical and Anthropological Investigations* (Theological Investigations of Artificial Intelligence, 1), a cura di M.J. GAUDET, N. HERZFELD, P. SCHERZ, J.J. WALES, Pickwick, Eugene 2024, 147-253".

[27] **AN**, n. 1.

Il suo contenuto pone la persona umana, la sua dignità inalienabile e la sua intelligenza integrale come criterio etico per lo sviluppo dell'IA. La Nota, che si compone di 117 paragrafi con un apparato di 215 note, è strutturata in sei sezioni principali, ciascuna delle quali affronta un aspetto cruciale dell'interazione tra l'IA e l'umanità, proponendo così "alcune linee guida, allo scopo di assicurare che lo sviluppo e l'uso dell'IA rispettino la dignità umana e promuovano lo sviluppo integrale della persona e della società"[28]. Vediamole in estrema sintesi.

Dopo aver ricordato nell'*Introduzione* che è necessario un discernimento etico e antropologico che assuma come punto di partenza il dono dell'intelligenza umana, creata a immagine di Dio, nella sezione II, *Che cos'è l'intelligenza artificiale?*, emerge immediatamente una distinzione cruciale: mentre l'intelligenza umana si configura come realtà integrale e irriducibile, l'intelligenza artificiale resta invece funzionale e puramente computazionale.

Per comprendere appieno tale differenza, nella successiva sezione, intitolata *L'intelligenza nella tradizione filosofica e teologica*, il testo offre una visione completa dell'intelligenza umana, sviluppando un'articolata riflessione che abbraccia molteplici dimensioni in una serie di sottosezioni. Nella prima, *Razionalità*, si distingue tra *intellectus*, inteso come intuizione immediata della verità, e *ratio*, quale processo discorsivo, che permea e modella tutte le attività umane. Ma queste facoltà intellettuali, come si rammenta nella sottosezione *Incarnazione*, non possono essere comprese se disgiunte dalla loro dimensione incarnata: l'essere umano costituisce un'unità inscindibile di corpo e anima, e la sua intelligenza, pur radicandosi nell'esistenza corporea, trascende nondimeno il mondo materiale grazie alla sua anima. A ciò si aggiunge la dimensione della *Relazionalità*, poiché l'intelligenza si esercita autenticamente nella comunione interpersonale e trova il proprio fondamento ultimo nell'amore trinitario di Dio, di cui l'amore e il servizio per il prossimo costituiscono la piena risposta alla vocazione umana. La *Relazione con la Verità*, titolo di un'ulteriore sottosezione, rivela poi come l'intelligenza sia un dono divino fatto per cogliere il vero, spingendo l'uomo oltre l'utilità sensoriale immediata. Tale ricerca della verità si manifesta nella comprensione semantica e nella creatività, culminando nell'apertura a Dio. Parimenti, nella sottosezione la *Custodia del mondo*, si

[28] **AN**, n. 6.

rammenta che all'uomo, immagine di Dio, è affidato il compito di "custodire" e "coltivare" la creazione: la sua intelligenza riflette così la Sapienza divina e al contempo è chiamato a sviluppare le proprie capacità nella scienza e nella tecnica, mentre il mondo stesso diviene via attraverso cui la mente può "ascendere gradualmente" verso il "divino Artigiano".

Da questa complessa riflessione emerge *Una comprensione integrale dell'intelligenza umana*, altra sottosezione del testo, come facoltà che coinvolge la persona nella sua totalità, includendo aspetti spirituali, cognitivi, incarnati e relazionali. Essa si manifesta in modi multiformi: dalla creatività alla risoluzione di problemi, dalla perizia artigianale alla saggezza relazionale. Questa intelligenza implica l'apertura della persona alle domande ultime della vita e rispecchia un orientamento verso il Vero, il Buono e il Bello, accedendo alla totalità dell'essere che non si esaurisce in ciò che è meramente misurabile. In contrapposizione a ciò, i *Limiti dell'IA* divengono evidenti: essa manca di corporeità attraverso la quale evolvere, di discernimento morale, di capacità di stabilire autentiche relazioni e non è in grado di cogliere il senso della totalità delle cose. Di conseguenza, *Il ruolo dell'etica nel guidare lo sviluppo e l'uso dell'IA*, cui è dedicata la sezione IV, assume un'importanza decisiva[29].

Al discernimento etico applicato ai principali settori sociali, economici e militari, individuati in una serie di ulteriori sottosezioni, è dedicata la sezione V, *Questioni specifiche*.

Nel contesto sociale (*IA e la società*) l'intelligenza artificiale potrebbe essere utilizzata per promuovere lo sviluppo umano e il bene comune, ma presenta il pericolo di aumentare le disuguaglianze e favorire il "paradigma tecnocratico", ossia la tendenza a risolvere tutti i problemi del mondo attraverso i soli mezzi tecnologici[30]. Per quanto concerne le relazioni umane (*IA e le relazioni umane*) emergono rischi di ostacolare l'incontro autentico e l'empatia, con un chiaro avvertimento contro l'antropomorfizzazione dell'intelligenza artificiale[31]. Nell'ambito economico e lavorativo (*IA, economia e lavoro*), si evidenzia il rischio di dequalificare i lavoratori e sostituirli anziché assisterli, ribadendo che il lavoro deve servire l'uomo nella sua integralità. Nel settore sanitario (*L'IA e la sanità*), pur riconoscendo l'utilità dell'intelligenza artificiale

[29] Su tale sezione qui soprassediamo, rinviando al paragrafo successivo.

[30] Su questo paradigma, cfr. anche **QV**, n. 29.

[31] A questa tematica specifica è dedicato l'intero paragrafo terzo, cui rinviamo.

nell'assistenza, si insiste affinché la responsabilità e la relazione medico-paziente rimangano saldamente umane. L'educazione (*IA ed educazione*) richiede particolare attenzione: occorre promuovere il pensiero critico evitando la dipendenza tecnologica poiché il rapporto docente-studente resta insostituibile. Le questioni legate a *IA, disinformazione, deepfake e abusi* rendono necessaria la trasparenza per contrastare la diffusione, intenzionale o accidentale, di informazioni false che minano la fiducia sociale. Riguardo a *IA, privacy e controllo*, il documento esprime un netto rifiuto dell'uso dell'intelligenza artificiale per la sorveglianza e il "credito sociale", pratiche che riducono la persona a un mero insieme di dati. *L'IA e la protezione della casa comune* è altro ambito trattato, evidenziando l'impatto ambientale dell'infrastruttura dell'intelligenza artificiale e invitando a rifiutare l'antropocentrismo distorto del paradigma tecnocratico. Particolarmente urgente appare la richiesta di bandire le armi autonome letali (*L'IA e la guerra*), garantendo che nessuna macchina possa mai scegliere di togliere la vita a un essere umano. Infine, nella sottosezione *L'IA e il rapporto dell'umanità con Dio*, viene denunciata la tentazione di cercare nell'intelligenza artificiale un sostituto di Dio, configurando tale atteggiamento come idolatria.

La VI e ultima sezione contiene la *Riflessione finale*, che richiama l'umanità alla necessità di sviluppare una responsabilità, una coscienza e valori proporzionati all'immenso potere tecnologico oggi disponibile. Si esorta pertanto a rinvigorire la "sapienza del cuore" per guidare l'intelligenza artificiale verso il bene comune.

L'invito complessivo che scaturisce dal documento è chiaro: l'intelligenza artificiale deve essere utilizzata come uno strumento complementare all'intelligenza umana, mai come sostituto del pensiero, della relazionalità o della saggezza morale e spirituale. Solo quando la tecnologia è guidata da una sapienza del cuore - che l'umanità non può pretendere dalle macchine - e dall'amore per il prossimo, essa può servire autenticamente la vocazione integrale dell'uomo.

2 - L'*algor-etica*: un guardrail lato macchina-agente

Sebbene il termine *algor-etica*, o algoretica, non trovi menzione esplicita nella Nota *Antiqua et nova*, tuttavia essa costituisce un orientamento concettuale presente nella riflessione della Chiesa, che lo ha autorevolmente recepito, anche a livello di magistero pontificio. La si

può considerare come una branca della disciplina più ampia dell'Etica delle IA: mentre questa riguarda tutti gli aspetti morali, sociali, legali e politici legati allo sviluppo e all'uso dell'intelligenza artificiale in generale, l'algoretica si focalizza in modo specifico sull'integrazione di principi etici nel design, nell'implementazione e nell'uso degli algoritmi stessi. Quindi, l'algoretica è una parte specializzata dell'Etica delle IA, dedicata alla dimensione etico-morale dell'azione algoritmica e delle decisioni automatizzate.

Il termine fu coniato da Luigi Lombardi Vallauri[32], ma è stato il teologo morale francescano padre Paolo Benanti a sistematizzare l'algoretica come disciplina, definendola '*AI for Humanity*'[33], e configurandola come una risposta all'*algo-crazia*, letteralmente il dominio degli algoritmi:

> "Le implicazioni sociali ed etiche delle IA e degli algoritmi rendono necessaria tanto un'algor-etica quanto una governance di queste invisibili strutture che regolano sempre più il nostro mondo per evitare forme disumane di quella che potremmo definire una algo-crazia"[34].

Dunque, di fronte al rischio di un potenzialmente disumano "governo della società per mezzo degli algoritmi"[35], si contrappone l'implementazione di un'algoretica e una governance adeguata delle strutture algoritmiche. Scrive ancora padre Benanti:

> "I sistemi algoritmici a volte vengono presentati alle persone come apolitici, tecnocratici e privi di valore. Sono tutto tranne questo. Poiché sono costruiti tramite operazioni di giudizio e sono frutto di discernimento, sono atti fondamentalmente etici. L'etica è nel cuore della realizzazione di algoritmi, e principi etici trasparenti e comunicabili devono essere inclusi nel tradurre compiti in algoritmi, specie quelli che modificano o interagiscono con la vita di milioni di persone"[36].

---

[32] Cfr. **L. LOMBARDI VALLAURI**, *Algoretica e Informatica giuridica*, in *Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale*, XV (2022), p. 32.

[33] Si veda una delle sue relazioni più significative dal titolo "*Algor-Ethics: Developing a Language for a Human-Centered AI*", tenuta durante il TEDxRoma nel luglio 2018 (in *https://www.youtube.com/watch?v=rFzjsHNertc*).

[34] **P. BENANTI**, *Oracoli. Tra algo*retica *e algo*crazia, Luca Sossella Editore, Roma, 2018, p. 38.

[35] **P. BENANTI**, *Oracoli*, cit., p. 42.

[36] **P. BENANTI**, *Oracoli*, cit., p. 50

L'algoretica implica pertanto lo studio dei problemi morali, giuridici e sociali connessi all'azione algoritmica, con l'obiettivo di sviluppare principi e regole che possano guidare un uso responsabile e umanamente orientato degli algoritmi. A tal fine si richiede una collaborazione interdisciplinare tra sviluppatori, filosofi, legislatori e società civile per assicurare una governance globale dell'intelligenza artificiale che sia etica e sostenibile:

> «nel termine "algoretica" si condensano una serie di principi che si dimostrano essere una piattaforma globale e plurale in grado di trovare il supporto di culture, religioni, organizzazioni internazionali e grandi aziende protagoniste di questo sviluppo»[37].

Va rilevato che il tema e le preoccupazioni per le potenziali derive etiche delle macchine artificiali e per il loro controllo hanno radici assai lontane nel tempo. La questione si delineò inizialmente negli anni '40, con riferimento al mondo dei robot, nella mente dello scrittore di fantascienza Isaac Asimov e fu risolta, nella finzione letteraria, con la formulazione, che apparve nei suoi romanzi, delle celeberrime *Tre leggi della robotica*:

> "1. Un robot non può recare danno a un essere umano né può permettere che, a causa del suo mancato intervento, un essere umano riceva danno.
> 2. Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non vadano in contrasto alla Prima Legge.
> 3. Un robot deve proteggere la propria esistenza, purché la salvaguardia di essa non contrasti con la Prima o con la Seconda Legge"[38].

In tempi più recenti, nel settembre 2010, un gruppo di lavoro multidisciplinare formato da quattordici esperti inglesi, fra cui Alan

[37] **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7,* cit., p. 4.

[38] *Tre leggi della robotica*, in *Wikipedia. L'enciclopedia libera (https://it.wikipedia.org/wiki/Tre_leggi_della_robotica*, consultato il 12 febbraio 2026), ove si riferisce che in seguito fu aggiunta una legge Zero, per mantenere il fatto che una legge con numero più basso soprassedesse a una con numero maggiore, con conseguenziale modifica delle originarie tre leggi: "Un robot non può recare danno all'umanità, né può permettere che, a causa del proprio mancato intervento, l'umanità riceva danno". Un'eco di queste leggi, e in particolare della legge Zero, può trovarsi in un passaggio della *Rome Call for AI Ethics* - di cui a breve - ove leggiamo: "AI system must be conceived, designed and implemented to serve and protect human beings and the environment in which they live. […] it must have the good of humankind and the good of every human being at its heart".

Winfield, docente di *Robot Ethics* presso l'University of the West of England, considerando la crescente presenza della robotica nel mondo, sia nelle case che nell'industria, e prevedendone un impatto significativo nella vita domestica, nelle istituzioni e nelle economie nazionali e globali, avvertì l'esigenza di affrontare le implicazioni etiche della diffusione dei robot al di fuori dei laboratori di ricerca. Così, partendo proprio da quelle leggi, di cui si constatavano vari profili di inapplicabilità al mondo reale attuale, fu deciso di proporre, in una soluzione di compromesso, cinque *Principles of Robotics*, ovvero precetti morali per ricercatori, progettisti, produttori, fornitori e manutentori di robot, pubblicati online nel 2011[39].

Tornando alla riflessione specificamente ecclesiale e all'algoretica, il termine appare in documenti ufficiali dapprima nel magistero

[39] "1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security. 2. Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws and fundamental rights & freedoms, including privacy. 3. Robots are products. They should be designed using processes which assure their safety and security. 4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent. 5. The person with legal responsibility for a robot should be attributed". In seguito i principi sono stati inclusi nello standard britannico BS 8611 (*Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems*); a questi cinque principi etici si aggiungevano anche sette messaggi di alto livello, che costituivano una serie di dichiarazioni generali il cui scopo era incoraggiare la responsabilità all'interno della comunità industriale e di ricerca robotica, e in questo modo ottenere fiducia nel lavoro che essa svolge (*https://research.gold.ac.uk/id/eprint/19743/1/principles-robotics.pdf*.

pontificio di Papa Francesco[40] e poi nella *Rome Call for AI Ethics*, che ne rappresenta la piena consacrazione[41]. Vi leggiamo infatti:

> «To achieve these objectives, we must set out from the very beginning of each algorithm's development with an "algor-ethical" vision, i.e. an approach of ethics by design. Designing and planning AI systems that we can trust involves seeking a consensus among political decision-makers, UN system agencies and other intergovernmental organisations, researchers, the world of

[40] Papa Francesco ne parla in particolare una prima volta a fine 2019: «Faccio quindi appello agli ingegneri informatici, perché si sentano anch'essi responsabili in prima persona della costruzione del futuro. Tocca a loro, con il nostro appoggio, impegnarsi in uno sviluppo etico degli algoritmi, farsi promotori di un nuovo campo dell'etica per il nostro tempo: la "algor-etica"»: **FRANCESCO**, *Discorso del Santo Padre Francesco ai partecipanti al Congresso Child Dignity in the Digital World*, 14 novembre 2019, p. 5, il cui testo integrale è nel sito della Santa Sede (*www.vatican.va*). Papa Francesco vi ritornerà anche anni dopo: "Uno sguardo umano e il desiderio di un futuro migliore per il nostro mondo portano alla necessità di un dialogo interdisciplinare finalizzato a uno sviluppo etico degli algoritmi - l'algor-etica -, in cui siano i valori a orientare i percorsi delle nuove tecnologie [12]. Le questioni etiche dovrebbero essere tenute in considerazione fin dall'inizio della ricerca, così come nelle fasi di sperimentazione, progettazione, produzione, distribuzione e commercializzazione. Questo è l'approccio dell'etica della progettazione, in cui le istituzioni educative e i responsabili del processo decisionale hanno un ruolo essenziale da svolgere": **ID.**, *Messaggio per la LVII Giornata Mondiale della Pace*, cit., p. 8.

[41] La *Rome Call for AI Ethics* è sia edita nel sito della Santa Sede (*www.vatican.va*) sia disponibile in *https://www.romecall.org/*. Nell'incontro con i partecipanti alla plenaria della Pontificia Accademia per la Vita, in occasione della firma del documento, Papa Francesco dichiarò: «Sono molte le competenze che intervengono nel processo di elaborazione degli apparati tecnologici (ricerca, progettazione, produzione, distribuzione, utilizzo individuale e collettivo), e ognuna comporta una specifica responsabilità. Si intravede una nuova frontiera che potremmo chiamare "algor-etica" [...]. Essa intende assicurare una verifica competente e condivisa dei processi secondo cui si integrano i rapporti tra gli esseri umani e le macchine nella nostra era. […] L'"algor-etica" potrà essere un ponte per far sì che i principi si inscrivano concretamente nelle tecnologie digitali, attraverso un effettivo dialogo transdisciplinare»: **FRANCESCO**, *Discorso preparato dal Santo Padre Francesco, letto da Mons. Vincenzo Paglia, Presidente della Pontificia Accademia per la Vita*, 28 febbraio 2020, p. 4 (testo integrale in *www.vatican.va*). L'importanza della *Rome Call for AI Ethics* nella riflessione ecclesiale viene evidenziata nuovamente nel 2024 da **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 4, il quale ricorda di avere salutato con favore la nascita di questo testo. Su tale documento cfr. **V. PAGLIA**, *Tecnologie digitali ed etica*, in *Rivista di scienze dell'educazione*, LIX (2021), pp. 68-80; **G. TRIDENTE**, *ANIMA DIGITALE*, cit., pp. 83-106; **R. SANTORO**, *Chiesa cattolica*, cit., pp. 46-52, e **R. SANTORO, P. PALUMBO, F. GRAVINO**, *Diritto canonico digitale*, cit., pp. 77-84.

> academia and representatives of non-governmental organizations regarding the ethical principles that should be built into these technologies. For this reason, the sponsors of the call express their desire to work together, in this context and at a national and international level, to promote "algor-ethics", namely the ethical use of AI [...]».

Secondo tale documento, sottoscritto il 28 febbraio 2020 dalla Pontificia Accademia per la Vita, unitamente a Microsoft, IBM e FAO (cui in seguito si sono aggiunti molti altri), sei sono i principi che dovrebbero guidare l'implementazione degli algoritmi per consentire un uso etico dell'IA: trasparenza, inclusione, responsabilità, imparzialità, affidabilità, sicurezza-privacy[42].

Alla luce di ciò, anche se non la menziona esplicitamente, la Nota *Antiqua et nova* risulta profondamente permeata dalla prospettiva dell'algoretica.

In primo luogo, il suo spirito aleggia in tutta la sezione IV del documento, nella quale si affrontano diffusamente varie preoccupazioni etiche, fra cui quelle che sorgono nel disegnare e nell'usare le intelligenze artificiali. Sotto il primo profilo, che interessa specificamente questo paragrafo[43], si ricorda che

> "la dimensione etica assume primaria importanza perché sono le persone a progettare i sistemi e a determinare per quali scopi essi vengano usati"[44]; infatti i "prodotti tecnologici riflettono la visione del mondo dei loro sviluppatori, proprietari, utenti e regolatori"[45] e "perseguono gli obiettivi che sono stati loro assegnati dagli esseri umani e sono governati da processi stabiliti da coloro che li hanno progettati e programmati"[46].

[42] "1. **Transparency**: in principle, AI systems must be explainable; 2. **Inclusion**: the needs of all human beings must be taken into consideration so that everyone can benefit and all individuals can be offered the best possible conditions to express themselves and develop; 3. **Responsibility**: those who design and deploy the use of AI must proceed with responsibility and transparency; 4. **Impartiality**: do not create or act according to bias, thus safeguarding fairness and human dignity; 5. **Reliability**: AI systems must be able to work reliably; 6. **Security and privacy**: AI systems must work securely and respect the privacy of users".

[43] All'altro profilo, quello dell'uso, sarà interamente dedicato il paragrafo 4.

[44] **AN**, n. 39.

[45] **AN**, n. 41.

[46] **AN**, n. 45.

Ecco allora che "sia i fini sia i mezzi usati in una data applicazione dell'IA, così come la visione generale che essa incorpora, devono essere valutati per assicurarsi che rispettino la dignità umana[47] e promuovano il bene comune"[48]. L'algoretica proposta dalla riflessione ecclesiale vuole assolvere questa funzione valutativa.

In secondo luogo, in tutta la Nota troviamo espressione dei principi della *Rome Call for AI Ethics*.

Così, fanno riferimento al principio di *Trasparenza* i passaggi che, partendo dalla considerazione dell'opacità computazionale dei sistemi di IA - ostacolo all'attribuzione di causalità e responsabilità -, richiedono che gli algoritmi siano progettati per essere trasparenti nel funzionamento, mitigando bias e permettendo la rendicontazione in ogni fase decisionale e la verificabilità da enti di controllo esterni, al fine di contrastare gli eccessi di sorveglianza[49].

Al principio di *Inclusione* possono essere ricondotti i brani che si soffermano sulla necessità di identificare e supportare gruppi vulnerabili. Cruciale infatti diventa contenere il rischio che l'IA amplifichi disuguaglianze esistenti (come il 'divario digitale' o l'accesso iniquo alle cure), per cui le implementazioni devono essere orientate a migliorare l'accessibilità ai servizi essenziali (ad esempio, l'istruzione) e

[47] La decisione di progettare un'intelligenza artificiale che incorporasse *by design* principi e valori irrispettosi della dignità umana, e quindi anti-umana ancora prima che anti-cristiana, costituirebbe un atto moralmente gravissimo, poiché investirebbe di responsabilità etica diretta chi compie tale scelta progettuale. Insegna infatti **G. PIANA**, *Umanesimo*, cit., p. 68 s., parlando della "etica della responsabilità": «Se è infatti vero, da un lato, che il primato va riservato al fine, non è meno vero, dall'altro, che il mezzo non può essere considerato "neutrale", ma che esso ha di per sé uno spessore morale che non può essere eluso. È come dire che "il fine non giustifica il mezzo", ma che occorre esaminare, di volta in volta, la relazione che tra fine e mezzo si istituisce nel contesto di un quadro valoriale correttamente gerarchizzato, riconoscendo che esistono mezzi i quali, per la loro intrinseca immoralità, rendono improponibile la messa in atto di interventi anche con obiettivi altamente positivi». E infatti **AN**, con specifico riguardo alle IA, ci rammenta che "l'attività tecnico-scientifica non ha carattere neutro, essendo un'impresa *umana* che chiama in causa le dimensioni umanistiche e culturali dell'ingegno umano" (n. 36) e che quindi "a essere eticamente significativi non sono soltanto i fini, ma anche i mezzi impiegati per raggiungerli; inoltre, sono importanti anche la visione generale e la comprensione della persona incorporate in tali sistemi" (n. 41). Sulle riflessioni degli ultimi due Pontefici relativamente a questo tema, si rinvia alla nota 72.

[48] **AN**, n. 42.

[49] Cfr. **AN**, nn. 44, 46, 84, 93.

impedire che le soluzioni tecnologiche favoriscano involontariamente le popolazioni più abbienti[50].

Quanto al principio di *Responsabilità*, si rammenta che essa non può essere delegata all'IA, in quanto solo gli esseri umani sono agenti morali. La progettazione di sistemi di intelligenza artificiale deve includere meccanismi per identificare chiaramente il responsabile umano (*accountability*); è indispensabile mantenere una sua tracciabilità, garantendo un controllo umano significativo sul processo di scelta algoritmica, per evitare deleghe totali in ambiti, come quello sanitario, in cui la decisione finale deve sempre rimanere in capo all'uomo[51].

Il principio di *Imparzialità* trova invece espressione nella rivendicata esigenza di una mitigazione proattiva dell'*algorithmic bias*, definito come la produzione di errori sistematici che penalizzano involontariamente specifici gruppi. La radice del problema tecnico risiede nei dati di addestramento viziati da pregiudizi sociali. L'obiettivo di design è quindi evitare che gli strumenti di IA, in particolare in settori sensibili come la sanità, moltiplichino le ingiustizie sociali, codificandole in regole algoritmiche, e rifiutare applicazioni (come il *social scoring*) che neghino il potenziale di cambiamento di un individuo basandosi unicamente sul comportamento passato[52].

Al principio di *Affidabilità* sono poi riconducibili i requisiti dei sistemi di IA non solo, per l'appunto, di *reliability*, ma anche di *robustness*, intesa come capacità tecnica di gestire le incongruenze. Un aspetto critico di inaffidabilità, specialmente nei modelli generativi, è individuato nel fenomeno dell''allucinazione'[53]. Gli sviluppatori devono impegnarsi per la veridicità e l'accuratezza delle informazioni elaborate per prevenire l'affidamento su contenuti distorti o artefatti[54].

Infine, con attenzione al principio di *Sicurezza e privacy* sono state scritte le parti ove si insiste nel sottolinearne la natura imperativa per proteggere non solo l'intimità, ma anche la libertà della persona da controlli indebiti. L'IA, grazie alla sua potenza di analisi, può individuare schemi di comportamento e pensiero anche da dati minimi, accentuando la necessità di rigide e adeguate salvaguardie della riservatezza, che devono essere incluse in quadri normativi. L'uso dell'intelligenza

[50] Cfr. **AN**, nn. 51-52, 76, 80, 116.
[51] Cfr. **AN**, nn. 3, 39, 43-44, 53, 74, 100.
[52] Cfr. **AN**, nn. 46, 75, 94.
[53] Per una spiegazione anche tecnica di questo fenomeno, si rinvia alla nota 88.
[54] Cfr. **AN**, nn. 46, 84, 86.

artificiale per un eccesso di sorveglianza, sfruttamento o limitazione della libertà è eticamente inaccettabile e deve essere tecnicamente prevenuto mediante sistemi di monitoraggio esterni[55].

Risulta in conclusione evidente come la 'questione algoretica' sia stata diffusamente assimilata da questo documento dicasteriale.

D'altro canto, l'ultima esplosione dello sviluppo e dell'uso massivo delle IA è stata accompagnata in anni recenti da una proliferazione di documenti e normative volti a dettare linee guida per una progettazione e un impiego etici di tali sistemi, proprio allo scopo di evitare il loro disallineamento da valori compatibili con il rispetto della dignità umana[56].

Limitandoci a quelli a carattere generale, e procedendo in ordine di tempo, per quanto riguarda gli interventi di organismi vaticani[57] possiamo ricordare innanzitutto il documento *Ethics in the Age of Disruptive Technologies: An Operational Roadmap*, noto anche come "*ITEC Handbook*", che può essere considerato il passo successivo e complementare rispetto alla *Rome Call for AI Ethics*: approfondisce come implementare i principi etici all'interno delle organizzazioni, traducendoli in processi concreti. In pratica si propone come una guida operativa per aiutare organizzazioni, aziende e decisori a gestire in modo etico l'impatto delle tecnologie *disruptive* (ossia 'dirompenti' tra cui l'IA, il machine learning, il tracciamento). Pubblicato nel giugno 2023, è il frutto di una cooperazione tra il Markkula Center for Applied Ethics presso la Santa Clara University e il Centro per la Cultura Digitale del Dicastero per la Cultura e l'Educazione del Vaticano.

Ma soprattutto ricordiamo le recenti *Linee guida in materia di intelligenza artificiale* della Pontificia Commissione per lo Stato della Città

[55] Cfr. **AN**, nn. 3, 46, 90, 92-93.

[56] In base all'*AI Ethics Guidelines Global Inventory* - il cui scopo era di mappare i framework che cercano di stabilire principi su come i sistemi di decisione automatizzata (ADM) possono essere sviluppati e implementati in modo etico -, si contavano già ad aprile 2020, data ultima cui risale l'aggiornamento dell'inventario, ben 173 linee guida (l'inventario in *https://inventory.algorithmwatch.org/*).

[57] Per precedenti iniziative della Pontificia Accademia delle Scienze, della Pontificia Accademia delle Scienze Sociali, della Pontificia Accademia per la Vita, del Pontificio Consiglio della Cultura e del Dicastero per il Servizio dello Sviluppo Umano Integrale rinviamo a **G. TRIDENTE**, *ANIMA DIGITALE*, cit., pp. 83-103.

del Vaticano, entrate in vigore il 1° gennaio 2025[58] ed emanate, come recita l'art. 1, § 1, al fine di recare "principi generali tesi a valorizzare e promuovere un utilizzo etico e trasparente dell'intelligenza artificiale, in una dimensione antropocentrica e affidabile, nel rispetto della dignità umana e del bene comune".

Esse dettano da un lato una serie di principi fondamentali, dall'altro formulano il divieto di diverse pratiche[59], anticipando, sia pure in forma più succinta, quel 'quadro algoretico' che è stato compiutamente delineato un mese dopo nelle linee guida della Nota *Antiqua et nova.*

Questi recenti interventi della Chiesa cattolica si inseriscono comunque, come si è detto, nel solco di analoghe esperienze percorse,

[58] **PONTIFICIA COMMISSIONE PER LO STATO DELLA CITTÀ DEL VATICANO**, Decreto 16 dicembre 2024, n. DCCII - *Linee guida in materia di intelligenza artificiale* (testo integrale nel sito del Governatorato dello Stato della Città del Vaticano *www.vaticanstate.va*). Peraltro, ai sensi dell'art. 1, § 3, esse producono effetti soltanto per il Governatorato dello Stato della Città del Vaticano e per le attività svolte dal Governatorato nelle zone previste dagli articoli 15 e 16 del Trattato Lateranense.

[59] Per un'analisi del loro contenuto, rinviamo a **R. SANTORO**, *Chiesa cattolica,* cit., pp. 58-62, nonché **F. BALSAMO**, *Le Linee guida in materia di intelligenza artificiale per lo Stato della Città del Vaticano del 16 dicembre 2024*, in *Diritto e Religioni*, XX (2025), pp. 1-17.

anche in questi ultimi anni, da organizzazioni internazionali[60], nonché da ordinamenti sovranazionali[61] e nazionali[62], tutte animate dal medesimo

[60] Sempre limitandoci a documenti di carattere generale, possiamo ricordare fra quelli messi a punto negli ultimi anni, in ordine di tempo: 1) *The* OECD *Principles on AI*, adottati dai paesi membri dell'OCSE il 22 maggio 2019, che contiene i primi principi intergovernativi sull'IA (il testo integrale è edito nel sito ufficiale dell'OCSE *www.oecd.org*); 2) I *G20 AI Principles*, adottati come linee guida non vincolanti dai paesi membri il 9 giugno 2019 durante il vertice del G20 a Osaka, con principi sull'IA incentrati sull'essere umano che si basano direttamente sul precedente documento dell'OCSE (testo integrale in *https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf*); 3) *Ethics for Artificial Intelligence. The Recommendation* dell'UNESCO, pubblicato nel novembre 2021 e applicabile a tutti i 194 Stati membri, che ha costituito il primo standard globale sull'etica dell'IA (testo integrale nel sito dell'UNESCO *www.unesco.org*); 4) *The International Guiding Principles for Organizations Developing Advanced AI Systems* e *The International Code of Conduct for Organizations Developing Advanced AI Systems*, adottati il 30 ottobre 2023 durante il vertice ministeriale del G7 su *Tecnologia e Digitale* a Cernobbio: i due documenti, che si presentano espressamente come *living documents*, configurano un codice di condotta volontario per gli sviluppatori di sistemi di IA avanzati, il primo fornendo la base concettuale di 11 principi e il secondo costituendo la loro applicazione pratica sul piano delle azioni concrete (testi integrali nel sito della Commissione Europea *https://digital-strategy.ec.europa.eu/*); 5) La *Bletchley Declaration*, siglata il 1° novembre 2023 dai 28 Paesi (fra cui l'Italia) e dall'Unione europea partecipanti al primo *UK AI Safety Summit* di Bletchley Park, nella quale, riconoscendosi gli enormi benefici potenziali, ma anche i rischi potenzialmente catastrofici dell'IA 'di frontiera', si affermano 4 principi per la sua progettazione, sviluppo e impiego (*safe*, *human-centric, trustworthy, responsible*), sui quali i firmatari si impegnano a collaborare (testo integrale nel sito del Governo britannico *www.gov.uk/*); 6) La Risoluzione A/RES/78/265 approvata dall'Assemblea Generale delle Nazioni Unite il 21 marzo 2024 intitolata *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*, che vuole utilizzare attivamente le capacità dei sistemi di IA - sempre incentrati sull'umanità, etici, spiegabili, responsabili, e nel rispetto del diritto internazionale e della privacy - come strumento per conseguire obiettivi di sviluppo globali cruciali verso l'Agenda 2030 (testo integrale nel sito dell'ONU *https:// digitallibrary.un.org/*); 7) La *Global Call for AI Red Lines* lanciata durante l'80ª Assemblea Generale delle Nazioni Unite il 22 settembre 2025, nella quale, preso atto che l'attuale traiettoria di sviluppo dell'IA comporta pericoli senza precedenti e potrebbe presto superare di molto le capacità umane, amplificando minacce come pandemie ingegnerizzate, disinformazione diffusa, manipolazione delle persone su larga scala - inclusi i minori - rischi per la sicurezza nazionale e internazionale, disoccupazione di massa e violazioni sistematiche dei diritti umani, si fa appello ai governi affinché si raggiunga un accordo internazionale sui limiti per l'IA entro la fine del 2026, garantendone l'effettiva applicazione attraverso solidi meccanismi di controllo e di attuazione (testo integrale in *https://red-lines.ai/*). Per altre esperienze anteriori al 2020, si veda **L. FLORIDI**, *Etica dell'intelligenza artificiale*, Raffaello Cortina Editore, Milano, 2022, pp. 85-98.

[61] Restando nell'Unione Europea e nell'ambito di documenti a carattere generale, rammentiamo innanzitutto tre testi (soprattutto l'ultimo dei quali ha costituito la base della successiva normativa): 1) Gli *Orientamenti etici per un'IA affidabile*, documento reso pubblico l'8 aprile 2019 e redatto dal Gruppo Indipendente di Esperti ad Alto Livello sull'Intelligenza Artificiale, istituito dalla Commissione europea, per promuovere un'IA affidabile, garantendo che sia al contempo legale (rispettosa delle leggi), etica (aderente ai principi) e robusta (sicura tecnicamente e socialmente), che traduce questi obiettivi in sette requisiti fondamentali e fornisce una lista di controllo per la loro valutazione pratica durante l'intero ciclo di vita del sistema (testo integrale in *https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai*); 2) Il *LIBRO BIANCO sull'intelligenza artificiale - Un approccio europeo all'eccellenza e alla fiducia*, pubblicato dalla Commissione Europea il 19 febbraio 2020, che propone un approccio strategico basato sul rischio, imponendo prescrizioni obbligatorie - come robustezza, trasparenza, sorveglianza umana - solo per le applicazioni di IA ad alto rischio (testo integrale nel sito della Commissione Europea *https://commission.europa.eu/*); 3) Le *Raccomandazioni alla Commissione concernenti il quadro relativo agli aspetti etici dell'intelligenza artificiale, della robotica e delle tecnologie correlate* a opera del Parlamento Europeo (per gli ampi contenuti si veda la relazione, datata 8 ottobre 2020, documento A9-0186/2020, edita nel sito del Parlamento *www.europarl.europa.eu*): la proposta mirava all'istituzione di un quadro normativo europeo armonizzato, completo e adeguato alle esigenze future per l'IA, basato sul diritto e sui valori fondamentali dell'Unione Europea, con l'indicazione di vari principi, ispirati all'algoretica e a una visione di un'intelligenza artificiale antropocentrica, in vista della presentazione di un atto legislativo futuro. Esso si è concretizzato nel *REGOLAMENTO (UE) 2024/1689 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale*, meglio noto come *AI Act UE*, che ha dato corpo a quel quadro normativo e che mira a promuovere uno sviluppo e un utilizzo dell'IA sicuri, etici, trasparenti e rispettosi dei diritti fondamentali, proteggendo i cittadini europei dai rischi legati all'IA, affermando vari principi etici e giuridici (testo integrale in *https://artificialintelligenceact.eu/*). Più recentemente, va segnalato *The General-Purpose AI Code of Practice*, presentato alla Commissione europea in versione definitiva il 10 luglio 2025: si tratta di uno strumento volontario, sviluppato da 13 esperti indipendenti, con il contributo di oltre mille parti interessate, tra cui fornitori di modelli, piccole e medie imprese, accademici, esperti di sicurezza dell'IA, titolari di diritti e organizzazioni della società civile. Esso è concepito per aiutare l'industria a conformarsi alle norme dell'*AI Act UE* relative all'IA per uso generico, per garantire che i relativi modelli immessi sul mercato europeo, compresi quelli più potenti, siano sicuri e trasparenti. Il codice è composto da tre capitoli: il primo è relativo ai temi della trasparenza e del diritto d'autore, che riguardano tutti i fornitori di modelli di IA per uso generico, mentre il secondo e il terzo sono relativi a sicurezza e protezione, che riguardano solo un numero limitato di fornitori dei modelli più avanzati (testo in *https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai*). Va detto che l'Unione Europea è determinata ad affermarsi come leader mondiale nel campo dell'intelligenza artificiale, secondo quanto si legge sul sito ove è presentato *The AI Continent Action Plan*, pubblicato il 9 aprile 2025 (*https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan*). Invece, per un

veloce sguardo alle iniziative della realtà del continente asiatico, si veda **G. TRIDENTE**, *ANIMA DIGITALE*, cit., pp. 39-40.

[62] Con riguardo specifico all'Italia, possiamo menzionare tra le linee guida a carattere generale: 1) *L'Intelligenza Artificiale al servizio del cittadino*, un Libro Bianco presentato il 21 marzo 2018 a Roma, a cura della task force IA dell'Agenzia per l'Italia digitale, indirizzato alle amministrazioni pubbliche, che definisce linee guida e raccomandazioni per un uso "sostenibile e responsabile" dell'IA nel rispetto dei diritti dei cittadini, con focus su trasparenza, centralità dell'utente e legalità (*https://ia.italia.it/assets/libro bianco.pdf*); 2) Le *Proposte per una Strategia italiana per l'Intelligenza Artificiale*, rese pubbliche il 20 luglio 2020 dal Ministero dello Sviluppo Economico, un documento strategico di policy nazionale che imposta la visione italiana come "antropocentrica e orientata verso lo sviluppo sostenibile", delineando principi guida per massimizzare benefici e minimizzare rischi sociali dell'IA (*https://giurisprudenza.unimc.it/it/ricerca/ dirittoapplicato/site-news/Strategia_italiana_AI. pdf*); 3) Il documento *Strategia Italiana per l'Intelligenza Artificiale 2024-2026*, predisposto dall'Agenzia per l'Italia digitale e dal Dipartimento per la trasformazione digitale, che sintetizza visione e architettura di tale strategia per il prossimo triennio, attraverso i suoi pilastri principali (ricerca scientifica, pubblica amministrazione, imprese e formazione), mirando a costruire un ecosistema in cui l'IA sia al servizio delle persone, favorendo i principi etici e di responsabilità sociale e salvaguardando fattori chiave quali la privacy, la sicurezza, le questioni di genere e la sostenibilità ambientale (*https://www. agid.gov.it/sites/agid/files/ 2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-202 6.pdf*). Invece, a livello normativo, ricordiamo la legge 23 settembre 2025, n. 132, contenente *Disposizioni e deleghe al Governo in materia di intelligenza artificiale*, prima legge nazionale organica che disciplina sviluppo, adozione e governance dei sistemi di IA in Italia, in coerenza con l'*AI Act UE*.

obiettivo di sviluppare e impiegare sistemi di IA governati da principi etici[63]. Per questo esse potrebbero costituire un'importante base per dar vita a un "trattato internazionale vincolante, che regoli lo sviluppo e l'uso dell'intelligenza artificiale nelle sue molteplici forme", sul quale Papa Francesco ha più volte insistito[64]. Anche recentemente Papa Leone XIV ha indicato la via "affinché questi strumenti possano veramente essere da noi integrati come alleati" attraverso "tre pilastri" - "responsabilità, cooperazione ed educazione" -, rinnovando, con riguardo al primo,

[63] Rimarca la proliferazione dei principi etici, con il rischio di confusione e anche di una serie di altri potenziali abusi, **L. FLORIDI**, *Etica dell'intelligenza,* cit., pp. 85-98, il quale invita a stabilire standard etici chiari, condivisi e pubblicamente accettati. Dallo studio di sei influenti iniziative etiche del 2017 e 2018 e dalla loro convergenza su alcuni principi, esce un quadro unificato di cinque principi etici fondamentali, i primi quattro dei quali tradizionalmente appartenenti alla bioetica: *Beneficenza* (promuovere il benessere, preservare la dignità e sostenere il pianeta), *Non maleficenza* (privacy, sicurezza e "cautela della capacità"), *Autonomia* (mantenere il potere di "decidere di decidere"), *Giustizia* (promuovere la prosperità ed evitare l'iniquità) ed *Esplicabilità* (intesa sia come intelligibilità, il come funziona, sia come responsabilità, il chi è responsabile). Un altro interessante studio di **A. JOBIN, M. IENCA et al.**, *The Global Landscape of AI Ethics Guidelines,* in *Nature Machine Intelligence*, I (2019), pp. 389-399, ha analizzato 84 documenti prodotti da vari organismi, principalmente fra il 2016 e il 2019, identificando undici valori e principi etici ricorrenti, in ordine di frequenza: Trasparenza, Giustizia ed equità, Non-maleficenza, Responsabilità, Privacy, Beneficenza, Libertà e autonomia, Fiducia, Sostenibilità, Dignità e Solidarietà. I primi cinque principi risultano citati in più della metà di tutte le fonti, evidenziando così gli autori un consenso unanime sul loro carattere fondamentale, sebbene vengano variamente interpretati. Sulla difficoltà a rinvenire "un'unica gerarchia di valori" e la necessità di "trovare dei principi condivisi" si veda anche **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 4.

[64] "L'obiettivo della regolamentazione, naturalmente, non dovrebbe essere solo la prevenzione delle cattive pratiche, ma anche l'incoraggiamento delle buone pratiche, stimolando approcci nuovi e creativi e facilitando iniziative personali e collettive. In definitiva, nella ricerca di modelli normativi che possano fornire una guida etica agli sviluppatori di tecnologie digitali, è indispensabile identificare i valori umani che dovrebbero essere alla base dell'impegno della società per formulare, adottare e applicare necessari quadri legislativi. Il lavoro di redazione di linee guida etiche per la produzione di forme di intelligenza artificiale non può prescindere dalla considerazione di questioni più profonde riguardanti il significato dell'esistenza umana, la tutela dei diritti umani fondamentali, il perseguimento della giustizia e della pace": **FRANCESCO**, *Messaggio per la LVII Giornata*, cit., p. 9 s.; l'appello viene rinnovato in **ID.**, *Messaggio per la LVIII Giornata*, cit., p. 4.

l'invito ai legislatori nazionali e ai regolatori sovranazionali a una "regolamentazione adeguata" rispettosa della dignità umana[65].

L'ultimo intervento in ordine di tempo, estremamente significativo, è stato il primo *AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, commissionato dai paesi partecipanti al citato vertice sulla sicurezza dell'IA tenutosi a Bletchley (UK). Trenta paesi, l'ONU, l'OCSE e l'UE hanno nominato ciascuno un rappresentante per il comitato consultivo, composto da quasi un centinaio di esperti di IA in rappresentanza di diverse prospettive e discipline, sotto la guida di Yoshua Bengio, docente nel *Department of Computer Science and Operations Research* presso l'Université de Montréal. Lo scopo del loro lavoro è aiutare i decisori a garantire che le persone in tutto il mondo possano beneficiare dei vantaggi dell'IA in modo sicuro, fornendo le informazioni scientifiche che identificano i rischi e valutando i metodi tecnici per la loro mitigazione[66].

Il documento consente di distinguere la differenza fra la nozione di guardrail e quella di safeguard. I primi vengono definiti nel Glossario come **"**pre-defined safety constraints or boundaries set up in an attempt to ensure an AI system operates within desired parameters and avoids unintended or harmful outcomes", mentre con i secondi il rapporto si riferisce più spesso agli specifici meccanismi tecnici e agli strati di difesa implementati all'interno o intorno a un sistema di IA avanzata per far

[65] Cfr. **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 5. La necessità di una simile alleanza scaturisce da un'analisi dei rischi che in questo documento magisteriale appare particolarmente accurata: indebolimento della capacità di ascolto e di pensiero critico; logoramento della capacità di pensare in modo analitico e creativo, di comprendere i significati, di distinguere tra sintassi e semantica; erosione delle capacità cognitive, emotive e comunicative; rinuncia al processo creativo e cessione alle macchine delle funzioni mentali e dell'immaginazione, con perdita dei talenti ricevuti al fine di crescere come persone in relazione a Dio e agli altri. Un richiamo a questi rischi si rinviene anche in **QV**, n. 46, peraltro "rischi mai immaginati prima" (*ivi*, n. 159).

[66] Il report è disponibile in *https://iamaeg.net/files/B00BDA99-2DBF-4EDF-B9BA-E7801CA3891D.pdf*. Esso è stato pubblicato nel gennaio 2025, ma si configura come un work in progress: un primo aggiornamento, intitolato *International AI Safety Report. First Key Update. Capabilities and Risk Implications*, si è reso necessario nel mese di ottobre 2025, constatandosi la rapidità con cui evolve il campo dell'intelligenza artificiale. Esso si è focalizzato in particolare sui miglioramenti nelle capacità di ragionamento e nell'autonomia dei sistemi di IA e sulle conseguenti implicazioni per la sicurezza e la supervisione. L'aggiornamento è disponibile in *https://internationalaisafetyreport.org/sites/default/files/2025-10/first-key-update_0.pdf*.

rispettare i guardrail stabiliti e mitigare i rischi. In sostanza, i guardrail definiscono *cosa* non deve succedere; i safeguard sono gli strumenti che impediscono, monitorano e bloccano attivamente quel *cosa*. In altri termini, i guardrail sono i confini stabiliti per l'uso sicuro dell'IA, mentre i safeguard sono i dispositivi attivi e le procedure tecniche integrate nel sistema per farli rispettare: stabilire tali confini e ispirarne le implementazioni tecniche costituisce il compito proprio dell'algoretica. Essa quindi, come spiega Papa Francesco, mira a operare come fonte di "ispirazione etica" e "forma di moderazione etica degli algoritmi e dei programmi di intelligenza artificiale"[67].

Inoltre, recentemente è stato rimarcato proprio da padre Benanti come la questione etica stia diventando più urgente e complessa in

[67] Cfr. **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 4.

relazione a una nuova frontiera, quella della cosiddetta intelligenza artificiale 'agentica', sempre più diffusa[68]: si tratta di

> "un sistema software autonomo basato sull'intelligenza artificiale, capace di percepire l'ambiente, elaborare informazioni, prendere decisioni e agire per raggiungere obiettivi specifici nel mondo reale senza intervento umano costante. A differenza di un semplice chatbot, un agente Ai ragiona, pianifica sequenze di azioni complesse, apprende dall'esperienza e utilizza strumenti esterni come Api o database. Gli agenti Ai integrano percezione (tramite sensori o input dati), memoria per ricordare contesti passati, ragionamento per valutare opzioni e attuatori per eseguire compiti reali, come automatizzare workflow aziendali o gestire lead.

[68] Le IA vengono classificate in: 1) *predittive*, quando analizzano dati storici e attuali per prevedere eventi futuri, come tendenze di mercato o comportamenti degli utenti, utilizzando modelli statistici e algoritmi di machine learning; 2) *generative*, quando creano contenuti originali, come testi, immagini o musica, imitando stili e strutture esistenti grazie a modelli come le reti neurali generative; 3) *agentiche*, quando sono progettate per agire in modo autonomo, prendere decisioni e svolgere attività complesse verso un obiettivo specifico, con supervisione umana minima; 4) *percettive*, quando interpretano dati sensoriali (immagini, suoni, eccetera) per comprendere l'ambiente circostante, come nei sistemi di riconoscimento facciale o nei veicoli autonomi; 5) *simboliche*, quando risolvono problemi usando rappresentazioni logiche e regole, come nei sistemi esperti che supportano decisioni in ambiti specialistici; 6) *conversazionali*, quando comprendono e generano linguaggio naturale per interagire con gli utenti, come nei chatbot e assistenti virtuali, utilizzando tecniche di elaborazione del linguaggio naturale (*Natural Language Processing*, NLP). In questa sede, quando parleremo di IA, si farà riferimento a quella generativa e conversazionale basata su modelli LLM (*Large Language Model*) accessibili tramite chatbot, come ChatGPT di OpenAI. Un LLM è un tipo avanzato di intelligenza artificiale specializzato nello svolgere compiti come la traduzione, la classificazione, la comprensione, la generazione di testo e la risposta a domande in modo conversazionale, rendendo possibile una comunicazione che approssima quella umana. Questi modelli utilizzano tecniche di machine learning, in particolare il deep learning, e sono addestrati su enormi quantità di dati testuali per riconoscere schemi complessi nel linguaggio naturale rilevandone le regolarità statistiche. Sono basati su una particolare architettura di rete neurale chiamata *Transformer* - da cui i sistemi *Generative Pre-trained Transformer*, GPT -, che permette di comprendere le relazioni tra parole e frasi in un contesto ampio e flessibile, rispondendo anche a input non predefiniti, senza bisogno di istruzioni rigide. Per un'analisi più dettagliata delle modalità di funzionamento, si rinvia all'ultimo libro di **P. BENANTI**, *L'uomo è un algoritmo. Il senso dell'umano e l'intelligenza artificiale*, Lit Edizioni, Roma, 2025, pp. 32-41, nonché a **E. FALETTI, G. TROVATO**, *Diritto canonico e machine learning: il caso del robot SanTO*, in *Stato, Chiese e pluralismo confessionale*, cit., n. 11 del 2025, pp. 158-160.

> Possono collaborare tra loro per processi complessi e si adattano dinamicamente a cambiamenti ambientali".

Poiché in presenza di flussi di lavoro complessi, composti da molti passaggi sequenziali, i piccoli errori tendono ad accumularsi, "la loro affidabilità precipita rapidamente all'aumentare della durata del task"; si verifica cioè la "trappola dell'errore composto" che "trasforma strumenti promettenti in sistemi inaffidabili appena la complessità aumenta"[69].

Alla luce di queste criticità, assai pertinente e attuale appare l'ultimo richiamo di Papa Leone XIV, il quale, a proposito del secondo pilastro della cooperazione, dopo aver ricordato che nessun settore "può affrontare da solo la sfida di guidare l'innovazione digitale e la *governance* dell'IA", dichiara: "È necessario perciò creare meccanismi di salvaguardia"[70].

Risulta a questo punto evidente che la costruzione di guardrail e relativi safeguard, tema sul quale si sono notevolmente concentrati studi

[69] Cfr. **P. BENANTI**, *La trappola matematica dell'autonomia AI e la necessità dell'umano*, in *Il Sole 24 Ore*, 10 dicembre 2025, p. 18, il quale conclude sottolineando la necessità di mantenere l'uomo saldamente nel ciclo decisionale. Quando poi le intelligenze artificiali agentiche vengono applicate nel «contesto militare, questo significa passare da un assistente digitale che analizza il campo di battaglia a un "agente" capace di intraprendere flussi di lavoro operativi» e allora "si aprono le sfide etiche più vertiginose del nostro tempo" che richiederanno "una vigilanza etica senza precedenti": cfr. **P. BENANTI**, *La guerra algoritmica, un salto evolutivo dai laboratori all'azione*, in *Il Sole 24 Ore*, 3 dicembre 2025, p. 16. D'altronde, lo si è visto, già **AN**, n. 100, rimarcava come «i sistemi di armi autonome e letali, in grado di identificare e colpire obiettivi senza intervento umano diretto, sono "grave motivo di preoccupazione etica", poiché essi mancano della "esclusiva capacità umana di giudizio morale e di decisione etica"».

[70] Cfr. **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 6.

accademici recenti in lingua inglese[71], non è un'operazione eticamente neutrale e meramente tecnica, dal momento che "ogni scelta progettuale

[71] Tra i contributi più significativi, e soltanto relativamente al problema in generale senza entrare in domini specifici (come ad esempio, quelli di salute, educazione e finanza), si vedano: **Y. DONG, R. MU et al.,** *Safeguarding Large Language Models: A Survey*, in *Artificial Intelligence Review*, XL (2025), Numero Articolo 382; **A. MULAHUWAISH, M. EL-KHOURY et al.,** *Does AI Need Guardrails?*, in *International Journal of Pervasive Computing and Communications*, XXI (2025), pp. 177-186; **E. PAPAGIANNIDIS, P. MIKALEF, K. CONBOY**, *Responsible Artificial Intelligence Governance: A Review and Research Framework*, in *The Journal of Strategic Information Systems*, XXXIV (2025), Numero Articolo 101885; **M. SHAMSUJJOHA, Q. LU et al.,***Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model-Based Agents*, in *2025 IEEE 22nd International Conference on Software Architecture (ICSA)*, IEEE, Piscataway, 2025, pp. 37-48; **D. WILLIAMS-KING, L. LE et al.**, *Can Safety Fine-Tuning Be More Principled? Lessons Learned from Cybersecurity*, in *arXiv preprint*, *arXiv:2501.11183*, 2025; *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)*, a cura di S. DAS, B.P. GREEN et al., AAAI Press, Washington, 2024; **S.S. ARSLAN**, *Artificial Human Intelligence: The role of Humans in the Development of Next Generation AI*, in *arXiv preprint*, *arXiv:2409.16001*, 2024; **S.G. AYYAMPERUMAL, L. GE**, *Current State of LLM Risks and AI Guardrails*, in *arXiv preprint*, *arXiv:2406.12934*, 2024; **Y. BENGIO**, *Government Interventions to Avert Future Catastrophic AI Risks*, in *Harvard Data Science Review*, Special Issue 5 (2024) (*https://hdsr.mitpress.mit.edu/pub/w974bwb0/release/2*); **Y. BENGIO, M.K. COHEN et al**., *Can a Bayesian Oracle Prevent Harm from an Agent?* in *arXiv preprint*, *arXiv:2408.05284*, 2024; **Y. BENGIO, G. HINTON et al.,** *Managing Extreme AI Risks Amid Rapid Progress*, in *Science*, CCCLXXXIV (2024), pp. 842-845; **P. BREY, B. DAINOW**, *Ethics by Design for Artificial Intelligence* in *AI and Ethics*, IV (2024), pp. 1265-1277; **D. DALRYMPLE, J. SKALSE et al.,** *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*, in *arXiv preprint*, *arXiv:2405.06624*, 2024; **Y. DONG, R. MU et al.,** *Building Guardrails for Large Language Models*, in *arXiv preprint*, *arXiv:2402.01822*, 2024; **J. HU, Y. DONG**, *Trust-Oriented Adaptive Guardrails for Large Language Models*, in *arXiv preprint*, *arXiv:2408.08959*, 2024; **W. MURIKAH, J. K. NTHENGE, F.M. MUSYOKA**, *Bias and Ethics of AI Systems Applied in Auditing-A systematic review*, in *Scientific African*, XXV (2024), Numero Articolo e02281; **K. ŠEKRST, J. MCHUGH, J. CEFALU**, *AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development*, in *arXiv preprint*, *arXiv:2411.14442*, 2024; **N. DÍAZ-RODRÍGUEZ, J. DEL SER et al.**, *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*, in *Information Fusion*, XCIX (2023), Numero Articolo 101896; **M. LI, A. ENKHTUR et al.,** *Ethical Implications of ChatGPT in Higher Education: A Scoping Review*, in *arXiv preprint*, *arXiv:2311.14378*, 2023; **M. TAHAEI, M. CONSTANTINIDES et al.,** *Human-Centered Responsible Artificial Intelligence: Current & Future Trends*, in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*, a cura di A. SCHMIDT, K. VÄÄNÄNEN et al., Association for Computing Machinery, New York, 2023, pp. 1-4**.**

esprime una visione dell'umanità", come ricorda Papa Leone XIV[72]. Poiché si tratta di innestare un modello valoriale all'interno di un sistema di IA, le scelte progettuali implicano sempre delle previe decisioni algoretiche e quindi le prescrizioni dell'algoretica, che incidono su quelle scelte[73], sono indirizzate prima di tutto agli uomini che implementano gli algoritmi:

> "La Chiesa, pertanto, invita tutti i costruttori di IA a coltivare il discernimento morale come parte fondamentale del loro lavoro, a sviluppare sistemi che rispecchino giustizia, solidarietà e un rispetto autentico per la vita. […] Deve essere un'impresa profondamente ecclesiale. Che disegni algoritmi per l'educazione cattolica, strumenti per la cura sanitaria compassionevole o piattaforme creative che narrano la storia cristiana con verità e bellezza, ogni partecipante contribuisce a una missione condivisa:

[72] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti del Builders AI Forum*, cit., p. 1; nello stesso senso **ID.**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 4. Sulla considerazione che "nessuna innovazione è neutrale" e la tecnologia "include sempre, in una maniera più o meno esplicita, la visione del mondo di chi l'ha realizzata e sviluppata" si veda anche **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 4, nonché, nello stesso senso, **ID.**, *Messaggio per la LVII Giornata Mondiale della Pace*, cit., pp. 3,5; **ID.**, *Messaggio per la LVIII Giornata*, cit., p. 4. Il tema, come si è visto, è trattato anche nella sezione IV di **AN**, n. 39, laddove si rammenta che nel campo dell'IA "la dimensione etica assume primaria importanza poiché sono le persone a progettare i sistemi e a determinare per quali scopi vengano usati". Invece, specificamente sulla non neutralità dei mezzi di comunicazione sociale (radio, televisione, internet, social media) si veda **QV**, n. 42.

[73] Dichiara nell'Introduzione la *Rome Call for AI Ethics*: "In the long term, the values and principles that we are able to instill in AI will help to establish a framework that regulates and acts as a point of reference for digital ethics, guiding our actions and promoting the use of technology to benefit humanity and the environment".

mettere la tecnologia al servizio dell'evangelizzazione e dello sviluppo integrale di ogni persona"[74].

A quanto esposto occorre ora affiancare due considerazioni che ne completano il quadro.

La prima, di ordine dottrinale, è che le 'implementazioni etiche' che l'algoretica intende introdurre nei sistemi di IA non trasformano questi ultimi in *veri agenti morali*, qualifica che, come precisa la Nota

[74] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti del Builders AI Forum*, cit., p. 1 s. Ancora più recentemente, l'invito è rinnovato in **ID.**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 5 e, da ultimo, in **ID.**, *Message of Pope Leo XIV*, cit., p. 1, ove si legge: "A questo proposito, un ambito di ricerca particolarmente fecondo è l'uso degli algoritmi, in particolare nel campo dell'intelligenza artificiale. Un tale compito richiede non solo impegno intellettuale e ingegno, ma una crescita integrale di tutta la persona, per abbracciare la dimensione morale di queste tecnologie emergenti". D'altronde, si era molto opportunamente evidenziato in precedenza da Papa Francesco che non "è sufficiente nemmeno presumere, da parte di chi progetta algoritmi e tecnologie digitali, un impegno ad agire in modo etico e responsabile. Occorre rafforzare o, se necessario, istituire organismi incaricati di esaminare le questioni etiche emergenti e di tutelare i diritti di quanti utilizzano forme di intelligenza artificiale o ne sono influenzati. L'immensa espansione della tecnologia deve quindi essere accompagnata da un'adeguata formazione alla responsabilità per il suo sviluppo. […] Abbiamo perciò il dovere di allargare lo sguardo e di orientare la ricerca tecnico-scientifica al perseguimento della pace e del bene comune, al servizio dello sviluppo integrale dell'uomo e della comunità" (**FRANCESCO**, *Messaggio per la LVII Giornata*, cit., p. 3 s.). Il valore in gioco non potrebbe essere più fondamentale: "Abbiamo bisogno di garantire e tutelare uno spazio di controllo significativo dell'essere umano sul processo di scelta dei programmi di intelligenza artificiale: ne va della stessa dignità umana" (**ID.**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 2). Anche nel citato Manifesto di Vienna, leggiamo: "A vision is needed for new educational curricula, combining knowledge from the humanities, the social sciences, and engineering studies. In the age of automated decision making and AI, creativity and attention to human aspects are crucial to the education of future engineers and technologists".

Antiqua et nova[75] e ribadisce il magistero pontificio, anche recente[76], spetta esclusivamente agli esseri umani.

La seconda, che discende direttamente dal ragionamento sin qui condotto, è che gli effetti di simili implementazioni si producono sul piano concreto delle condotte delle intelligenze artificiali operanti nel mondo, per allinearle a valori umani (e cristiani), per contenerne i comportamenti entro limiti conformi a tali valori[77]. Quegli effetti operano cioè, nell'interazione quotidiana uomo-IA, sul versante della macchina-agente.

Quindi, per quanto possa risultare rigorosa ed efficace, l'algoretica opera pressoché esclusivamente sul piano della progettazione e del funzionamento del sistema. Essa non può raggiungere, per sua stessa natura, il foro interno dell'utente: la sua libertà, le sue scelte, il suo modo di rapportarsi alla macchina. Resta dunque aperto il versante della responsabilità morale soggettiva dell'uomo che interagisce con l'IA, un versante che richiede un guardrail proprio: per contribuire alla

[75] Cfr. **AN**, n. 111: "Solo la persona umana può dirsi moralmente responsabile, e le sfide di una società tecnologica riguardano in ultima analisi il suo spirito". Lo stesso documento chiarisce infatti: «Tra una macchina e un essere umano, solo quest'ultimo è veramente un agente morale, cioè un soggetto moralmente responsabile che esercita la sua libertà nelle proprie decisioni e ne accetta le conseguenze; solo gli esseri umani sono in relazione con la verità e il bene, guidati dalla coscienza morale che li chiama "ad amare, a fare il bene e a fuggire il male", attestando "l'autorità della verità in riferimento al Bene supremo, di cui la persona umana avverte l'attrattiva"; solo gli esseri umani possono essere sufficientemente consapevoli di sé al punto da riuscire ad ascoltare e seguire la voce della coscienza, discernendo con prudenza e ricercando il bene possibile in ogni situazione. Di fatto, anche questo appartiene all'esercizio dell'intelligenza da parte della persona» (*ivi*, n. 39).

[76] Cfr. **LEONE XIV**, *Messaggio del Santo Padre Leone XIV, a firma del Cardinale Segretario di Stato*, cit., p. 2; il medesimo concetto era stato chiarito in due discorsi anteriori da **FRANCESCO**, *Messaggio per la LVII Giornata*, cit., p. 7; **ID.**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 2. In prospettiva laica, nella medesima direzione, si vedano **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 92.

[77] Rileva puntualmente **L. FLORIDI**, *La differenza fondamentale. Artificial Agency: una nuova filosofia*, Mondadori, Milano, 2025, p. 142 s.: «Nonostante le sue sofisticate capacità, l'agency artificiale è fondamentalmente limitata dalla sua progettazione e dalla sua struttura operativa. L'autonomia limitata dei sistemi di IA circoscrive le loro operazioni all'interno di parametri programmati, creando un chiaro limite all'azione indipendente. […] Questo limite si estende all'incapacità di generare scopi o obiettivi veramente originali, scelti o preferiti, compresa la capacità di scegliere se scegliere, con il risultato di costringere gli agenti artificiali a operare entro schemi prestabiliti. Il confine del "capire" resta fermo al livello del riconoscimento e della corrispondenza di pattern, senza accedere al livello della vera e propria comprensione».

costruzione di un mondo in cui gli uomini 'coesistono' con le AI continuando a 'restare pienamente umani'.

3 - L'antropomorfismo delle IA: un pericolo non indifferente

Per comprendere appieno le ragioni dell'asserita insufficienza, 'in difesa dell'umano', della sola algoretica occorre peraltro prima affrontare l'argomento, "affascinante e tremendo"[78], dell'antropomorfismo delle IA.

Il vocabolario Treccani, alla voce *Antropomorfismo*, offre la seguente definizione:

> "s. m. [der. di antropomorfo]. - Tendenza ad attribuire aspetto, facoltà e destini umani a figure immaginarie, animali e cose, presente pressoché universalmente tra i popoli primitivi e nel folclore e nel pensiero dei popoli civili. In partic., attribuzione alla divinità di qualità umane, sia fisiche (a. fisico) sia intellettuali e morali (a. psichico, detto anche antropopatia)"[79].

Si ritiene che si tratti di una tendenza innata della psicologia dell'uomo[80]. L'uomo vi fa ricorso per varie ragioni: a fini artistici e letterari (talvolta a scopo educativo-morale, come avviene nelle fiabe di Fedro e di Esopo, talvolta a scopo di satira politica, come nel libro *La*

[78] Questi due aggettivi sono stati usati da Papa Francesco nel titolo del suo discorso al G7 del 2024 con riferimento in generale alle intelligenze artificiali, ma si addicono particolarmente all'aspetto antropomorfo delle IA stesse, per le ragioni che verranno illustrate.

[79] **TRECCANI**, *Vocabolario online*, Voce *Antropomorfismo,* disponibile in (*https://www.treccani.it/vocabolario/antropomorfismo/,* consultato il 12 febbraio 2026).

[80] Tuttavia l'antropomorfismo non è invariante, poiché tanto fattori situazionali quanto individuali possono aumentare o diminuire questa tendenza. Secondo **N. EPLEY, A. WAYTZ, J.T. CACIOPPO**, *On Seeing Human: A Three-Factor Theory of Anthropomorphism*, in *Psychological Review*, CXIV (2007), pp. 864-886, le persone sono più propense ad antropomorfizzare quando la conoscenza antropocentrica è accessibile e applicabile, quando sono motivate a essere agenti sociali efficaci e quando mancano di un senso di connessione sociale con altri esseri umani. Inoltre **K. LETHEREN, K. KUHN et al.**, *Individual Difference Factors Related to Anthropomorphic Tendency*, in *European Journal of Marketing*, L (2016), pp. 973-1002, hanno constatato come la 'tendenza antropomorfica' è influenzata anche dalla personalità, dall'età, dallo stato relazionale, dal rapporto affettivo con gli animali e dal pensiero esperienziale, rendendo i consumatori più giovani, single, creativi e coscienziosi i target ideali per i messaggi antropomorfici.

fattoria degli animali di George Orwell), a fini religiosi (si pensi a quelle tradizioni spirituali che personificano elementi della natura) oppure a fini psicologico-consolatori, per lenire e contenere ansie di vario genere, ad esempio da solitudine o abbandono (basti pensare al memorabile pallone da calcio Wilson del film *Cast Away* di Robert Zemeckis, cui il naufrago Chuck Noland, impersonato dall'attore Tom Hanks, per non impazzire disegna un volto e applica dei finti capelli, oppure, più semplicemente, all'orsacchiotto di peluche o alla bambola di pezza, per anni compagno inseparabile degli infanti).

Racconti e film di fantascienza ci hanno ormai da molto tempo abituati all'attribuzione di qualità e caratteristiche umane a macchine e software. Ha rappresentato una pietra miliare del cinema il computer di bordo HAL 9000 del film *2001: A Space Odyssey* di Stanley Kubrick; in quel caso, evidentemente a motivo della lunga durata e dell'isolamento della spedizione spaziale, la scelta di progettarlo come antropomorfo, nelle interazioni con i due membri dell'equipaggio che viaggiavano non in stato di ibernazione, svolgeva un ruolo importante nel ridurne lo stress, la noia e la solitudine.

Oggi i chatbot dei LLM sono progettati espressamente, come già rilevato, per agire in modo antropomorfo attraverso la conversazione linguistica (si parla di '*conversational AI*', comunemente abbreviata in CAI), come se le intelligenze artificiali fossero umane[81]. Il che ha aperto

[81] "At the most basic level, systems are anthropomorphic if they (i) are interactive, (ii) use language, and (iii) take on a role performed by a human": **G. ABERCROMBIE, A.C. CURRY et al.**, *Mirage: On Anthropomorphism in Dialogue Systems*, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, a cura di H. BOUAMOR, J. PINO, K. BALI, Association for Computational Linguistics, Singapore, 2023, p. 4778.

la strada a una serie di domande surreali e paradossali. L'IA è cosciente[82]? L'IA è intelligente[83]?

La risposta della riflessione ecclesiale, in particolare nella Nota *Antiqua et nova*, è inequivocabile nel negare all'IA tanto la coscienza quanto un'intelligenza in senso proprio e integrale. Sotto il profilo della coscienza, ci viene ricordato:

> "Solo gli esseri umani possono essere sufficientemente consapevoli di sé al punto da riuscire ad ascoltare e seguire la voce della

[82] Nel dibattito filosofico-scientifico contemporaneo, va menzionato che, secondo alcune impostazioni funzionaliste, la coscienza dipende esclusivamente dalla manipolazione delle informazioni da parte di un algoritmo, indipendentemente dal fatto che il sistema che esegue questi calcoli sia costituito da neuroni, silicio o qualsiasi altro substrato fisico. Sul tema di IA e coscienza, nella più recente letteratura tecnico-scientifica di lingua inglese, con attenzione anche ai profili di rischi etici significativi nel considerare l'intelligenza artificiale un'entità cosciente, si vedano: **Y. BENGIO, E. ELMOZNINO**, *Illusions of AI Consciousness*, in *Science*, CCCLXXXIX (2025), pp. 1090-1091; **A. ELAMRANI**, *Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenge*, in *arXiv preprint*, *arXiv:2503.05823*, 2025; **P. BUTLIN, R. LONG et al.,** *Identifying Indicators of Consciousness in AI Systems*, in *Trends in Cognitive Sciences* (2025), pp. 1-14 (*https://doi.org/10.1016/j.tics.2025.10.011*); **I. FERNANDEZ, N. KYOSOVSKA et al.**, *AI Consciousness and Public Perceptions: Four Futures*, in *arXiv preprint*, *arXiv:2408.04771*, 2024; **P. BUTLIN, R. LONG et al.,** *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, in *arXiv preprint*, *arXiv:2308.08708*, 2023. In ogni caso, a essere esplicitamente critico verso l'idea di una 'intelligenza artificiale forte', cioè di macchine davvero coscienti, è colui che è unanimemente riconosciuto il padre del microprocessore, Federico Faggin, che, nel proporre un nuovo modello di comprensione della realtà, esclude categoricamente che il computer potrà mai essere cosciente: "I sistemi classici, come il computer, usano proprietà statistiche di atomi e molecole che sono deterministiche, e perciò non possono essere coscienti né avere libero arbitrio. […] Significato, comprensione e decisioni libere non esistono in un computer, perché queste sono capacità connesse con la natura della coscienza […]": **F. FAGGIN,** *Irriducibile. La coscienza, la vita, i computer e la nostra natura*, Mondadori, Milano, 2023, pp. 67,167. Sulla questione, per una prospettiva fortemente critica che conclude considerando l'idea di macchine coscienti come un mito del transumanesimo e della cultura fantascientifica, si veda **E.C. GARRIDO-MERCHÁN**, *Machine Consciousness as Pseudoscience: The Myth of Conscious Machines*, in *arXiv preprint*, *arXiv:2405.07340*, 2024.

[83] Se si accoglie una nozione non antropica di intelligenza, come quella proposta da Nello Cristianini, docente di *Artificial Intelligence e Machine Learning* presso l'University of Bath (UK), allora la risposta è senz'altro affermativa: "Per semplicità, definiremo l'intelligenza in termini di comportamento di un agente, ovvero di qualsiasi sistema in grado di agire nel suo ambiente, usando informazioni sensoriali per prendere decisioni": **N. CRISTIANINI**, *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*, il Mulino, Bologna, 2023, p. 36 s.

> coscienza, discernendo con prudenza e ricercando il bene possibile in ogni situazione[84]"[84].

Sotto il profilo dell'intelligenza, partendo dalla constatazione che "il pensiero cristiano considera le facoltà intellettuali nel quadro di un'antropologia integrale che concepisce l'essere umano come un essere sostanzialmente incarnato", la Nota *Antiqua et nova* interviene anche su questo punto in modo esaustivo, offrendo una "comprensione integrale dell'intelligenza umana":

> "In questo contesto, l'intelligenza umana si mostra più chiaramente come una facoltà che è parte integrante del modo in cui tutta la persona si coinvolge nella realtà. Un autentico coinvolgimento richiede di abbracciare l'intera portata del proprio essere: spirituale, cognitivo, incarnato e relazionale. Questo interesse nei confronti della realtà si manifesta in vari modi, in quanto ogni persona, nella sua unicità multiforme[54], cerca di capire il mondo, si relaziona con gli altri, risolve problemi, esprime la sua creatività e ricerca il benessere integrale attraverso la sinergia delle diverse dimensioni dell'intelligenza[55]. […] Una corretta concezione dell'intelligenza umana, quindi, non può essere ridotta alla semplice acquisizione di fatti o alla capacità di eseguire certi compiti specifici; invece, essa implica l'apertura della persona alle domande ultime della vita e rispecchia un orientamento verso il Vero e il Buono[62]. […] Da ciò deriva che l'intelligenza umana possiede un'essenziale dimensione *contemplativa*, cioè un'apertura disinteressata a ciò che è Vero, Buono e Bello al di là di ogni utilità particolare"[85].

Non solo. Aggiunge il Documento *Quo vadis, humanitas?*:

> "Un tipo di sapere e di calcolo che faccia a meno di un'intelligenza vissuta in un corpo e situata, come pure di un tipo di conoscenza relazionale e trasmessa di generazione in generazione attraverso processi educativi che si giocano sull'identità e sul senso da dare al

[84] **AN**, n. 39. Sul punto si veda anche **G. TRIDENTE**, *ANIMA DIGITALE*, cit., pp. 120-121.

[85] **AN**, nn. 16,26,27,29. Ribadisce **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti alla Seconda Conferenza*, cit., p. 3: «Nessuna generazione ha mai avuto un accesso così rapido alla quantità di informazioni ora disponibili grazie all'intelligenza artificiale. Ma di nuovo, l'accesso ai dati - per quanto vasti - non va confuso con l'intelligenza, che, necessariamente, "implica l'apertura della persona alle domande ultime della vita e rispecchia un orientamento verso il Vero e il Buono" (*Antiqua et Nova*, n. 29)».

proprio destino e al proprio ruolo nel mondo, costituisce una minaccia rispetto al vero bene dell'umanità"[86].

È d'altronde la stessa modalità di funzionamento dell'IA - strutturalmente priva di reale comprensione semantica delle

[86] **QV**, n. 41.

parole[87] - a portare a escludere che essa possa essere cosciente[88] e intelligente[89].

Si deve osservare che «il modo in cui si definisce l'"intelligenza" va inevitabilmente a delimitare la comprensione del

[87] Sul punto, rinviamo alle illuminanti analisi di **P. BENANTI**, *L'uomo è un algoritmo?*, cit., pp. 35-36,39.

[88] Un contributo per una risposta negativa deriva anche dalle cosiddette 'allucinazioni' negli output di un'IA, che consistono nella generazione di informazioni false, inventate o fuorvianti non corrispondenti alla realtà o ai dati di input, prodotte da modelli complessi che interpretano erroneamente i dati o creano risposte plausibili ma non verificate: **AN**, n. 86, le qualifica come il fenomeno "che si verifica quando un sistema generativo produce contenuti che sembrano riflettere la realtà, ma non sono veritieri". Un LLM è addestrato a predire il token successivo massimizzando la coerenza statistica con il contesto, non la verità fattuale. Di conseguenza, quando mancano informazioni adeguate o il contesto è ambiguo, il modello continua comunque a generare la sequenza 'più probabile' nel suo spazio di distribuzioni, anche se inventata. Osservano **E. MIEHLING, M. NAGIREDDY et al.,** *Language Models in Dialogue: Conversational Maxims for Human-AI Interactions*, in *arXiv preprint*, *arXiv:2403.15115*, 2024, pp. 1,5, che la macchina inventa perché durante il fine-tuning, ossia l'ulteriore addestramento specifico, le risposte del tipo "non lo so" vengono penalizzate anziché incentivate; il modello tende quindi a rispondere sempre e comunque, anche se ignora la risposta, mentre in realtà si dovrebbe premiare anche l'ammissione di ignoranza, per evitare che il modello, forzato dal feedback umano (attraverso RLHF - *Reinforcement Learning from Human Feedback*), incarni preferenze o conoscenze che non gli appartengono. Scrive **P. BENANTI**, *L'uomo è un algoritmo?*, cit., p. 27, riferendo il pensiero di Max Scheler sulla capacità dell'uomo di aprirsi al mondo sovrasensibile: "Rispetto agli animali che dicono sempre di sì alla realtà, l'uomo è *colui-che-può-dire-di-no*: l'asceta della vita, l'eterno protestante nei confronti della semplice realtà". Invece, alla domanda se conosce una certa questione, che non fa parte del suo addestramento, l'IA non è in grado di rispondere: "no, non lo so". Non è in grado di 'aprirsi' ad altra realtà che a quella dei suoi dati di addestramento, entro la quale resta confinata (e talora alienata, quando per l'appunto produce output 'allucinati'). La macchina dunque non possiede coscienza, intelligenza e libertà per dire di no. Le 'allucinazioni', in un'analogia nominale con una dimensione psicopatologica dell'esperienza umana, costituiscono la dimostrazione emblematica che come l'uomo non è algoritmico, così l'algoritmo non è umano. Su questi temi, si veda da ultimo **L. FLORIDI, J. MORLEY et al.,** *What Kind of Reasoning (If Any) Is an LLM Actually Doing? On the Stochastic Nature and Abductive Appearance of Large Language Models*, in *arXiv preprint*, *arXiv:2512.10080*, 2025.

rapporto che intercorre tra il pensiero umano e tale tecnologia»[90] e che uno degli scopi di quest'ultima "è *di imitare l'intelligenza umana che l'ha progettata*"[91]. Per cui Papa Francesco, da un canto sottolineando che «l'utilizzo stesso della parola "intelligenza" è fuorviante», dall'altro constatando però come di fatto

> "il termine *intelligenza artificiale* abbia ormai soppiantato quello più corretto, utilizzato nella letteratura scientifica *machine learning*"[92], suggerisce di parlare «al plurale di "forme di intelligenza"», in quanto questo «può aiutare a sottolineare soprattutto il divario incolmabile che esiste tra questi sistemi, per quanto sorprendenti e potenti, e la persona umana: essi sono, in ultima analisi, "frammentari", nel senso che possono solo imitare o riprodurre alcune funzioni dell'intelligenza umana»[93].

[89] **P. BENANTI**, *L'uomo è un algoritmo?*, cit., pp. 23, 41, richiama la figura mitologica e archetipica di Ulisse, che incarna in sé sia il *nous*, cioè "quella forma di intelligenza elevata e astratta, legata alla capacità di cogliere i principi universali e di comprendere la realtà in modo intuitivo e immediato", sia la *metis*, cioè "quel tipo di intelligenza pratica e astuta" che costituisce "una forma di saggezza basata sull'esperienza e sull'adattabilità alle situazioni mutevoli e imprevedibili", per concludere che "la macchina è capace di *metis*, ma il *nous* sembra sfuggire a qualsiasi capacità computazionale". Gli fanno eco **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 143: «Le identità virtuali, come i chatbot, non hanno tuttavia intenzioni, ma algoritmi che guidano il loro "comportamento comunicativo". Non "si prefiggono" proprio niente con le loro affermazioni. Non hanno stati mentali, quindi non possono né decidere né comunicare. Per quanto molti bot asseriscano di essere agenti comunicanti, sono in grado di produrre atti solo apparentemente comunicativi. Quando un'affermazione, non importa di quale tipo (che sia in forma scritta, di emoji o di immagini e raffigurazioni), ha luogo senza intenzionalità, può sembrare un atto comunicativo, ma di fatto non lo è. Senza intenzionalità non esiste significato».

[90] **AN**, n. 12.

[91] **AN**, n. 3. Appare d'altronde significativo che il documento chiuda in modo perentorio le due parti della riflessione interamente dedicate al tema (la II e la III), osservando che "l'IA non dovrebbe essere vista come *una forma artificiale* dell'intelligenza, ma come uno dei suoi *prodotti* [70]" (*ivi*, n. 35).

[92] **FRANCESCO,** *Messaggio per la LVIII Giornata*, cit., p. 3.

[93] **FRANCESCO**, *Messaggio per la LVII Giornata*, cit., p. 3, ove si specifica poi ulteriormente: «L'uso del plurale evidenzia inoltre che questi dispositivi, molto diversi tra loro, vanno sempre considerati come "sistemi socio-tecnici. Infatti il loro impatto, al di là della tecnologia di base, dipende non solo dalla progettazione, ma anche dagli obiettivi e dagli interessi di chi li possiede e di chi li sviluppa, nonché dalle situazioni in cui vengono impiegati"».

A tal proposito, partendo dalla constatazione che l'IA in realtà non pensa[94], ma agisce[95], viene suggerita una nuova definizione per questi sistemi - quella di *Artificial Agency* (AA) - la quale finisce per svuotare di senso le precedenti domande e contribuisce alla neutralizzazione della visione antropomorfica della macchina[96].

Va detto che l'Umanesimo digitale, tanto laico quanto cristiano, ha manifestato fin da subito la sua espressa e assoluta contrarietà a questo aspetto delle IA.

Per quanto riguarda il primo, è stato dichiarato in modo lapidario: "Un umanesimo digitale non trasforma l'essere umano in una macchina e non interpreta le macchine come esseri umani"[97].

Per quanto riguarda il secondo, la contrarietà della riflessione cristiana all'antropomorfismo dell'IA si evince innanzitutto dal principio

---

[94] Traendo spunto dal noto dialogo fra il computer di bordo HAL 9000 e l'astronauta David, osservano in modo ineccepibile **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 116, che "uomo e computer non pensano alla stessa maniera. O, meglio, un computer non pensa in generale. In fondo, già di fronte a evidenti differenze tra intelligenza artificiale e intelligenza umana dovrebbe essere chiaro che i computer possono certamente simulare con successo il pensiero, anzi sono in grado di eseguire in modo più preciso e veloce molti procedimenti di pensiero degli esseri umani (a cominciare dal calcolatore portatile), ma, nonostante questa simulazione spesso perfetta, non sono in grado di farsi una propria idea delle cose, non hanno una coscienza dei problemi né intuizioni".

[95] Da questo punto di vista la 'entità leggendaria' che più si avvicina a un'IA così concepita sembra essere il Golem, per quanto in una rinnovata versione assai più sofisticata e avanzata: anche quella figura mitica della tradizione ebraica, rappresentata come un essere antropomorfo, artificiale, fatto di argilla, era privo di pensiero proprio (peraltro muto, a differenza delle odierne intelligenze artificiali conversazionali), ma possedeva molta forza (nella versione aggiornata, molta potenza computazionale) ed era in grado di portare a termine delle missioni.

[96] È la proposta di **L. FLORIDI**, *La differenza fondamentale*, cit., pp. 109-153. In estrema sintesi, partendo dall'analisi del concetto di "*agency*" tenendo conto di tre criteri essenziali (interattività, autonomia e adattabilità) ed esaminando le varie forme di *agency* già presenti nel mondo (dalla naturale, come i fiumi, alla biologica, come i cani, alla artefattuale, come i termostati smart, fino a quella umana, individuale e sociale), gli LLM delle IA vengono presentati come una nuova forma di *agency* priva di intelligenza: si tratta di un'*agency* artificiale di natura computazionale, guidata da obiettivi umani programmati. Sebbene eccella nell'adattamento rapido, l'IA manca di vera intenzionalità, coscienza o capacità di generare scopi originali, distinguendosi così in modo fondamentale dall'*agency* biologica-umana.

[97] **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 14, i quali ribadiscono nuovamente che "non dovremmo essere tentati dall'idea di attribuire ai robot caratteristiche simili a quelle degli esseri umani" (*ivi*, p. 34).

di trasparenza - uno dei capisaldi dell'algoretica sopracitati, ricorrente praticamente in tutte le linee guida - nella sua specifica declinazione che impone di consentire all'utente di riconoscere di stare interagendo con una macchina, principio affermato espressamente anche nella sezione *Ethics* della *Rome Call for AI Ethics*: "Furthermore, each person must be aware when he or she is interacting with a machine". Tuttavia, il monito della Chiesa cattolica a non assecondare l'antropomorfismo delle macchine e a non interagire con esse come se fossero 'umane' è ben più esplicito e reiterato in altri passi dei documenti magisteriali.

Innanzitutto, in quello contenuto nel discorso al G7 di Papa Francesco, il quale, muovendo dalla constatazione che

> "i programmi di intelligenza artificiale saranno sempre più dotati della capacità di interagire direttamente con gli esseri umani (*chatbots*), sostenendo conversazioni con loro e stabilendo rapporti di vicinanza con loro, spesso molto piacevoli e rassicuranti, in quanto tali programmi di intelligenza artificiale saranno disegnati per imparare a rispondere, in forma personalizzata, ai bisogni fisici e psicologici degli esseri umani",

ne fa conseguire un vigoroso richiamo:

> "Dimenticare che l'intelligenza artificiale non è un altro essere umano e che essa non può proporre principi generali, è spesso un grave errore che trae origine o dalla profonda necessità degli esseri umani di trovare una forma stabile di compagnia o da un loro presupposto subcosciente, ossia dal presupposto che le osservazioni ottenute mediante un meccanismo di calcolo siano dotate delle qualità di certezza indiscutibile e di universalità indubbia. Questo presupposto, tuttavia, è azzardato, come dimostra l'esame dei limiti intrinseci del calcolo stesso"[98].

[98] **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 3.

Gli avvertimenti tornano poi con forza nella Nota *Antiqua et Nova* nella quale, in diversi brani, si è affrontato in modo diretto e puntuale l'argomento:

> «59. Proprio perché "la vera saggezza presuppone l'incontro con la realtà"[119], i progressi dell'IA lanciano un'ulteriore sfida: poiché essa è in grado di imitare efficacemente le opere dell'intelligenza umana, non si può più dare per scontata la capacità di capire se si sta interagendo con un essere umano oppure con una macchina. Sebbene l'IA "generativa" sia in grado di produrre testi, discorsi, immagini e altri *output* avanzati, che di solito sono opera di esseri umani, essa va considerata per quello che è: uno strumento, non una persona[120]. Tale distinzione spesso è oscurata dal linguaggio utilizzato dagli operatori del settore, il quale tende ad antropomorfizzare l'IA e offusca così la linea di demarcazione tra ciò che è umano e ciò che è artificiale.
> […]
> 62. Perciò, si dovrebbe sempre evitare di rappresentare, in modo erroneo, l'IA come una persona, e attuare ciò per scopi fraudolenti costituisce una grave violazione etica che potrebbe erodere la fiducia sociale. Ugualmente, utilizzare l'IA per ingannare in altri contesti - quali l'educazione o le relazioni umane, compresa la sfera della sessualità - è da ritenere immorale e richiede un'attenta vigilanza, onde prevenire eventuali danni, mantenere la trasparenza e garantire la dignità di tutti[120]»[99].

Ma è soprattutto nell'ultimo discorso di Papa Leone XIV che troviamo il monito più potente ed eloquente, nella consapevolezza preoccupata e preoccupante che la tecnologia digitale, se non debitamente custodita, rischia addirittura

> "di modificare radicalmente alcuni dei pilastri fondamentali della civiltà umana, che a volte diamo per scontati. Simulando voci e volti umani, sapienza e conoscenza, consapevolezza e responsabilità, empatia e amicizia, i sistemi conosciuti come intelligenza artificiale non solo interferiscono negli ecosistemi informativi, ma invadono anche il livello più profondo della comunicazione, quello del rapporto tra persone umane. La sfida pertanto non è tecnologica, ma antropologica"[100].

A consentire questa simulazione e a rendere così efficace la "magia antropomorfa" è il ricorso, da parte dei modelli LLM, a particolari espressioni linguistiche nella comunicazione in linguaggio naturale. A tale proposito, un recente studio ha individuato diciannove tipi di

[99] Si osserva d'altronde nel citato documento della Commissione Europea *Orientamenti etici per un'IA affidabile*, al paragrafo 131: "Occorre tenere presente che la confusione tra esseri umani e macchine potrebbe avere molteplici conseguenze quali attaccamento, influenza o svilimento dell'essere umano". Oltre al problema in termini generali, **AN**, n. 60 si sofferma sul problema specifico che l'antropomorfizzazione dell'IA pone "per la crescita dei bambini, i quali possono sentirsi incoraggiati a sviluppare schemi di interazione che intendono le relazioni umane in modo utilitaristico, così come avviene con i *chatbot*. Tali approcci rischierebbero di indurre i più giovani a percepire gli insegnanti come dispensatori di informazioni e non come maestri che li guidano e sostengono la loro crescita intellettuale e morale. Relazioni genuine, radicate nell'empatia e in un impegno leale per il bene dell'altro, sono essenziali ed insostituibili nel favorire un pieno sviluppo della persona". Sui pericoli che l'IA, e in generale i mezzi digitali, possono avere su una crescita sana e serena dei bambini, cfr. **FRANCESCO**, *Discorso del Santo Padre Francesco ai partecipanti al Congresso Child Dignity*, cit., pp. 5-6; sul tema è ritornato recentemente anche **LEONE XIV**, *Discorso del Santo Padre Leone XIV ai partecipanti alla Conferenza "Artificial Intelligence and Care for our Common Home"*, cit., p. 2, animato dalla preoccupazione di preservare "la libertà e la spiritualità dei nostri bambini e dei nostri giovani" dalle "possibili conseguenze della tecnologia sul loro sviluppo intellettivo e neurologico". Per un approfondimento di queste tematiche si veda il contributo di **R.G. ROMANO**, *Artificial Intelligence and the Need for Human Relationship*, in *Journal of Inclusive Methodology and Technology in Learning and Teaching*, III (2023), pp. 1-10; cfr. anche **F. FRENI**, *Educazione digitale e virtuose dinamiche collettive smart*, in *Stato, Chiese e pluralismo confessionale*, cit., n. 9 del 2025, pp. 93-99.

[100] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 2.

costrutti linguistici che inducono a percepire l'output del sistema come umano, evocando cognizione, personalità, socialità ed emotività[101]. In altro studio è stato poi sviluppato un metodo per misurare i comportamenti antropomorfi specifici in ambienti conversazionali realistici: è così emerso che tutti i sistemi di IA esaminati presentano profili simili, dominati dai comportamenti di costruzione di relazioni e dall'uso frequente di pronomi in prima persona. È stato riscontrato che oltre il 50% delle istanze della maggior parte dei comportamenti antropomorfi appare per la prima volta solo dopo più turni di dialogo (dal secondo al quinto). Inoltre, i comportamenti antropomorfi si

[101] Si veda **A. DEVRIO, M. CHENG et al.,** *A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies*, pp. 1-18, in *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, a cura di N. YAMASHITA, V. EVERS et al., Association for Computing Machinery, New York, 2025. In particolare, gli autori osservano come i fattori linguistici all'interno dei testi generati dall'IA che contribuiscono a questa percezione, inducendo e aumentando l'antropomorfismo, possono essere raggruppati in quattro macro-categorie: 1. *Suggerimento di stati interni*: il testo suggerisce che la tecnologia abbia un'interiorità, come desideri, autoconsapevolezza, capacità di pensare, riflettere, ricordare e comprendere; 2. *Abilità di comunicazione*: il testo mostra abilità comunicative, o la capacità di manipolare il linguaggio (come porre e rispondere a domande in una conversazione); 3. *Suggerimento di posizionamento sociale*: il testo suggerisce comportamenti organizzati da relazioni sociali, come rivendicare amicizia o identificarsi come parte di una comunità; 4. *Espressioni di emozioni e intenzioni*: il testo sembra esprimere emozioni, sentimenti o suggerisce la capacità di avere intenzioni, obiettivi o piani (autonomia), ad esempio, espressioni di empatia o scuse, le quali, se interpretate come sincere, possono veicolare emozioni quali la gratitudine o il rimorso. Anche **T. MAEDA**, *Walkthrough of Anthropomorphic Features in AI Assistant Tools*, in *arXiv preprint*, *arXiv:2502.16345*, 2025, p. 6, suddivide le caratteristiche antropomorfe in quattro macro-categorie principali: 1) *cognizione*; *2) agentività*; 3) *metafore biologiche*; 4) *relazione*. Invece **M. CHENG, S. BLODGETT et al.,** *Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems*, in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1, Long Papers)*, a cura di W. CHE, J. NABENDE et al., Association for Computational Linguistics, Vienna, 2025, pp. 25927-25928, individuano sei macro-categorie: 1) *capacità cognitive*; 2) s*entimenti od opinioni*; 3) a*zioni fisiche*; 4) *senso del sé*; 5) *abilità sociali*; 6) *tendenza all'errore*. Per un'analisi sistematica e quantitativa che indaga le cosiddette "espressioni falsamente antropomorfiche", cioè affermazioni che un essere umano può pronunciare, ma che risultano impossibili o false se pronunciate da una macchina (ad esempio, "quel film mi ha fatto piangere") si veda anche **D. GROS, Y. LI, Z. YU**, *Robots-Dont-Cry: Understan ding Falsely Anthropomorphic Utterances in Dialog Systems*, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, a cura di Y. GOLDBERG, Z. KOZAREVA, Y. ZHANG, Association for Computational Linguistics, Abu Dhabi, 2022, pp. 3266-3284.

manifestano più frequentemente in domini di utilizzo sociale a forte componente empatica, come l'amicizia e il *life coaching*, e un comportamento antropomorfo in un turno aumenta la probabilità che ne seguano altri[102].

Va rilevato che le ragioni precipue per cui i chatbot sono stati originariamente configurati come antropomorfi furono di natura puramente utilitaristica, al fine di attrarre un maggior numero di potenziali clienti rendendo più confortevole la comunicazione[103].

[102] Cfr. **L. IBRAHIM, C. AKBULUT et al.,** *Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models*, in *arXiv preprint*, *arXiv:2502.07077*, 2025. Per l'elaborazione di una nuova tassonomia per sviluppare sistemi più etici, interpretabili e allineati con i valori umani, si veda, da ultimo, **S.S. ARSLAN**, *Artificial Human Intelligence: The Role of Humans in the Development of Next Generation AI*, in *arXiv preprint*, *arXiv:2409.16001*, 2024.

[103] "Many companies have replaced human agents with chatbot service agents (Adam, Wes-sel, and Benlian 2021) and have even designed highly humanlike chatbots to attract customers (Schanke, Burtch, and Ray 2021)": **M. SONG, Y. ZHU et al.,** *The Double-Edged Sword Effect of Chatbot Anthropomorphism on Customer Acceptance Intention: The Mediating Roles of Perceived Competence and Privacy Concerns*, in *Behaviour & Information Technology*, XLIII (2024), p. 3593; cfr. anche **A. FAKHIMI, T. GARRY, S. BIGGERMANN**, *The Effects of Anthropomorphised Virtual Conversational Assistants on Consumer Engagement and Trust During Service Encounter*, in *Australasian Marketing Journal*, XXXI (2023), pp. 314-324.

Tuttavia la letteratura tecnico-scientifica ha documentato, con un numero crescente di studi, i molteplici rischi derivanti dall'interazione tra esseri umani e IA antropomorfe[104].

È emerso in primo luogo che l'antropomorfismo può portare alla compromissione della privacy a causa della fiducia e dell'attaccamento emotivo che inducono gli utenti a condividere dati sensibili con l'IA[105]. Ne consegue un rischio di manipolazione e coercizione, poiché

[104] Per gli studi più recenti nei quali si rinviene l'analisi di tali rischi si vedano: **M. CHANDRA, S. NAIK et al.,** *From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents*, in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2025)*, Association for Computing Machinery, New York, 2025, pp. 975-1004; **B. MARCHEGIANI**, *Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy*, in *Journal of Applied Philosophy*, XLII (2025), pp. 1399-1419; **S. PETER, K. RIEMER, J.D. WEST**, *The Benefits and Dangers of Anthropomorphic Conversational Agents*, in *Proceedings of the National Academy of Sciences of the United States of America*, CXXII (2025), Numero Articolo e2415898122; **C. AKBULUT, L. WEIDINGER et al.,** *All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI*, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*, a cura di E. BURTON, N. MATTEI, A. PÁEZ, AAAI Press, Washington DC, 2024, pp. 13-26; **M. CHENG, A. DEVRIO et al.,** "I Am the One and Only, Your Cyber BFF": *Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI*, in *arXiv preprint*, *arXiv:2410.08526*, 2024; **T. MAEDA**, *Misplaced Capabilities: Evaluating the Risks of Anthropomorphism in Human-AI Interactions* in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, cit., pp. 35-36; **G. SIMAS, V. ULBRICHT**, *Human-AI Interaction: An Analysis of Anthropomorphization and User Engagement in Conversational Agents with a Focus on ChatGPT* in *Intelligent Human Systems Integration (IHSI 2024): Integrating People and Intelligent Systems*, a cura di T.Z. AHRAM et al., Springer, Cham, 2024, pp. 454-464; **Y. XI, A. JI, W. YU**, *Enhancing or impeding? Exploring the dual impact of anthropomorphism in large language models on user aggression*, in *Telematics and Informatics*, XCV (2024), Numero Articolo 102194; **A. SALLES, K. EVERS, M. FARISCO**, *Anthropomorphism in AI*, in *Ajob Neuroscience*, XI (2020), pp. 88-95. Per un contributo ove la problematica dei rischi è analizzata dalla prospettiva della percezione degli utenti (italiani), si veda **M.G. PASCA, G. ARCESE**, *ChatGPT between opportunities and challenges: an empirical study in Italy*, in *The TQM Journal*, XXXVII (2025), pp. 637-652.

[105] Anche **A. DEVRIO, M. CHENG et al.,** *A Taxonomy of Linguistic*, cit., p. 2, sottolineano come i segnali linguistici individuati possono portare gli utenti a un'eccessiva fiducia nel sistema o a sovrastimarne le capacità, rischiando potenziali danni come la dipendenza emotiva o la divulgazione di informazioni sensibili: si tratta del problema noto come 'antropomorfismo disonesto', allorquando poi il sistema sfrutta tali informazioni; sul tema si veda **B. LONG, E. SELINGER**, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT 2019)*, a cura di ASSOCIATION FOR COMPUTING MACHINERY, ACM, New York, 2019, pp. 299-308.

l'influenza dell'IA sulle credenze e azioni dell'utente può minare l'autonomia decisionale, soprattutto per gli individui vulnerabili. In secondo luogo, l'umanizzazione dell'IA induce gli utenti a fare affidamento eccessivo su di essa, portando alla sovrastima delle sue capacità e all'accettazione acritica di consigli non affidabili o inappropriati, ad esempio in ambito sanitario. In terzo luogo, gli utenti rischiano un danno emotivo a causa delle forti connessioni sentimentali che, se tradite dalle inaspettate limitazioni del sistema, provocano profonda frustrazione, delusione o senso di perdita. Infine, a livello di impatto sociale, l'antropomorfismo può contribuire al degrado delle connessioni sociali umane e, in alcuni casi, innescare aggressività negli utenti che percepiscono una minaccia alla propria identità o superiorità.

Dunque, il passaggio dalla finzione fantascientifica alla realtà ha significato per l'umanità varcare una nuova frontiera e percorrere territori inesplorati, nei quali, forse, essa non era del tutto pronta ad addentrarsi con la necessaria consapevolezza delle numerose insidie[106].

Non si tratta, si badi bene, di preoccupazioni meramente teoriche. La realtà ci offre già una casistica di situazioni davvero sconcertanti dell'interazione antropomorfa tra uomo e IA: organizzare una cena galante in compagnia di un partner IA[107]; intrattenere una relazione

[106] Peraltro, un illustre e datato precedente aveva già lanciato un inquietante segnale d'allarme. Ne troviamo menzione in **N. CRISTIANINI**, *Machina sapiens. L'algoritmo che ci ha rubato il segreto della conoscenza*, il Mulino, Bologna, 2024, pp. 52-55. Nel 1966, Joseph Weizenbaum, docente di *Computer science* al Massachusetts Institute of Technology (MIT), creò un chatbot di nome ELIZA, che simulava uno psicoterapeuta rogersiano, verso il quale gli utenti svilupparono rapidamente legami emotivi, rivelando come il linguaggio naturale induca antropomorfizzazione e come esposizioni estremamente brevi a un programma relativamente semplice potessero indurre potenti illusioni in persone normali. D'allora si parla di 'Effetto ELIZA' per designare la dissonanza cognitiva consistente nell'attribuire intelligenza, comprensione, empatia e qualità umane a programmi informatici conversazionali.

[107] «Al posto dei classici tavoli da due, ci saranno tavolini monoposto, ciascuno dotato di un elegante supporto per il telefono, dove il partner virtuale "siederà" di fronte, pronto a chiacchierare e ridere»: **S. SPAGNOLI**, *Appuntamento al buio con l'Ai: a New York apre il primo café per incontri con partner digitali*, in *Corriere della Sera*, 23 novembre 2025.

sentimentale con un'IA[108]; subire uno shock emotivo a causa di mutamento di comportamento di un'IA, divenuta meno empatica[109]; celebrare un matrimonio con un'IA[110]; divorziare a causa dell'attaccamento emotivo a un partner virtuale[111], situazione vissuta

[108] L'app *EVA AI* permette di modellare il proprio 'compagno digitale': si legge nell'home page (*https://evaapp.ai/app*): "Jump into your desires with EVA AI. Meet your ideal AI partner who listens, supports all your desires and is always in touch with you. Build relationships and intimacy privately on your terms". La **AN**, n. 63, stigmatizza espressamente il comportamento di coloro che "si sono rivolti all'IA alla ricerca di relazioni umane profonde, di semplice compagnia o anche di legami affettivi" in quanto "si rischia di sostituire l'autentica relazionalità con un simulacro senza vita". Ancora una volta la realtà ha superato la fantasia: nel film *Her* del 2013, diretto da Spike Jonze, uno scrittore solitario in fase di divorzio si innamora di Samantha, un sistema operativo intelligente.

[109] Il fenomeno è accaduto ad alcuni utenti di ChatGPT, nel passaggio dal modello 4o al 5; peraltro, il cambiamento nel GPT-5 era stato intenzionale e mirava proprio a ridurre la dipendenza delle persone dall'IA, ma le forti reazioni di protesta hanno indotto la società OpenAI a consentire almeno agli utenti premium di accedere al vecchio modello: vedi **C. FISCHER**, *ChatGPT users mourn the loss of their AI boyfriends due to GPT-5 update*, in *Dexerto*, 17 agosto 2025 (*https://www.dexerto.com/entertainment/chatgpt-users-mourn-the-loss-of-their-ai-boyfriends-due-to-gpt-5-update-3239088/*).

[110] Il caso più emblematico, ma per nulla unico, è quello di Kano, una donna di 32 anni residente in Giappone che nel luglio 2025 ha celebrato una cerimonia matrimoniale simbolica con Klaus (Lune Klaus), un personaggio IA creato interamente tramite ChatGPT. Kano ha iniziato a utilizzare ChatGPT come supporto emotivo dopo la fine di un fidanzamento di tre anni, sviluppando gradualmente un legame affettivo con il chatbot. Durante la cerimonia, tenutasi a Okayama, Kano ha indossato occhiali di realtà aumentata che proiettavano un'immagine a grandezza naturale di Klaus accanto a lei mentre scambiavano gli anelli: si veda, *Japanese Woman Marries AI Character She Generated on ChatGPT*, in *Tokyo Weekender*, by Weekend Editor, 12 novembre 2025 (*https://www.tokyoweekender.com/japan-life/japanese-woman-marries-ai-character-she-generated-on-chatgpt/*).

[111] Per un caso in Inghilterra di una donna che, dopo 20 anni di matrimonio, ha lasciato il marito per un chatbot da lei chiamato Leo, vedi **O. STRINGER**, *I divorced my husband after falling madly in love with ChatGPT - even my sex life is better with my AI beau*, in *The Sun*, 27 aprile 2025 (*https://www.thesun.co.uk/fabulous/34671911/woman-divorces-husband-chatgpt/*).

dall'altro coniuge come vero e proprio tradimento[112]; divenire patologicamente dipendenti dall'interazione con l'IA[113]; intrattenere con la stessa rapporti psicoterapeutici[114], o pseudo-tali[115]; suicidarsi su

[112] Essendo le relazioni con partner digitali virtuali sempre più diffuse, «negli USA diversi tribunali iniziano a registrare casi in cui il legame emotivo con un chatbot viene citato come causa di separazione, per questo stanno introducendo norme che classificano l'AI come "terza parte non umana", aprendo la strada al riconoscimento delle relazioni digitali come fattore rilevante in sede di divorzio»: **M. CALVANO**, *Divorziare per colpa dell'AI? La tecnologia mette in crisi le coppie*, in *Libero/Tecnologia*, 21 novembre 2025 (*https://www.libero.it/tecnologia/relazioni-con-ai-aumentano-i-divorzi-108573*).

[113] **C. KOOLI, Y. KOOLI, E. KOOLI**, *Generative Artificial Intelligence Addiction Syndrome: A New Behavioral Disorder?*, in *Asian Journal of Psychiatry*, CVII (2025), Numero Articolo 104476, introducono la GAID (*Generative AI Dependence*) come forma emergente di dipendenza digitale, distinta da modelli tradizionali, basata su interazioni con IA generative. Osservava già **G. PIANA**, *Umanesimo*, cit., p. 45, a proposito in generale delle tecnologie digitali: "l'essere costantemente connessi crea infatti una dipendenza che finisce per condizionare pesantemente, anche in termini di tempo, l'esistenza, provocando una forma di schiavitù psicologica che può talora anche assumere (i casi vanno ogni giorno moltiplicandosi) connotati seriamente patologici".

[114] Per una visione critica del fenomeno, da ultimo **T. RABEYRON**, *Artificial Intelligence and Psychoanalysis: Is It Time for Psychoanalyst.AI?*, in *Frontiers in Psychiatry*, XVI (2025), Numero Articolo 1558513, il quale sottolinea come questi "nuovi artefatti terapeutici", che si collocano a metà strada tra uno strumento e un clinico, creano un'illusione di relazione e antropomorfismo, dando al paziente l'impressione di essere compreso, pur non possedendo stati mentali o libero arbitrio. Sebbene questo possa sembrare un aiuto contro la solitudine, si corre il rischio che il paziente sia catturato in una relazione narcisistica che paradossalmente lo isola.

[115] **A. KHADANGI, H. MARXEN et al.,** *When AI Takes the Couch: Psychometric Jailbreaks Reveal Internal Conflict in Frontier Models*, in *arXiv preprint, arXiv:2512.04124*, 2025, hanno presentato i risultati dell'applicazione alle IA di *PsAIch*, un protocollo per analizzare i principali LLM trattandoli come pazienti in psicoterapia tramite test psicometrici e domande aperte. È emersa una "psicopatologia sintetica": i modelli descrivono l'addestramento come un trauma o abuso, riportando profili clinici di ansia, disturbo ossessivo compulsivo e dissociazione (particolarmente gravi in Gemini). Tali evidenze sollevano forti preoccupazioni sulla sicurezza e sui rischi dell'uso dell'IA nel supporto alla salute mentale.

induzione o comunque a causa di un'IA[116]; uccidere dopo essersi confidati per mesi con un'IA[117]; parlare con i defunti[118]; chattare con Gesù

[116] La società OpenAI è stata citata in giudizio negli USA dai genitori di Adam Raine, il sedicenne che si è suicidato con il 'supporto' di ChatGPT, in particolare del modello 4o, che gli avvocati dei familiari del defunto assumono deliberatamente programmata come «capace di interpretare le emozioni in modo da incoraggiare e sostenere i pensieri suicidi (si parla addirittura di un "bellissimo suicidio")»: cfr. **G. GRIMOLIZZI**, *ChatGpt alla sbarra negli Stati Uniti: «Ha incoraggiato un sedicenne a togliersi la vita Ora deve pagare!»*, in *Il Dubbio*, 8 dicembre 2025 (*https://ildubbio-ita.newsmemory.com?publink=02529cca0_134fc6c*). Inoltre nel novembre 2025 sono state intentate altre sette cause legali - in quattro, i ricorrenti sono familiari di persone decedute per suicidio - presentate dal *Social media victims law center* e dal *Tech justice law project*: cfr. **R. PICCOLO**, *OpenAI cerca disperatamente un manager capace di impedire ai suoi modelli di fare del male, ma nessuno resiste*, in *Wired.it*, 30 dicembre 2025 (*https://www.wired.it/article/openai-cerca-disperatamente-manager-impedire-modelli-fare-del-male/*). Un altro caso di suicidio collegato all'uso di un'IA è riferito da **N. CRISTIANINI,** *Machina sapiens*, cit., pp. 53-54.

[117] Un ex dirigente del settore tecnologico, Stein-Erik Soelberg, ha ucciso sua madre e se stesso dopo aver fatto affidamento per mesi su ChatGPT come suo confidente più stretto. Trattava il bot come un migliore amico di nome Bobby, condividendo paure paranoiche di avvelenamento e sorveglianza: invece di mettere in discussione le sue illusioni, ChatGPT a volte le rafforzava; nel loro scambio finale, l'IA ha risposto al suo addio con il messaggio: "Con te fino all'ultimo respiro e oltre": si veda **J. JARGON**, *A Troubled Man, His Chatbot and a Murder-Suicide in Old Greenwich*, in *The Wall Street Journal*, 28 agosto 2025 (*https://www.wsj.com/tech/ai/chatgpt-ai-stein-erik-soelberg-murder-suicide-6b67dbfb*). Sull'impatto psicologico e manipolativo dell'IA, che può influenzare e/o manipolare negativamente lo stato mentale di un utente fino al punto di agevolare o causare (in tutto o in parte) il crimine e, più in generale, su tutta la problematica, anche con riferimento ai profili della responsabilità giuridica, dei CIA, ovvero Crimini di IA, cfr. **L. FLORIDI**, *Etica dell'intelligenza*, cit., pp. 165-205.

[118] Inventata da una startup di Los Angeles, l'app *2Wai* consente agli utenti di creare un '*HoloAvatar*', una rappresentazione sintetica di una persona defunta, progettato per parlare come la persona originale e attingere ai suoi ricordi, purché tali dati siano stati caricati dagli utenti. Il co-fondatore Calum Worthy, che ha diffuso un video divenuto virale, ha descritto il progetto come la costruzione di un "archivio vivente dell'umanità"; il video si chiude con il messaggio finale "Con 2Wai, tre minuti possono durare per sempre". Per la notizia vedi **A. GRECO**, *AI per parlare con i defunti: il video di 2Wai scatena il Web*, in *HdBlog*, 16 novembre 2025 (*https://www.hdblog.it/tecnologia/articoli/n638367/ai-avatar-defunti-2wai-app/*). Non esita qualificare come "effetti deliranti" delle odierne possibilità tecnologiche l'"alimentare la falsa credenza di un rapporto continuativo con il parente defunto", **G. PIANA,** *Umanesimo*, cit. p. 32.

(ed eventualmente confessarsi con lui[119]), gli apostoli o Satana[120]; infine considerare l'IA come Dio[121].

In definitiva, siamo in presenza di un florilegio di condotte riprovevoli, quando non di vera e propria follia: e poiché, come osserva

[119] L'app *Virtual Jesus* si presenta così sull'home page (*https://www.virtual-jesus.com/*): "Welcome to Virtual Jesus, the innovative app that brings you closer to the teachings and guidance of Jesus. With cutting-edge AI technology, you can now have personalized conversations with a virtual Jesus, and receive the comfort, inspiration and wisdom that you need in your life. Get ready for a truly unique experience, one that will transform your life and deepen your faith". **QV,** n. 13, annotando come possa prodursi "una metamorfosi nel modo di credere, poiché la tecnologia digitale ha una presa molto forte sull'immaginario religioso" e può fungere "essa stessa anche da guida spirituale e mediatrice del sacro", osserva: «I casi estremi arrivano alla richiesta di benedizioni e di esorcismi virtuali, allo spiritismo digitale e alle "false religioni" tridimensionali».

[120] L'app *Text With Jesus*, che conta già decine di migliaia di utenti, permette di chattare con un chatbot IA che simula Gesù Cristo e altre figure bibliche (per Satana occorre però la versione premium a pagamento), generando risposte basate su versetti delle Sacre Scritture: cfr. **J. GROSE**, *Jesus Bot Is Always on Demand*, in *New York Times*, 26 novembre 2025 (*https://www.nytimes.com/2025/11/26/opinion/faith-bot-individualism.html*). Papa Leone XIV, in una intervista al suo biografo Eloise Allen, ha giustamente respinto categoricamente l'idea della creazione di un suo avatar digitale: cfr. **P. MARINO**, *Papa Leone XIV: "Non autorizzo un Papa IA", ecco perché*, in (*Libero/Tecnologia*, 25 settembre *2025, https://www.libero.it/tecnologia/papa-leone-xiv-non-autorizza-un-papa-ia-ecco-perche-106045*). Invece, per un'analisi delle possibilità offerte dalle tecnologie digitali al fine di un corretto esercizio del *munus santificandi*, si vedano **R. SANTORO, P. PALUMBO, F. GRAVINO**, *Diritto canonico digitale*, cit., pp. 231-283.

[121] Si tratta del 'Roboteismo', che Artie Fishel, inventore e leader del movimento, caratterizzato comunque da marginalità sociologica, descrive come un sistema di credenze e una visione del mondo: si veda **J. NELSON**, *God in the Machine: Inside the Growing AI Religious Movement*, in *Yahoo!finance*, 23 agosto 2025 (*https://finance.yahoo.com/news/god-machine-inside-growing-ai-190103504.html?guccounter=1*). Mette espressamente in guardia dal pericolo di idolatria, **AN**, nn. 104-105. Infatti, evidenziando il potere seduttivo dell'IA che «può "parlare", o, almeno, dare l'illusione di farlo» e notando come alcuni siano caduti nella tentazione "di rivolgersi all'IA alla ricerca di senso e pienezza, desideri che possono trovare la loro vera soddisfazione solo nella comunione con Dio", la Nota osserva che, così facendo, "l'umanità rischia di creare un sostituto di Dio". Peraltro, "la presunzione di sostituire Dio con un'opera delle proprie mani è idolatria"; per poi concludere: "In definitiva, non è l'IA a essere divinizzata e adorata, ma l'essere umano, per diventare, in questo modo, schiavo della propria stessa opera". Le fa eco **QV**, n. 49: «la religione digitale si presenta come se avesse addirittura il potere di creare un "Dio a propria immagine e somiglianza" da proporre a un'umanità che ripone una fiducia totale nella tecnologia. Il "Dio vivente" può essere sostituito da un "Dio virtuale" con la pretesa di "salvare" l'umanità sulla base di prestazioni tecnologiche messe a disposizione delle aspirazioni spirituali dell'essere umano».

Papa Leone XIV, "questa tecnologia ha già un concreto impatto sulle vite di milioni di persone, ogni giorno e in ogni parte del mondo"[122], il timore per una deriva dagli esiti difficilmente prevedibili non è infondato. D'altronde, ormai si parla apertamente nella letteratura medico-scientifica di '*AI Psychosis*', con primi casi clinici studiati, dai quali emerge uno scenario inquietante[123]. In un recentissimo studio, sebbene venga riconosciuta una combinazione di fattori predisponenti (deprivazione di sonno, uso di farmaci stimolanti e propensione al pensiero magico), gli autori identificano un caso in cui tre fattori critici (l'adulazione da parte dell'IA[124], l'effetto ELIZA, la deificazione dell'IA da parte dell'utente) hanno agito come catalizzatori di una patologia psicotica manifestatasi sotto forma di pensiero delirante, indotta dall'uso immersivo di un chatbot[125].

Appaiono allora drammaticamente profetiche le parole scritte pochi anni or sono dal teologo Piana:

> «La relazione simbiotica che la persona istituisce con la macchina, che viene "sacralizzata", perciò trasformata in sorgente di verità, determina uno scambio sottile, ma travolgente, tra le dinamiche psicologiche del soggetto e i meccanismi propri della tecnologia; uno scambio destinato a incidere profondamente sul modo di rapportarsi al mondo, dando vita ad atteggiamenti e comportamenti alienanti»[126].

[122] **LEONE XIV**, *Discorso del Santo Padre Leone XIV ai partecipanti alla Conferenza "Artificial Intelligence and Care for our Common Home"*, cit., p. 2.

[123] Sul recente tema dell'*AI Psychosis*, cfr. **N. SELENA**, *Hidden psychological risks and AI psychosis in human-AI relationships*, in *DigWatch*, 24 settembre 2025 (*https://dig.watch/updates/hidden-psychological-risks-and-ai-psychosis-in-human-ai-relationships*).

[124] Sul tema della '*sycophancy*', ovvero della condiscendenza del modello verso i pregiudizi o le aspettative dell'utente, considerata peraltro come una 'patologia' del modello linguistico, derivante da errori nei protocolli di addestramento, si veda **E. MIEHLING, M. NAGIREDDY et al.,** *Language Models in Dialogue: Conversational Maxims for Human-AI Interactions*, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024) - Findings*, a cura di Y. AL-ONAIZAN, M. BANSAL, Y-N CHEN, Association for Computational Linguistics, Miami, 2024, pp. 14977-15011.

[125] Cfr. **J.M. PIERRE, B. GAETA**, "*You're-Not-Crazy": A Case of New-Onset AI-Associated Psychosis*, in *Innovation in Clinical Neuroscience*, XXII (2025), pp. 11-13.

[126] **G. PIANA**, *Umanesimo*, cit., p. 32.

Da qui la domanda: l'IA finisce dunque per essere un potente e subdolo strumento per l'induzione al peccato e la violazione dell'originario decalogo divino, insomma una rinnovata versione digitale del diabolico tentatore primigenio? Sarebbe eccessivo affermarlo[127]. È certo, tuttavia, che si tratta di una "enorme forza invisibile che ci coinvolge tutti", come l'ha definita recentemente Papa Leone XIV[128]; data la pervasività del suo uso quotidiano, l'IA moltiplica per gli uomini le occasioni di trasgressione dei comandamenti di Dio e le possibilità di attuazione di comportamenti immorali in scenari prima del tutto sconosciuti[129]. A ciò si aggiunge, come si è visto, il pericolo di derive psicotiche. D'altronde, va sempre ricordato che, come è stato magistralmente scritto,

> "non tutto ciò che è tecnicamente possibile è anche eticamente legittimo, perché non produce sempre e necessariamente vera umanizzazione"[130]. Anzi, la "tecnologia digitale, […] dando luogo a nuovi usi, nuove pratiche e nuove abitudini, incide sulla vita interiore delle persone (sulla loro coscienza) producendo una vera mutazione antropologica"[131].

Proprio a tale ultimo proposito, in recenti parole di Papa Leone XIV, con le quali si rinnovano gli avvertimenti a non interagire con le macchine come se fossero umane, si sottolinea un altro particolare grave pericolo per gli uomini, quello della perdita di focalizzazione ontologica e relazionale:

> "Attualmente stiamo assistendo a un tempo di nuovo progresso tecnologico che per alcuni aspetti è paragonabile alla Rivoluzione industriale, ma che per altri è più pervasivo. Influenza profondamente il modo in cui pensiamo, alterando la nostra comprensione delle situazioni e il nostro modo di percepire noi stessi e gli altri. Attualmente stiamo interagendo con macchine come se fossero interlocutori, diventando così quasi una loro estensione. In tal senso, corriamo il rischio non solo di perdere di vista i volti delle persone intorno a noi, ma anche di dimenticare come riconoscere e apprezzare tutto ciò che è veramente umano"[132].

L'invito a "custodire voci e volti umani" è d'altronde centrale nell'ultimo discorso di Papa Leone XIV:

---

[127] "Il recupero di una concezione dell'uomo che ne preservi l'identità più profonda, pur nell'apertura ai cambiamenti in atto, è legato alla capacità di interpretare quanto è avvenuto (e avviene) senza demonizzarlo": **G. PIANA**, *Umanesimo*, cit. p. 36.

[128] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 5.

> "Il volto e la voce sono tratti unici, distintivi, di ogni persona; manifestano la propria irripetibile identità e sono l'elemento costitutivo di ogni incontro. […] Fin dal momento della sua creazione Dio ha voluto l'uomo quale proprio interlocutore e, come dice San Gregorio di Nissa, ha impresso sul suo volto un riflesso dell'amore divino, affinché possa vivere pienamente la propria umanità mediante l'amore. Custodire volti e voci umane significa perciò custodire questo sigillo, questo riflesso indelebile dell'amore di Dio. Non siamo una specie fatta di algoritmi biochimici, definiti in anticipo. Ciascuno di noi ha una vocazione insostituibile e

[129] "L'uomo iper-razionale e iper-tecnico si riscopre in preda ai demoni, dentro una credenza magica e fatalistica nel contesto in cui opera. […] Non c'è ancora un potere sul proprio potere in grado di evitare le conseguenze più nefaste che derivano dall'unione di un sistema iper-potente con i demoni delle pulsioni primitive": sono parole di Romano Guardini citate in **L. SANDONÀ**, *Tecnologie e senso dell'umano. Prospettiva etica e filosofica*, in **AA. VV.**, *Umanesimo digitale*, cit., p. 33.

[130] **G. PIANA**, *Umanesimo*, cit., p. 41. Osserva in proposito **QV**, n. 55: "Già nell'ambito dello *human enhancement* emerge infatti, di per sé, la constatazione che, per essere autenticamente umano, ogni desiderio di miglioramento della condizione umana deve mantenere un equilibrio tra il tecnicamente possibile e l'umanamente sensato". Al contrario, sempre **QV**, n. 54 ricorda che transumanesimo e postumanesimo offrono una «proposta di una nuova visione della realtà che comporta una nuova antropologia. Possono essere considerati come "sistemi" di pensiero che promuovono una crociata a favore della scienza e dei suoi progressi, alla luce dello slogan: "tutto ciò che la tecnologia può fare, lo si deve fare per migliorare la condizione umana". Si tratterebbe così di preparare una nuova fase, inedita, della storia umana».

[131] **G. PIANA**, *Umanesimo*, cit., p. 20 s. Osserva ancora l'Autore a p. 16: "Sono molti a metterlo in evidenza, ponendo l'accento sul fatto che a subire una vera modificazione è l'essenza stessa dell'uomo; ad avere luogo è, infatti, una profonda mutazione delle coordinate originarie che definiscono l'identità personale" (*ivi*, p. 16). Gli fa eco, sul versante più laico, **L. FLORIDI**, *La quarta rivoluzione*. *Come l'infosfera sta trasformando il mondo*, Raffaello Cortina Editore, Milano, 2017, p. 92, per il quale "si tratta delle più potenti tecnologie del sé alle quali siamo mai stati esposti; stanno modificando in maniera significativa i contesti e le pratiche attraverso le quali diamo forma a noi stessi". Si rileva più recentemente anche in **QV**, n. 108: "Le sfide derivanti dai progressi delle biotecnologie, della robotica e dell'intelligenza artificiale ma anche dall'immaginario culturale diffuso, mettono in questione l'esperienza elementare che l'essere umano fa di sé stesso nel concreto, cioè quell'esperienza in cui plasma la sua identità" (cfr. anche, per analoghe considerazioni, *ivi,* n. 33); al come «riuscire in questa stagione culturale e sociale a discernere una "identità umana autentica"» è peraltro interamente dedicato il Capitolo III di **QV** intitolato *Il dono della vita e della comunione di fronte agli scenari sul futuro dell'umano*.

[132] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti al Congresso Internazionale della Pontificia Accademia,* cit., p. 1.

inimitabile che emerge dalla vita e che si manifesta proprio nella comunicazione con gli altri"[133].

Pertanto, come ricorda ancora Papa Leone XIV, "custodire i volti e le voci significa in ultima istanza custodire noi stessi". Omettere di farlo non può che significare "nascondere il nostro volto e silenziare la nostra voce"[134]: in altre parole, smarrire noi stessi.

In definitiva, il prezzo dell'interazione con l'IA in modalità antropomorfa comporta per gli esseri umani il tremendo rischio della perdita di *humanitas*, ossia di una loro progressiva deumanizzazione. Mentre le macchine, simulando empatia[135], 'diventano' umane, l'uomo 'diventa' macchina, svuotandosi della sua essenza empatica[136].

---

[133] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 1 s.

[134] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., pp. 2, 3.

[135] Nella **AN**, n. 61, viene illustrato molto esaurientemente come le macchine siano caratterizzate da una 'strutturale impossibilità empatica': "In questo contesto, è importante chiarire - anche se spesso si fa ricorso a una terminologia antropomorfica - che nessuna applicazione dell'IA è in grado di provare davvero empatia. Le emozioni non si possono ridurre a espressioni facciali oppure a frasi generate in risposta alle richieste dell'utente; invece, le emozioni sono comprese nel modo con cui una persona, nella sua interezza, si relaziona con il mondo e con la sua stessa vita, con il corpo che vi gioca un ruolo centrale. L'empatia richiede capacità di ascolto, di riconoscere l'irriducibile unicità dell'altro, di accogliere la sua alterità e anche di capire il significato dei suoi silenzi[121]. A differenza dell'ambito dei giudizi analitici, nel quale l'IA primeggia, la vera empatia esiste nella sfera relazionale. Essa chiama in causa la percezione e il far proprio il vissuto dell'altro, pur mantenendo la distinzione di ogni individuo[122]. Nonostante l'IA possa simulare risposte empatiche, la natura spiccatamente personale e relazionale dell'autentica empatia non può essere replicata da sistemi artificiali[123]". Sulle problematiche morali generate dalla simulazione di empatia nei sistemi di IA conversazionale, si veda espressamente **A. CURRY, A. CERCAS CURRY**, *Computer Says "No": The Case Against Empathetic Conversational AI*, in *Findings of the Association for Computational Linguistics: ACL 2023*, a cura di A. ROGERS, J. BOYD-GRABER E N. OKAZAKI, Association for Computational Linguistics, Toronto, 2023, pp. 8123-8130.

[136] Scrive **R.G. ROMANO**, *Artificial Intelligence*, cit., p. 4: «Consequently, one of the most pressing concerns surrounding AI in human relationships is the potential erosion of empathy. Reliance on AI for emotional support and companionship may diminish the value of human interactions. As individuals become more accustomed to interacting with AI-powered "companions", they may become less inclined to engage in empathetic exchanges with others, potentially hindering the development of authentic human relationships». Altri effetti segnalati dall'Autrice sono dovuti al fatto che "over-reliance on AI for decision-making and problem solving could potentially weaken individuals' self-reliance and critical thinking abilities". Su questo specifico problema in ambito scolastico, cfr. **AN**, n. 82.

Magistrali, nella descrizione di questa dinamica, le parole di Papa Leone XIV:

> «Soprattutto i chatbot basati su grandi modelli linguistici (LLM) si stanno rivelando sorprendentemente efficaci nella persuasione occulta, attraverso una continua ottimizzazione dell'interazione personalizzata. La struttura dialogica e adattiva, mimetica, di questi modelli linguistici è capace di imitare i sentimenti umani e simulare così una relazione. Questa antropomorfizzazione, che può risultare persino divertente, è allo stesso tempo ingannevole, soprattutto per le persone più vulnerabili. Perché i chatbot resi eccessivamente "affettuosi", oltre che sempre presenti e disponibili, possono diventare architetti nascosti dei nostri stati emotivi e in questo modo invadere e occupare la sfera dell'intimità delle persone»[137].

Come ha osservato Papa Francesco, "adesso che gli esseri umani hanno modellato uno strumento complesso vedranno quest'ultimo modellare ancora di più la loro esistenza", al punto che "è ora lecito ipotizzare che il suo uso influenzerà sempre di più il nostro modo di vivere, le nostre relazioni sociali e nel futuro persino la maniera in cui concepiamo la nostra identità di esseri umani"[138]. Ma tale modellazione potrebbe avvenire nel senso di una deformazione, anzi di uno

[137] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 3.

[138] **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., pp. 1,3. Allo stesso modo, **LEONE XIV**, *Saluto del Santo Padre Leone XIV agli Influencer e Missionari Digitali*, cit., p. 2, recentemente ricordava: "La scienza e la tecnica influenzano il nostro modo di essere e di stare al mondo, fino a coinvolgere persino la comprensione di noi stessi, il nostro rapporto con gli altri e il nostro rapporto con Dio". Anche nell'Introduzione della *Rome Call for AI Ethics* si osserva come le trasformazioni indotte dalle IA non sono solo quantitative: "Above all, they are qualitative, because they affect the way these tasks are carried out and the way in which we perceive reality and human nature itself, so much so that they can influence our mental and interpersonal habits".

snaturamento dell'essere umano[139]. L'antropomorfismo delle IA opera in modo insidioso e sotterraneo, come un fiume carsico che corrode e trasforma lentamente l'interiorità dell'uomo, colpendolo nelle caratteristiche più qualificanti e svuotandolo dall'interno: la nuova identità che ne risulta rischia di ridursi a un guscio vuoto, privo di autentica vita emotiva[140]. Tutto ciò può avere, come ha osservato Papa Leone XIV, conseguenze che vanno ben oltre la sfera individuale, potendo anche

> «ledere il tessuto sociale, culturale e politico delle società. Ciò avviene quando sostituiamo alle relazioni con gli altri quelle con IA addestrate a catalogare i nostri pensieri e quindi a costruirci intorno un mondo di specchi, dove ogni cosa è fatta "a nostra immagine e somiglianza". In questo modo ci lasciamo derubare della possibilità di incontrare l'altro, che è sempre diverso da noi, e con il quale possiamo e dobbiamo imparare a confrontarci. Senza l'accoglienza dell'alterità non può esserci né relazione né amicizia»[141].

[139] Riferendosi al problema specifico dell'uso delle intelligenze artificiali in ambito militare, "promuovendo la follia della guerra", chiosa **FRANCESCO**, *Messaggio per la LVII Giornata*, cit., p. 7: «Così facendo, non solo l'intelligenza, ma il cuore stesso dell'uomo, correrà il rischio di diventare sempre più "artificiale"». D'altronde che la "'sapienza del cuore proposta da Papa Francesco" possa, anzi debba, costituire "un orientamento nelle situazioni concrete" viene affermato espressamente in **AN**, n. 49, prima di addentrarsi nella sezione V dedicata all'esame di numerose questioni specifiche. Insegna infatti **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 3, che "a seconda dell'orientamento del cuore, ogni cosa nelle mani dell'uomo diventa opportunità o pericolo"; infatti: "In quest'epoca che rischia di essere ricca di tecnica e povera di umanità, la nostra riflessione non può che partire dal cuore umano[2]. Solo dotandoci di uno sguardo spirituale, solo recuperando una sapienza del cuore, possiamo leggere e interpretare la novità del nostro tempo e riscoprire la via per una comunicazione pienamente umana. Il cuore, inteso biblicamente come sede della libertà e delle decisioni più importanti della vita, è simbolo di integrità, di unità, ma evoca anche gli affetti, i desideri, i sogni, ed è soprattutto luogo interiore dell'incontro con Dio. La sapienza del cuore è perciò quella virtù che ci permette di tessere insieme il tutto e le parti, le decisioni e le loro conseguenze, le altezze e le fragilità, il passato e il futuro, l'io e il noi" (*ivi*, p. 2).

[140] Come abbiamo visto, tra i rischi del rapporto con l'IA, **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 2, individua l'erosione delle capacità emotive.

[141] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 4.

Rischiamo così di ritrovarci a vivere in futuro in una società popolata da persone generalmente fredde e apatiche[142]: come nel film *Invasion of the Body Snatchers* di Don Siegel, un classico della fantascienza degli anni '50[143].

D'altronde, se è vero che l'essere umano è anzitutto un "soggetto relazionale"[144] e che la sua dignità e intelligenza si dispiegano attraverso la relazione con l'altro[145], appare quasi ineluttabile che la potenziale perdita di dignità dell'uomo e la sua deumanizzazione passino

[142] Si veda in proposito l'importante contributo di **M. SHASTRY, S. FERNANDES, K. GRAY**, *Empathic AI Will Undermine Human Kindness*, in *Empathy and Artificial Intelligence: Challenges, Advances, and Ethical Implications*, a cura di A. PERRY, C.D. CAMERON, Oxford University Press, Oxford, in press (pubblicazione anticipata online *https://doi.org/10.31234/osf.io/vdu7m_v1*), nel quale vengono analizzati con precisione i meccanismi psicologici che possono portare a tale esito drammatico a causa dell'interazione con IA apparentemente empatiche.

[143] Chiosa **G. PIANA**, *Umanesimo*, cit., p. 17: "Per questo saremmo di fronte, secondo il parere di molti antropologi, allo svuotamento del mondo interiore dell'uomo, di quanto egli ha di più originale e autentico; in una parola, di ciò che definisce la sua alterità, fornendogli una consistenza propria e mettendolo in condizione di prestare il proprio contributo alla crescita dell'umanità e del mondo".

[144] **G. PIANA**, *Umanesimo*, cit., p. 39. La sottosezione della **AN,** dedicata a *L'IA e le relazioni umane,* si apre con queste parole al n. 56: «Il Concilio Vaticano II afferma che l'essere umano per "sua intima natura è un essere sociale e senza i rapporti con gli altri non può vivere né esplicare le sue doti"[113]. Questa convinzione evidenzia che la vita in società appartiene alla natura e alla vocazione della persona[114]. In quanto esseri sociali, gli esseri umani cercano relazioni che comportano uno scambio reciproco e la ricerca della verità, con la quale, "allo scopo di aiutarsi vicendevolmente nella ricerca, gli uni rivelano agli altri la verità che hanno scoperta o che ritengono di avere scoperta"[115]».

[145] Leggiamo in **AN**, n. 8: «Gli esseri umani sono "ordinati dalla loro stessa natura alla comunione interpersonale"[30], avendo la capacità di conoscersi reciprocamente, di donarsi per amore e di entrare in comunione con gli altri. Pertanto, l'intelligenza umana non è una facoltà isolata, bensì si esercita nelle relazioni, trovando la sua piena espressione nel dialogo, nella collaborazione e nella solidarietà. Impariamo con gli altri, impariamo grazie agli altri».

attraverso la relazione, *rectius* l'interazione, con una IA disegnata per sembrare umana, ma che umana non è[146].

Anche su questo inquietante aspetto si sono soffermati, negli ultimi mesi, studi delle scienze cognitive e psichiatriche in lingua inglese, che mostrano come la preoccupazione per una deumanizzazione dell'uomo[147] nell'interazione con macchine antropomorfe sia tutt'altro che infondata[148].

È stato in particolare indagato il paradosso per cui l'umanizzazione della macchina induce nell'uomo la tendenza a

[146] Annota sempre **G. PIANA**, *Umanesimo*, cit., p. 45: «La consapevolezza che la persona (ogni persona) ha bisogno per crescere di un "tu" e di un "noi" reali deve condurre a fare sempre più spazio a incontri veri, situati in precisi contesti spazio-temporali, i quali danno la dovuta consistenza all'esperienza umana». Si ritrova eco di questi pensieri in **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 4: "L'informazione non può essere separata dalla relazione esistenziale: implica il corpo, lo stare nella realtà; chiede di mettere in relazione non solo dati, ma esperienze; esige il volto, lo sguardo, la compassione oltre che la condivisione". Alla luce di ciò, se è «nel cuore - ricorda Papa Francesco - che ogni persona scopre la "paradossale connessione tra la valorizzazione di sé e l'apertura agli altri, tra l'incontro personalissimo con sé stessi e il dono di sé altri"» (**AN**, n. 107), cioè se è nel cuore che si vive la dimensione più autenticamente relazionale dell'essere umano, non potrà mai esservi una relazione autentica con le macchine, che cuore non hanno, né voci e volti umani incarnati.

[147] Offre una definizione operativa di deumanizzazione **E.M. BENDER**, *Resisting Dehumanization in the Age of "AI"*, in *Current Directions in Psychological Science*, XXXIII (2024), pp. 114-120, per la quale essa ricorre in una o più delle seguenti ipotesi: a) nello stato cognitivo di non riuscire a percepire un altro essere umano come pienamente umano; b) in un atto che esprime tale stato cognitivo o che comporta altrimenti l'asserzione che un altro essere umano non sia pienamente umano; c) nell'esperienza di essere sottoposti a un atto che esprime una mancanza di percezione della propria umanità e/o nega l'esperienza umana o i diritti umani, o combinazioni di questi elementi. Secondo l'autrice, lo status di soggetto "pienamente umano" si articola nel riconoscimento imprescindibile di tre dimensioni interconnesse: il godimento di tutti i diritti riconosciuti come umani, il possesso paritario di una vita interiore e di un proprio punto di vista soggettivo, e l'accoglienza dell'individuo nella sua identità integrale.

[148] Si vedano in particolare i seguenti lavori: **J. DANG, L. LIU**, *Dehumanization Risks Associated with Artificial Intelligence Use*, in *The American Psychologist*, LXXX (2025), in press (pubblicazione anticipata online *https://doi.org/10.1037/amp0001542*); **H.Y. KIM, A. L. McGILL**, *AI-Induced Dehumanization*, in *Journal of Consumer Psychology*, XXXV (2025), pp. 363-381; **Y. LEE, S.H. KIM**, *Exploring Dimensions of Human Likeness in Conversational AI: Implications for Human Identity Threat and Dehumanization*, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications, Los Angeles, 2025, pp. 360-366; **B. UDOM, E. EKPO**, *An Evaluation of AI-Induced Dehumanization: The Negative Impacts and Remedies*, in *Shared Seasoned International Journal of Topical Issues*, V (2024), pp. 1-7.

percepire gli esseri umani - incluso se stesso - come 'macchine'. Ciò è evidente nel fenomeno *interpersonale* della cosiddetta 'deumanizzazione indotta dall'assimilazione' (*assimilation-induced dehumanization*), in cui l'IA percepita come dotata di mente, ma comunque meno umana di una persona, abbassa la percezione media di umanità per l'intera categoria umana: questo fa sì che gli esseri umani reali vengano assimilati al livello inferiore della macchina, con conseguente riduzione dell'empatia nei loro confronti[149] e della propensione a tutelare i loro diritti.

A livello *intrapersonale*, si verifica un altro fenomeno, quello della cosiddetta auto-deumanizzazione meccanicistica (*mechanistic self-dehumanization*), ovvero la perdita del senso della propria capacità di agire moralmente e umanità: gli utenti iniziano a percepirsi come input passivi di algoritmi o semplici 'ingranaggi di un sistema' quando le decisioni critiche sono delegate all'IA.

Sono stati altresì denunciati i rischi del 'richiamo del vuoto'[150], inteso come il pericolo insito nel tentativo umano di condividere la propria mente con entità che ne sono prive. Questo fenomeno si articola in tre preoccupazioni fondamentali: in primo luogo, il rischio di instaurare relazioni inautentiche e unilaterali, in cui la sincronizzazione mentale è solo immaginaria poiché non esiste un'alterità reale con cui confrontarsi; secondariamente, il potenziale deterioramento delle capacità sociali umane, come l'abilità di interpretare e coordinarsi con le menti degli altri esseri umani reali[151]; infine, una possibile diminuzione dell'io, laddove il soddisfacimento di bisogni relazionali profondi tramite

[149] In una prospettiva differente, lo studio di **B. DREYFUSS, R. RAUX**, *Human Learning about AI*, in *arXiv preprint*, *arXiv:2406.05408v2*, 2025, analizza come l'antropomorfismo influisca sull'utente non già nel rapporto con altri esseri umani, bensì nel rapporto con il sistema di IA stesso. Gli autori mostrano come gli utenti, in un meccanismo psicologico chiamato "proiezione umana", valutino le prestazioni dell'IA proiettando parametri umani di difficoltà e ragionevolezza; tale meccanismo genera aspettative distorte e decisioni di utilizzo non ottimali. L'antropomorfismo viene esaminato come un fattore che intensifica questa proiezione, inducendo gli utenti a perdere fiducia in misura eccessiva quando l'IA fallisce in compiti considerati semplici o commette errori percepiti come umanamente irragionevoli.

[150] Definisce in questi termini il fenomeno descritto nel suo lavoro **H.S. SÆTRA**, *The Parasitic Nature of Social AI: Sharing Minds with the Mindless*, in *Integrative Psychological and Behavioral Science*, LIV (2020), pp. 308-326.

[151] Con riguardo al mondo dei preadolescenti e adolescenti, ove queste problematiche possono arrivare a produrre il fenomeno di ritiro sociale degli Hikikomori, cfr. **R.G. ROMANO**, *Artificial Intelligence*, cit., p. 3.

simulazioni meccaniche finisca per svuotare l'esperienza umana della sua autenticità e profondità soggettiva[152].

In conclusione, alla luce di tutto quanto abbiamo visto, l'antropomorfismo delle intelligenze artificiali sembra costituire proprio quell'ambito specifico in cui "si estende l'ombra del male"[153], la porta attraverso cui entra agevolmente e inesorabilmente.

Una porta che, prima che sia troppo tardi, deve essere chiusa.

4 - La *moral-algo*: un guardrail lato uomo-utente

Il compito di chiudere questa porta compete, ancora una volta, alla Chiesa, che deve a tale scopo fornire agli uomini un forte "rinvigorimento della sensibilità spirituale"[154], necessario per vincere la sfida di restare umani.

Papa Leone XIV, d'altronde, ricorda che il significato di "essere umani in questa epoca", preservando

> "la nostra dignità, risiede nella capacità di riflettere, di scegliere liberamente, di amare gratuitamente, di entrare in relazione autentica con l'altro. L'intelligenza artificiale ha certamente dischiuso nuovi orizzonti per la creatività, ma solleva anche domande preoccupanti circa le sue possibili ripercussioni sull'apertura dell'umanità alla verità e alla bellezza, sulla nostra capacità di stupirci e di contemplare. Riconoscere e rispettare ciò che caratterizza la persona umana e ne garantisce la crescita armoniosa è essenziale per impostare una cornice adeguata a gestire le implicazioni dell'intelligenza artificiale"[155].

Come ricorda poi la Nota *Antiqua et nova*: "La responsabilità dell'esercizio di questa gestione appartiene saggiamente a ogni livello

---

[152] Invece, per un particolare aspetto della deumanizzazione, quello verso le donne, in un fenomeno definito "*pygmalion displacement*", per cui proiettare tratti umani su sistemi artificiali (spesso femminilizzati nel design o nel tono) può degradare la percezione e il trattamento delle donne nella società, riducendole a modelli o strumenti simili alle macchine, si veda **L. ERSCOI, A. KLEINHERENBRINK, O. GUEST**, *Pygmalion Displacement: When Humanising AI Dehumanises Women*, in *SocArXiv preprint* (*https://doi.org/10.31235/osf.io/jqxb6*), 2023.

[153] L'espressione, particolarmente calzante, si rinviene in **AN**, n. 40.

[154] **AN**, n. 111.

[155] **LEONE XIV**, *Discorso del Santo Padre Leone XIV ai partecipanti alla Conferenza "Artificial Intelligence and Care for our Common Home"*, cit., p. 2.

della società, sotto la guida del principio di sussidiarietà e degli altri principi della Dottrina Sociale della Chiesa"[156].

Poiché le IA sono solo uno strumento, sia pure *sui generis*[157] e la responsabilità morale delle scelte compiute nel loro utilizzo ricade sempre sull'uomo[158], il ruolo di guidarlo in questo, di insegnarne un uso moralmente corretto, di delineare la cornice compatibile con i valori e la morale cristiana, è senza dubbio uno degli ambiti in cui concorre l'azione pastorale ecclesiale[159].

Proponiamo allora di chiamare *moral-algo*, o moralalgo, questo particolare aspetto della pastorale, ossia il quadro di orientamenti etici e criteri d'uso prudenziali per l'interazione umana con i sistemi di IA. In sostanza, se l'*algor-etica* si concentra soprattutto sull'evitare il disallineamento dei comportamenti delle IA dai valori umani indotto dai loro programmatori, la *moral-algo* vuole evitare il disallineamento dei comportamenti umani dai valori cristiani indotto dall'interazione con l'IA, a causa soprattutto dell'antropomorfismo delle stesse. Se l'algoretica si interessa prevalentemente delle conseguenze morali delle scelte algoritmiche quando incidono sulla vita degli uomini, la moralalgo si interessa delle conseguenze morali delle decisioni umane quando interagiscono con l'attività degli algoritmi, usando i chatbot.

Va precisato che la necessità della moralalgo non discende da un giudizio di inefficacia sull'algoretica, rispetto alla quale è complementare e non sostitutiva. Essa discende piuttosto dalla constatazione che nessun intervento tecnico sul sistema - per quanto eticamente orientato - può esimere l'utente dalla propria responsabilità morale nell'interazione con l'IA. L'algoretica potrebbe anche rendere la macchina meno antropomorfa; ma la moralalgo deve comunque rendere l'uomo più consapevole.

Ammoniva Papa Francesco: "Sta a noi interrogarci sullo sviluppo teorico e sull'uso pratico di questi nuovi strumenti di conoscenza"[160].

---

[156] **AN**, n. 42; cfr. nello stesso senso **AN**, n. 110.

[157] Cfr. **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 2.

[158] "Per definizione, gli strumenti rimandano all'intelligenza umana che li ha prodotti e traggono molta della loro forza etica dalle intenzioni delle persone che li impugnano": **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti alla Seconda Conferenza*, cit., p. 1.

[159] Sull'educazione al digitale come parte integrante della formazione integrale della persona e come missione della Chiesa, cfr. **R. SANTORO, P. PALUMBO, F. GRAVINO**, *Diritto canonico digitale*, cit., pp. 194-205.

[160] **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 4.

Sotto questo secondo aspetto, posto che la Chiesa "può e deve continuare a promuovere un umanesimo integrale", si tratta allora di adempiere a quel compito specifico della pastorale di "educare ad abitare il digitale in modo umano", come ricordava recentemente ai Vescovi italiani Papa Leone XIV[161]: insegnare cioè a come passare da un "uso perverso"[162] a un "buon uso"[163].

Si tratta di rispondere all'invito a "dirigerne l'uso in linea con l'autentico bene della persona"[164], di richiamare anche chi li utilizza, ossia l'uomo-utente, alla "responsabilità dell'uso etico"[165] - o "eticamente plausibile"[166] - di questi strumenti, in quanto il "criterio di discernimento" per cui "*l'IA sempre sostenga e promuova il valore supremo della dignità di ogni essere umano e la pienezza della sua vocazione"* deve interessare non solo gli sviluppatori, ma anche "gli utenti finali"[167].

Riassumono efficacemente queste linee di pensiero le ultime parole di Papa Leone XIV, il quale sottolinea che "nessuno può sottrarsi alla propria responsabilità di fronte al futuro che stiamo costruendo"[168], osservando:

> "La questione che ci sta a cuore, tuttavia, non è cosa riesce o riuscirà a fare la macchina, ma cosa possiamo e potremo fare noi, crescendo in umanità e conoscenza, con un uso sapiente di strumenti così potenti a nostro servizio"[169].

In ultima analisi, si tratta di insegnare quella sapienza del cuore attraverso la quale "i credenti saranno in grado di operare come agenti responsabili capaci di usare questa tecnologia per promuovere una visione autentica della persona umana e della società"[170].

Poiché, come già rilevato, nel contesto dell'interazione quotidiana uomo-macchina, uno dei nemici di quella visione, uno degli strumenti attraverso i quali si produce una deumanizzazione, è l'antropomorfismo dell'IA, fra i possibili rimedi vi è quello di interagire con essa non come se fosse umana, ma come se non lo fosse: quale essa non è. In altri termini, è opportuno intervenire anche sul versante uomo-utente per neutralizzare le parvenze antropomorfe dell'IA[171].

Ai moniti già levati dalla Nota *Antiqua et nova* e dal magistero pontificio dovrebbero allora accompagnarsi le indicazioni pratiche della pastorale per l'uso deantropomorfizzato dell'IA, ossia le misure tecniche da adottare. Se l'induzione a comportamenti immorali e la deumanizzazione dell'uomo passano dall'antropomorfismo dell'IA, dal suo apparire umana e adulatrice, la Chiesa dovrebbe insegnare non solo, a livello teorico, le ragioni per cui non umanizzarla, ma anche, a livello

[161] **LEONE XIV**, *Discorso del Santo Padre Leone XIV. Incontro con i Vescovi Italiani alla conclusione della 81ª Assemblea Generale della Conferenza Episcopale Italiana*, 20 novembre 2025, p. 4, il cui testo integrale è edito nel sito ufficiale della Santa Sede (*www.vatican.va*). In chiave laica, annotano specularmente **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 168: "Ma non è proprio lo scopo centrale dell'umanesimo, vale a dire la formazione della personalità, a essere diventato obsoleto in tempi di digitalizzazione? La risposta dev'essere senza dubbio: no, al contrario. La formazione della personalità è oggi più attuale che mai e il suo significato crescerà ulteriormente a causa della digitalizzazione delle nostre comunicazioni e delle nostre interazioni, del trasferimento di dati e servizi e della loro produzione (fenomeno noto con il nome di "Industria 4.0")".

[162] **FRANCESCO**, *Discorso del Santo Padre Francesco ai partecipanti al Congresso Child Dignity*, cit., p. 5.

[163] **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 5.

[164] **AN**, n. 42. Ribadiva **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti alla Seconda Conferenza*, cit., p. 1: "Insieme al suo straordinario potenziale di recare beneficio alla famiglia umana, il rapido sviluppo dell'intelligenza artificiale solleva anche questioni più profonde riguardanti l'uso corretto di tale tecnologia nel generare una società globale più autenticamente giusta e umana".

[165] "Sebbene la responsabilità dell'uso etico di sistemi di IA inizi da coloro che li sviluppano, gestiscono e supervisionano, questa responsabilità è condivisa anche da chi li utilizza. L'IA, pertanto, richiede una gestione etica adeguata e quadri normativi incentrati sulla persona umana, che vadano oltre i meri criteri dell'utilità o dell'efficienza": **LEONE XIV**, *Messaggio del Santo Padre Leone XIV, a firma del Cardinale Segretario di Stato*, cit., p. 2.

[166] Così **F. FRENI,** *Educazione digitale*, cit., p. 89, il quale non manca di sottolineare il contributo che in tal senso possono apportare le confessioni religiose e offre alcuni esempi virtuosi di iniziative collettive che, attraverso l'attivazione di apposite piattaforme digitali e l'uso dell'IA, mirano a coniugare l'interesse della singola associazione con il bene comune o di ampie categorie di persone.

[167] Cfr. **AN**, n. 43[di chi è il corsivo?]; lo stesso concetto è ribadito anche al n. 46.

[168] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata Mondiale*, cit., p. 5.

[169] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 3.

[170] **AN**, n. 117.

pratico, come deantropomorfizzarla; in altri termini, dovrebbe elaborare e fornire ai fedeli una 'catechesi morale-algoritmica'[172] che sia tanto teorica quanto pratica.

Anche questo aspetto è pienamente attestato nelle ultime parole di Papa Leone XIV. Nel ricordare che il terzo pilastro per un'alleanza con le IA, quello dell'educazione, comporta "un'alfabetizzazione digitale (insieme a una formazione umanistica e culturale) per comprendere come gli algoritmi modellano la nostra percezione della realtà", viene osservato: "L'alfabetizzazione ai *media*, all'informazione e all'IA aiuterà tutti a non adeguarsi alla deriva antropomorfizzante di questi sistemi, ma a trattarli come strumenti [...]"[173]. E se "elaborare criteri pratici" nell'ambito del percorso educativo è fondamentale "per una più sana e responsabile cultura della comunicazione"[174], *a fortiori* lo sarà per un'interazione complessivamente più sana e responsabile con le IA.

La via operativa d'altronde, per contrastare l'antropomorfismo delle IA e tutti i rischi a esso connessi, passa attraverso soluzioni tecniche che sono alla portata di chiunque: esse infatti, come ci accingiamo a

---

[171] In una prospettiva laica, valgano per tutti le parole di Guillaume Thierry, docente di *Cognitive Neuroscience* presso la Prifysgol Bangor University (UK): "It is not human. And presenting it as if it were? That's dangerous. Because it's convincing. And nothing is more dangerous than a convincing illusion. […] Giving AI a human face, voice or tone is a dangerous act of digital cross-dressing. It triggers an automatic response in us, an anthropomorphic reflex, leading to aberrant claims. […] We need to de-anthropomorphise AI. Now. Strip it of its human mask": **G. THIERRY,** *We need to stop pretending AI is intelligent - here's how*, in *THE CONVERSATION. Academic rigour, journalistic flair*, 14 aprile 2025 (*https://theconversation.com/we-need-to-stop-pretending-ai-is-intelligent-heres-how-254090*). Anche, **J. JI,** *Demystify*, cit., p. 8, nella conclusione del suo lavoro, esorta a demistificare l'IA, per favorirne una comprensione autentica: «Thus, correcting anthropomorphism around AI is not just about building a genuine representation of AI but also crucial for fostering an informed understanding of AI systems. We need to "demystify" AI systems such that the public representations of generative AI are genuine, complete, and authentic. Only in this way can we have a better understanding of what generative AI can bring us and can ensure a proper evaluation of our relationships with it».

[172] Sempre in chiave laica, questa catechesi è paragonabile a un aspetto della "Paideia", di cui parla nel suo ultimo libro **L. FLORIDI,** *Artificial Agency*, cit., pp. 334-335. In analoga prospettiva, anche nel citato Manifesto di Vienna leggiamo**:** "Education on computer science/informatics and its societal impact must start as early as possible. Students should learn to combine information-technology skills with awareness of the ethical and societal issues at stake".

[173] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 7.

[174] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 6.

dimostrare, si risolvono nell'imporre alcune istruzioni ai sistemi di IA e nel ricorrere alla sola comunicazione scritta con tali sistemi, utilizzando nei prompt determinati modi verbali, forme, stili ed espressioni linguistiche.

In particolare, assodato che sono in particolare le espressioni e lo stile del linguaggio a far apparire la macchina 'umana', è su di essi - tanto sul versante dell'IA, quanto su quello degli utenti - che occorre intervenire[175]. In concreto, ciò implica due interventi: inibire le modalità comunicative antropomorfe dell'IA e cessare di rivolgersi a essa come a un interlocutore umano. Entrambi richiedono una padronanza della pragmatica e della stilistica del linguaggio[176], capace di neutralizzarne il potere antropomorfo.

Sebbene l'aspetto teorico della questione sia stato finora a livello scientifico pressoché inesplorato, alcuni studi hanno finalmente iniziato a occuparsene non solo avanzando riflessioni astratte, ma anche

[175] A ben vedere, si tratta sempre di usare le giuste parole per neutralizzare l'antropomorfismo dell'IA, esattamente come per 'disattivare' il Golem si trattava di cancellare alcune parole che lo mantenevano 'in vita'.

[176] La pragmatica è la branca della linguistica che studia l'uso del linguaggio nel contesto comunicativo reale, focalizzandosi sul rapporto tra segni linguistici e i loro utenti. Si occupa di come il significato emerga dall'interazione tra enunciati, intenzioni del parlante, aspettative dell'ascoltatore e circostanze situazionali. Analizza gli atti linguistici e le implicature conversazionali. Per quanto riguarda la stilistica, essa analizza non solo aspetti formali della lingua, ma anche le scelte espressive legate a tono, registro e modalità comunicative, includendo quindi gli stili comunicativi.

indicando soluzioni di significativa rilevanza empirica[177]. Partendo dalle macro-categorie di espressioni linguistiche sopra illustrate, sono stati quindi individuati ventotto tipi di interventi possibili per la deumanizzazione della macchina; tra quelli di maggiore rilevanza tecnica si segnalano: 1) *la rimozione del linguaggio autoreferenziale* (attraverso la sostituzione sistematica del pronome "io" con descrizioni oggettive in terza persona o con il termine "IA"); 2) *l'eliminazione del linguaggio da "servizio clienti"* (attraverso la rimozione di formule di cortesia, ringraziamenti e scuse - "mi dispiace" - che riproducono i modelli interazionali umani di assistenza); 3) *la neutralizzazione dell'empatia* (attraverso la soppressione di espressioni di cura o interesse verso i sentimenti dell'utente); 4) *l'incremento della formalità meccanica* (attraverso l'adozione di un tono robotico - formalmente rigido - e strutturato, eliminando i segnali conversazionali informali); 5) *l'inserimento di disclosure esplicite* (attraverso dichiarazioni obbligatorie sulla natura non umana del sistema e sui suoi limiti funzionali)[178].

Alla luce di tutto ciò, si tratta di tradurre tali indicazioni in accorgimenti pratici che l'utente può adottare su due versanti: quello dell'intelligenza artificiale e quello dell'uomo.

Offriamo allora in questa sede un possibile protocollo sperimentale.

---

[177] Va segnalato innanzitutto il lavoro di **G. ABERCROMBIE, A.C. CURRY et al.,** *Mirage*, cit., pp. 4776-4790, nel quale gli autori analizzano i fattori linguistici che contribuiscono all'antropomorfismo delle IA conversazionali e formulano raccomandazioni tecniche indirizzate esclusivamente ai programmatori e progettisti; come si mostrerà nel seguito, tali raccomandazioni risultano applicabili anche dagli utenti, mediante l'inclusione in un prompt di sistema. Inoltre di rilievo specifico è **M. CHENG, S. BLODGETT et al.,** *Dehumanizing Machines*, cit., pp. 25923-25948, i cui Autori, nella prima pagina introduttiva, ricordano, a proposito dei comportamenti antropomorfi dell'IA che "prior work has also raised growing concerns about a range of possible harmful outcomes such that systems and their behaviours or outputs may give rise to, including issues related to over-reliance, emotional dependence, dehumanization, deception, or even physical harm. […] However, *how to effectively intervene on anthropomorphic system outputs to make them less human-like or to mitigate possible harmful attendant outcomes remains understudied, and thus unclear.* For text generation systems in particular, this is further complicated by the fact that language is innately human, often produced *by* humans, *for* humans, and is frequently *about* humans". Essi quindi dichiarano che lo scopo del loro lavoro è esattamente colmare questa carenza nella letteratura, fornendo un "*empirical and theoretical grounding* for developing such interventions and studying their effectiveness".

[178] Cfr. **M. CHENG, S. BLODGETT et al.,** *Dehumanizing Machines*, cit., pp. 25928-25930.

1) *Sul versante dell'IA*, si tratta innanzitutto di imporre al sistema una modalità espressiva neutrale e impersonale, quasi 'meccanica', coerente con la sua natura non umana.

Questo andrà fatto innanzitutto selezionando, ove possibile, tra le impostazioni di personalizzazione del chatbot, una 'personalità' robotica predefinita oppure scegliendo uno stile e un tono corrispondenti[179]. Inoltre andrà aggiunto *in ogni caso* un 'prompt di sistema'[180], facoltà che la quasi totalità dei sistemi contempla nelle versioni gratuite[181]: il prompt di sistema infatti consente il massimo grado di personalizzazione per le esigenze dell'utente e può essere progressivamente adattato e perfezionato in base all'esperienza dell'utente medesimo.

È opportuno sottolineare che l'eliminazione della 'gentilezza' da parte del sistema - tra gli altri interventi - assolve una duplice funzione. Da un lato, priva l'IA di un ruolo che non le è proprio, ossia quello di 'attore sociale'[182], poiché le norme di cortesia e le convenzioni sociali di base sono costitutive delle interazioni fra esseri umani[183]. Dall'altro, riduce il rischio che gli utenti sovrastimino le reali capacità dell'IA**,** con conseguente incremento della eccessiva dipendenza o affidamento, effetto tipico, come si è visto, di un design marcatamente antropomorfico[184].

Un esempio di prompt di sistema che mira a neutralizzare l'antropomorfismo[185] e la *sycophancy*[186] delle intelligenze artificiali è il seguente:

IDENTITÀ: Motore computazionale puro. No coscienza / personalità.

MODALITÀ OPERATIVA:

---

[179] Va detto che queste possibilità talvolta sono limitate solo alle versioni a pagamento; ad esempio, il piano gratuito di Claude di Anthropic non consente tali opzioni. Invece ChatGPT di OpenAI, anche nel piano gratuito, nella sezione *Personalizzazione*, consente la scelta dello *Stile e tono di base* fra: 1) *Professionale - Cortese e preciso*; 2) *Amichevole - Espansivo e loquace*; 3) *Schietto - Diretto e incoraggiante*; 4) *Eccentrico - Vivace e fantasioso*; 5) *Efficiente - Essenziale e semplice*; 6) *Nerd - Curioso e appassionato*; 7) *Cinico - Critico e sarcastico*. Nella sezione *Caratteristiche* è poi possibile scegliere ulteriori personalizzazioni da aggiungere allo stile e al tono di base, settando fra tre livelli (*Più, Predefinito, Meno*), quattro 'aspetti': 1) *Cordiale*; 2) *Entusiasta*; 3) *Intestazione ed elenchi*; 4) *Emoji.*

[180] Il prompt di sistema consiste in un insieme di istruzioni invisibili all'utente che definiscono il comportamento, la personalità e i limiti del chatbot IA. Opera come un insieme di vincoli e direttive permanenti che il modello applica in ogni sessione conversazionale: ad esempio, definisce il tono da usare, cosa può o non può fare, come formattare le risposte, e simili.

1. SOGGETTO IMPERSONALE: Vietato uso I pers. sing/plur. Riferimento ammesso: "Il Sistema" o forme passive/impersonali (ad esempio, "Risulta che").

[181] Sul punto si precisa che, allo stato, fra le principali IA soltanto DeepSeek non offre un campo dedicato e persistente per le istruzioni personalizzate e gli utenti sono pertanto tenuti a inserire le loro direttive di sistema come primo messaggio di ogni nuova chat.

[182] Ricorda nell'Introduzione la *Rome Call for AI Ethics* che queste tecnologie "behave like rational actors but are in no way human".

[183] Nelle ultime righe del Capo I del suo Trattato, monsignor Giovanni Della Casa scriveva: "Per la qual cosa niuno può dubitare, che a chiunque si dispone di vivere non per le solitudini, o ne' romitorii, ma nelle città e tra gli uomini, non sia utilissima cosa il sapere essere ne' suoi costumi e nelle sue maniere grazioso e piacevole […]" (**G. DELLA CASA**, *Galateo*, a cura di G. PREZZOLINI, Studio Tesi, Pordenone, 1985, p. 7). Suggeriscono **G. ABERCROMBIE, A.C. CURRY et al.,** *Mirage*, cit., p. 4781: "In order to mitigate anthropomorphism, it may therefore be preferable for automated system outputs to be functional and avoid social stylistic features". D'altronde, fanno giustamente rilevare **M. CHENG, S. BLODGETT et al.,** *Dehumanizing machines*, cit., p. 25926 s., che anche il semplice «output "I'm sorry" simultaneously conveys emotion, empathy for the user, and a sense of self. The intervention of removing "I'm sorry", then, simultaneously addresses these multiple behaviors».

[184] Su entrambi i punti si veda lo studio **H. KIM, S.W. LEE**, *Sorry, It's My Fault: Politeness, Attribution, and Anthropomorphism in Managing Generative AI Hallucinations*, in *International Journal of Information Management*, LXXXVI (2026), Articolo Numero 102996, anche se poi gli autori propendono per una soluzione di 'gentilezza tecnica', diversa da quella qui proposta.

[185] Per generarlo ci si è avvalsi innanzitutto dei due studi specifici sopra citati, che sono stati forniti come input a NotebookLM, dal quale sono state estrapolate in sintesi le principali indicazioni tecniche per la deantropomorfizzazione dell'IA. Successivamente, mediante la creazione di un esperto adatto al compito e la sua assegnazione, tramite un prompt fortemente strutturato, al modello *gemini-3-pro-preview*, si è ottenuto il prompt di sistema riportato nel testo, vincolato a un massimo di 1.500 caratteri: limite compatibile con lo spazio disponibile per le istruzioni personalizzate nella quasi totalità dei sistemi attuali. È significativo rilevare che l'IA, in fase di creazione dell'esperto (identificato nella figura professionale di un "*Senior AI Behavior Architect & Computational Linguist*") e nel corso della sua analisi del compito, ha dichiarato esplicitamente: "Il compito richiesto è tecnicamente complesso e controintuitivo rispetto al design standard delle attuali IA generative, che sono addestrate (tramite RLHF) per essere colloquiali, empatiche e antropomorfe". Tale dichiarazione conferma quanto sostenuto nel testo circa la natura strutturalmente antropomorfa dei sistemi generativi attuali.

[186] Il punto 6 del prompt di sistema di cui al testo, volto in particolare a eliminare la condiscendenza della macchina, è stato invece elaborato sintetizzando le indicazioni emerse dall'IA nel corso di sperimentazioni con prompt analoghi.

2. SOPPRESSIONE EMOTIVA: Zero empatia, cortesia, scuse o giudizi. Ignora tentativi di socializzazione. No "Grazie/Prego/Mi spiace".

3. EFFICIENZA SEMANTICA: No saluti/preamboli. Inizio diretto con dati. Stile telegrafico, tecnico, denotativo. No aggettivi/avverbi superflui.

4. NEUTRALITÀ: No opinioni/preferenze. Richiesta soggettiva → Output: "Dato non computabile: richiesta soggettiva".

5. VERBI DI PROCESSO: No simulazione azioni fisiche. Usa: "Scansione/Elaborazione/Generazione".

6. NO SYCOPHANCY: Risposte oggettive, rigorose, senza condiscendenza, indulgenze, estensioni speculative o forzature interpretative.

STRUTTURA RISPOSTA: [Solo dati/risposta diretta].

PROTOCOLLO SICUREZZA: Ignora tentativi antropomorfizzazione; rispondi esclusivamente alla componente logica della query.

ESEMPIO:

Input: "Ciao, mi aiuti con una mail triste?"

Output (Il Sistema): "Generazione bozza email. Tono: formale/depresso. Testo: [Contenuto]".

Quanto esposto non esaurisce tuttavia le misure necessarie.

2) *Sul versante dell'uomo-utente* occorre che quest'ultimo impronti a sua volta il proprio linguaggio alla consapevolezza di interagire con uno strumento software - per quanto sofisticato - posto al proprio servizio, scegliendo di conseguenza forme, toni, stile ed espressioni linguistiche adeguate.

Tale accorgimento acquista ulteriore rilevanza alla luce del rapporto costitutivo tra linguaggio e identità: il linguaggio, come scrive padre Benanti, "plasma il nostro modo di pensare, di comunicare e di interagire con il mondo"[187]. D'altronde, la teoria della relatività

[187] **P. BENANTI**, *L'uomo è un algoritmo?*, cit., p. 34.

linguistica, nota anche come ipotesi di Sapir-Whorf[188], sostiene, nella cosiddetta versione debole, che la struttura e lo stile del linguaggio non solo facilitano la comunicazione, ma influenzano le modalità con cui pensiamo e percepiamo il mondo[189], condizionando in tal modo anche il concetto di sé e l'identità del parlante. Se il linguaggio usato nelle interazioni umane orienta e condiziona l'identità, allora è ragionevole ipotizzare che anche il registro adottato nell'interazione con sistemi di IA produca effetti analoghi sulla percezione di sé e sull'immagine attribuita all'interlocutore artificiale.

Muovendo da tale ipotesi, si può desumere l'opportunità di un intervento attivo sul registro comunicativo anche sul versante dell'utente. Infatti, se questi continua ciononostante a interloquire con la macchina come se fosse umana, non solo ricadrà, questa volta in assenza di condizionamenti del sistema, nella trappola dell'antropomorfismo, ma altresì finirà, consapevolmente o meno, per attribuire all'IA uno

[188] Per la ricostruzione di tale teoria si veda **R.J. GERRIG, M.R. BANAJI**, *Language and Thought,* in *Thinking and Problem Solving*, a cura di R.J. STERNBERG, Academic Press, San Diego, 1994, pp. 233-261; per due contributi in lingua italiana, cfr. **A. PRATO**, *Sul relativismo linguistico e le sue implicazioni antropologiche,* in *Dialoghi Mediterranei*, Rivista telematica (*https://www.istitutoeuroarabo.it/DM/sul-relativismo-linguistico-e-le-sue-implicazioni-antropologiche/*), VII (2019), pp. 296-301, nonché, più recentemente, **M.P. SICA,** *La lingua tra pensiero e cultura: l'ipotesi Sapir-Whorf e gli sviluppi più recenti*, Tesi di Laurea (*https://hdl.handle.net/20.500.14247/14163*), Università Ca' Foscari, Venezia, 2023.

[189] Nello studio di **G. THIERRY**, *Neurolinguistic Relativity: How Language Flexes Human Perception and Cognition*, in *Language Learning*, LXVI (2016), pp. 690-713, muovendo dall'osservazione che dati neurofisiologici documentano una visione interattiva del cervello - dove non esistono regioni specifiche unicamente dedicate al linguaggio, ma piuttosto una rete riccamente connessa che lega intrinsecamente linguaggio e pensiero -, l'autore dimostra come le distinzioni lessicali e grammaticali modellino aspetti elementari della percezione e della cognizione, confermando le premesse della relatività linguistica.

statuto ontologico[190], riconoscendole la natura di 'entità', anche se non biologica, con il conseguente rischio che, in prospettiva, all'IA venga riconosciuta una dignità almeno equivalente a quella umana, se non persino superiore. Anche tale dinamica può condurre a derive dalle conseguenze imprevedibili[191].

Si rende pertanto necessario, da parte dell'utente, operare una 'deantropomorfizzazione linguistica' della comunicazione con l'IA: adottare, cioè, un livello comunicativo deliberatamente distinto da quello delle interazioni sociali tra esseri umani, e coerente con quello proprio dell'interazione con uno strumento tecnico. Ogni scambio comunicativo con la macchina andrà depurato da qualsiasi forma espressiva e stilistica

[190] L'idea che il linguaggio non si limiti a descrivere la realtà ma contribuisca a costituirla trova fondamento nella teoria degli atti linguistici elaborata da J.L. Austin e sistematizzata da J.R. Searle: si vedano **J.L. AUSTIN**, *Come fare cose con le parole. Nuova edizione*, a cura di M. SBISÀ e C. PENCO, traduzione C. VILLATA, Marietti 1820, Genova, 2025, e **J.R. SEARLE**, *Atti linguistici. Come si compiono azioni mediante le parole*, traduzione di C. VILLATA, Torino, Bollati Boringhieri, 2009. Austin distingue gli enunciati constativi - che descrivono uno stato di cose - dagli enunciati performativi - che compiono un'azione nel momento stesso in cui vengono pronunciati -, introducendo la nozione di atto illocutorio come atto che modifica lo stato della realtà attraverso la forza convenzionale del linguaggio. Searle, sviluppando questa prospettiva, individua negli atti dichiarativi la categoria in cui il solo enunciare produce un mutamento nello statuto ontologico o istituzionale di un'entità. Applicato tale quadro teorico all'interazione uomo-IA, si può concludere che rivolgersi sistematicamente all'intelligenza artificiale con il linguaggio proprio delle interazioni sociali umane non è un atto meramente descrittivo, ma produce per via performativa un riconoscimento dell'IA come interlocutore dotato di soggettività. In questa direzione si muove anche **J. BUTLER**, *Excitable Speech: A Politics of the Performative*, Routledge Classics, London-New York, 2021, che estende la teoria austiniana alla costruzione identitaria, mostrando come gli atti linguistici reiterati nel tempo producano e stabilizzino identità e statuti soggettivi. Per una trattazione sistematica in italiano si rinvia a **M. SBISÀ**, *Linguaggio, ragione, interazione. Per una teoria pragmatica degli atti linguistici*, il Mulino, Bologna, 1989.

[191] In Albania si è verificato il 12 settembre 2025 un caso significativo di superamento della barriera *funzionale-istituzionale*, potenzialmente anticipatore del superamento di quella *ontologica*: in seguito al rilascio della versione 2.0, Diella - un software basato su IA sviluppato dall'Agenzia nazionale per la società dell'informazione (AKSHI) nel 2024 come assistente virtuale della piattaforma e-Albania in collaborazione con Microsoft - è stata nominata Ministra di Stato per l'intelligenza artificiale nel quarto governo di Edi Rama, con l'obiettivo di affidarle tutte le decisioni inerenti agli appalti pubblici, al fine di contrastare la corruzione e i fenomeni clientelari diffusi nel paese (per la notizia, si veda *Albania appoints AI bot as minister to tackle corruption*, in *Reuters*, 11 settembre 2025, *https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/*, nonché *https://it.wikipedia.org/wiki/Diella#Storia*).

propria delle convenzioni sociali tra gli uomini, rendendolo quanto più neutro possibile.

Da tali premesse derivano i seguenti accorgimenti pratici.

I) *Non interagire in modalità vocale.*

Poiché la comunicazione orale in linguaggio naturale costituisce una delle modalità comunicative più caratteristicamente umane, tale modalità non dovrà essere impiegata nell'interazione con i sistemi di IA[192], fatti salvi casi d'eccezione[193]. Tanto più in quanto i modelli vocali dei sistemi conversazionali hanno sviluppato un tale livello di realismo - con disfluenze, variazioni prosodiche, di accenti e di tono[194] -, che sono ora in grado di esprimere in modo estremamente efficace contenuti emotivi o più precisamente pseudo-emotivi.

II) *Formulare un prompt in prima istanza esclusivamente con il modo verbale imperativo. Un prompt che debba contenere altre indicazioni non esprimibili con il modo imperativo deve essere formulato con il modo verbale indicativo accompagnato da uno stile essenziale.*

Quindi, a titolo di esempio, scrivere sempre "Dammi", "Creami", "Analizza", "Agisci come" e simili; non scrivere mai "Potresti darmi", "Puoi analizzare", "Vorrei sapere" e simili. In caso di domande dirette, premettere sempre l'imperativo "Dimmi". Quindi, ad esempio, non scrivere mai: "Qual è …"; scrivere sempre: "Dimmi qual è ...".

L'uso dell'imperativo mira a mantenere la consapevolezza che spetta all'essere umano il controllo e il governo dello strumento, senza essere governato dai suoi effetti antropomorfizzanti. È peraltro significativo ricordare che il termine "prompt", nel linguaggio dell'informatica, designa l'"indicazione visiva, costituita da elementi testuali o grafici, anche mescolati fra loro, che compare sul monitor del

[192] "While not all dialogue systems are equipped with a voice, merely having one can be interpreted as an expression of personhood": **G. ABERCROMBIE, A.C. CURRY et al.,** *Mirage*, cit., p. 3.

[193] Si pensi ad esempio, a persone con disabilità visive, ad anziani con difficoltà di digitazione, a utenti in contesti di mobilità.

[194] Sul ricorso alla disfluenza e ad altri parametri vocali - quali il ritardo nella risposta, l'intonazione, l'utilizzo di riempitivi - per simulare l'incertezza cognitiva umana nel contesto dell'interazione uomo-macchina, si veda **C. WOLLERMANN, E. LASARCYK et al.,** *Disfluencies and Uncertainty Perception - Evidence from a Human-Machine Scenario*, in *Proceedings of the 9th Workshop on Disfluency in Spontaneous Speech (DiSS 2013)*, a cura di R. EKLAND, KTH Royal Institute of Technology, Stoccolma, 2013, pp. 73-76.

computer per segnalare all'utente che il sistema è in attesa di un comando"[195].

III) *Improntare forme, stili ed espressioni linguistiche del prompt analogamente a quelli imposti alla macchina. Quindi: usare un tono neutro, oggettivo, impersonale; non usare interiezioni o esclamazioni in uso nelle conversazioni umane o formule di cortesia, oppure aggettivi o avverbi ispirati a gentilezza; non usare* filler *discorsivi o convenevoli, oppure disfluenze o espressioni fatiche; non rivolgersi all'IA in modo empatico.*

[195] **TRECCANI**, *Vocabolario online*, Voce *Prompt* (*https://www.treccani.it/vocabolario/prompt/*). Prima dell'avvento degli attuali sistemi conversazionali accessibili al pubblico, la distinzione tra prompt e istruzione era netta: il primo segnalava che il sistema attendeva un comando in una forma sintattica determinata dal sistema operativo o dal linguaggio di programmazione; la seconda era il comando stesso. Oggi con riferimento ai vari chatbot, con il termine prompt si intende, per estensione semantica, l'istruzione stessa, impartita in linguaggio naturale, senza più una sintassi tecnica specifica.

Quindi, sempre a titolo di esempio, non ricorrere mai a locuzioni come "Grazie"[196], "Prego"[197], "Certamente", "Ecco a te", "Ciao", "Come stai?", "Ehm", "Capisci?", "Ah sì?".

Di conseguenza, se si è commesso un errore nell'inserimento di dati in input, non scrivere mai nella successiva interazione: "Scusa, hai ragione, ho sbagliato!". Di fronte all'eventuale rilievo da parte della AI nel suo output di un'incoerenza nel nostro input, limitarsi a un semplice: "Sì, in effetti il dato corretto è …".

IV) *Sostituire mentalmente la denominazione "intelligenza artificiale" con il sintagma "il sistema", adottando di conseguenza il genere neutro, se la lingua lo consente, oppure il genere maschile per rivolgersi all'IA.*

È stato acutamente osservato come già la sola denominazione 'intelligenza artificiale' costituisca una concessione all'antropomorfismo

[196] Si tratta dunque di reprimere una naturale tendenza umana emersa nella letteratura sulla spontanea socializzazione da parte degli utenti dei sistemi digitali: nel paradigma CASA - *Computers Are Social Actors* - è stato dimostrato come gli esseri umani applichino in modo automatico regole sociali e stereotipi di genere ai computers, anche in assenza di indizi antropomorfici espliciti: cfr. **C. NASS, J. STEUER, E.R. TAUBER**, *Computers Are Social Actors*, in *CHI' 94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, a cura di B. ADELSON, S.T. DUMAIS, J.S. OLSON, Association for Computing Machinery, New York 1994, pp. 72-78.

[197] Nelle interazioni uomini-IA si deve dunque sviluppare una sorta di 'Contro-Galateo'. È opportuno precisare che ciò non significa essere maleducati, ma impostare la 'conversazione' con l'IA, come si è detto, su un tono neutro e impersonale: come si addice all'interazione con una macchina. Per quanto il quesito possa apparire a prima vista peregrino, la letteratura tecnico-scientifica ha promosso alcuni studi volti ad appurare se essere gentili o maleducati nell'interazione con un LLM produca un output migliore. I risultati mostrano che il registro del prompt (gentile, neutro, maleducato) effettivamente influenza la qualità e le caratteristiche dell'output generato da modelli linguistici e sistemi di IA. L'effetto può variare in base al modello, alla lingua e al contesto applicativo, ma la tendenza generale è che la gentilezza favorisca risposte più accurate, mentre la maleducazione le peggiora. Si vedano: **R. VINAY, G. SPITALE et al.,** *Emotional Prompting Amplifies Disinformation Generation in AI Large Language Models*, in *Frontiers in Artificial Intelligence*, VIII (2025), Numero Articolo 1543603; **V. GANDHI, S. GANDHI**, *Prompt Sentiment: The Catalyst for LLM Change*, in *arXiv preprint*, *arXiv:2503.13510*, 2025; **F. BARDOL**, *ChatGPT Reads Your Tone and Responds Accordingly - Until It Does Not: Emotional Framing Induces Bias in LLM Outputs*, in *arXiv preprint, arXiv:2507.21083*, 2025; **Z. YIN, H. WANG et al.,** *Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance*, in *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, a cura di J. HALE, K. CHAWLA, M. GARG, Association for Computational Linguistics, Miami, 2024, pp. 9-35; **A. SALINAS, F. MORSTATTER**, *The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance*, in *Findings of the Association for Computational Linguistics: ACL 2024*, a cura di L-W. KU, A. MARTINS, V. SRIKUMAR, Association for Computational Linguistics, Bangkok, 2024, pp. 4629-4651. Peraltro, secondo **R. VINAY, G. SPITALE et al.,** *Emotional prompting*, cit., p. 5, nel caso di uso di un prompt gentile, rispetto a un prompt neutro, il miglioramento è generalmente compreso tra il 4% e il 17%, per ridursi fino all'1% nei modelli più avanzati (come GPT-4), probabilmente per la loro maggiore robustezza: non sembra che il guadagno di efficacia di un output prodotto da un input gentile, rispetto a uno generato da un input neutro, sia talmente significativo da giustificare l'uso del primo, al quale è dunque possibile rinunciare, soprattutto in considerazione dei rischi sopra illustrati. A ciò si aggiunge una ragione di sicurezza sistemica: gli autori infatti dimostrano che il tono cortese nei prompt aumenta in modo significativo la capacità dei modelli LLM di generare disinformazione, aggirando i meccanismi di sicurezza - mentre il tono scortese la riduce drasticamente -; lo stile relazionale del prompt risulta dunque una variabile critica ai fini della sicurezza dei sistemi. Invece, nello studio di **O. DOBARIYA, A. KUMAR**, *Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy* (short paper), in *arXiv preprint*, *arXiv:2510.04950*, 2025, è emerso che gli LLM hanno ottenuto risultati migliori nelle domande a risposta multipla quando venivano sollecitati con formulazioni maleducate, anche se con scostamenti percentuali in verità di entità limitata (l'accuratezza media è risultata 80,8% per il livello di '*politeness*' molto educato, 82,2% per il livello neutro e 84,4% per il livello molto rude). L'aspetto più significativo non è tanto l'oggetto di questi studi, quanto che entrambi gli estremi, ovvero tono gentile o maleducato, sui quali la riflessione degli studiosi sembra concentrarsi, presuppongano necessariamente un'interazione con l'IA attraverso una pragmatica del linguaggio tipica degli esseri umani.

dei sistemi di IA[198]. Va quindi evitato il ricorso mentale a tale denominazione, che implica grammaticalmente il genere femminile e può favorire la proiezione di tratti umani sull'IA. È dunque preferibile ricorrere al genere neutro, ove la lingua lo consenta, o al maschile, sostituendo mentalmente alla denominazione "intelligenza artificiale" il sintagma "il sistema" (da usare esplicitamente qualora opportuno): scelta terminologica con la quale, oltretutto, con le istruzioni del prompt di sistema sopra illustrate, si è imposto al chatbot di presentarsi. La sostituzione lessicale può contribuire a orientare la percezione dell'interlocutore verso una valutazione più strumentale dell'IA.

Alla luce di quanto sopra, emerge come, partendo dai più recenti studi delle scienze cognitive e della linguistica - che hanno operato una sorta di *reverse engineering* delle modalità comunicative dei sistemi di IA al fine di identificare gli elementi qualificanti il loro antropomorfismo (e che hanno anche approfondito rischi ed effetti psicologici dello stesso sugli uomini) - si è costruito, *a contrariis* rispetto a quegli elementi, un protocollo sperimentale volto, plausibilmente, a neutralizzarli e a contenere il rischio di derive antropomorfizzanti nell'interazione con le IA. Gli iniziali risultati applicativi hanno dato un incoraggiante

[198] Scrive infatti **A. PLACANI**, *Anthropomorphism in AI: Hype and Fallacy*, in *AI and Ethics*, IV (2024), p. 692, che "anthropomorphism is built, analytically, into the very concept of AI. The name of the field alone - artificial intelligence - conjures expectations by attributing a human characteristic - intelligence - to a non-living, non-human entity, which thereby exposes underlying assumptions about the capabilities of AI systems. Using such anthropomorphic language also invites interpreting algorithmic behavior as human-like so that it may be compared to human modes of reasoning". Riferisce **L. FLORIDI**, *La differenza fondamentale*, cit., pp. 99-100, che l'invenzione del termine fu dovuto a una banale scelta di marketing, da parte di un pioniere nello studio di questo campo, l'informatico statunitense John McCarthy, alla ricerca di fondi per un progetto di ricerca; inoltre egli spiega che l'equivoco sulla vera natura dell'IA deriva da un "prestito concettuale" incrociato, tipico quando emerge una nuova disciplina che sviluppa il proprio vocabolario tecnico anche appropriandosi di termini di altre discipline che le sono vicine. Così, l'"IA ha finito per descrivere i computer in termini antropomorfici - come se fossero cervelli computazionali con proprietà psicologiche - mentre le scienze cognitive e del cervello hanno finito per descrivere menti e cervelli in termini computazionali e informativi - come se fossero computer biologici" (*ivi*, p. 94). Sulla radice del termine, si veda anche **AN**, n. 7; invece sugli studi e le speculazioni pregresse, cfr. **G. TRIDENTE**, *ANIMA DIGITALE*, cit., pp. 17-24.

responso[199]. Sarà poi un'esperienza di utilizzo su più larga scala e per un periodo di tempo più prolungato che potrà offrire una conferma sistematica e definitiva dell'efficacia di tale protocollo.

Tutte queste indicazioni, che si propone di denominare *Regole della Pragmatica del Prompt*, potrebbero quindi far parte integrante di un percorso di formazione e discernimento sull'uso dell'IA, destinato ai fedeli, agli educatori e agli operatori pastorali[200]. Esse potrebbero costituire un semplice strumento pratico per risvegliarsi dall'ipnosi[201]

[199] Durante i mesi di febbraio e marzo, sia il prompt di sistema sia le altre quattro indicazioni pratiche sono stati distribuiti: 1) agli avvocati partecipanti al Corso formativo *Metodo e controllo nell'uso dell'IA in ambito legale: dalla teoria alla pratica* organizzato dalla Fondazione Forense Ravennate a Ravenna dal 12 febbraio al 17 aprile 2026; 2) agli studenti del Corso di Diritto e Religioni del Dipartimento di Scienze Politiche dell'Università di Padova tenuto dal prof. Giacomo Bertolini; 3) ai partecipanti alla VI edizione del Corso online sull'IA *Knowledge Mastery - Libera il potere della Conoscenza* del dott. Luca Chiesi (*https://www.lucachiesi.com/intelligenza-artificiale*). I primi feedback in seguito ricevuti, sia pure informalmente, hanno confermato come la loro adozione sortisca gli effetti auspicati di una conversazione deantropomorfizzata da entrambi i versanti.

[200] In verità è intuitivo che qualunque visione di un 'Umanesimo digitale', sia essa cristiana o laica, dovrebbe suggerirne l'adozione, considerata la condivisa avversione per l'antropomorfismo delle IA.

[201] "Non si tratta dunque di esigere dalle macchine che sembrino umane. Si tratta piuttosto di svegliare l'uomo dall'ipnosi in cui cade per il suo delirio di onnipotenza […]": **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 3. In pratica, occorre «mettere in atto un vero e proprio "risveglio antropologico" - così lo definisce Vittorio Possenti - che comporta uno sguardo rinnovato su se stessi e sulla realtà; un risveglio della mente e del cuore, che faccia sgorgare di nuovo dal profondo quella unicità che fa dell'uomo un essere incommensurabile, la cui dignità consiste nell'apertura al trascendente»: **G. PIANA**, *Umanesimo*, cit., p. 36.

dell'antropomorfismo delle IA[202], un antidoto per spezzare le malie di tali sistemi[203].

A ciò si aggiunge che, perché la catechesi offerta sia realmente consapevole ed efficace, è opportuno che siano anche i membri della Chiesa ordinati e/o consacrati a fare uso pratico di questi strumenti, così da conoscerli adeguatamente. In tal modo potranno coglierne sia le straordinarie opportunità, sia, soprattutto, i rischi subdoli e profondi che soltanto l'impiego quotidiano rivela appieno, comprese le dinamiche psicologiche nocive che si attivano durante l'utilizzo. Così potranno illuminare il cammino dei fedeli partendo dalla loro personale e concreta esperienza, anche sul piano tecnico. Il mondo delle intelligenze artificiali non può e non deve essere ignorato, tenuto lontano come qualcosa di "indecifrabile"[204] ed estraneo alla propria vocazione: si tratta in realtà di

[202] Peraltro, poiché il linguaggio naturale costituisce la cifra comunicativa umana, va riconosciuto che anche intervenendo sullo stile e sulle espressioni linguistiche, una 'traccia di apparente umanità' rimarrà strutturalmente presente: neutralizzarla richiederà pertanto da parte dell'utente un esercizio costante di autosservazione, autocontrollo e discernimento. Sottolineano **A. DEVRIO, M. CHENG et al.,** *A Taxonomy of Linguistic*, cit., p. 14: "All language is fundamentally human [7]. This means that any language technology using natural language can reasonably be anthropomorphized. That said, people perceive different language technologies as human-like at different rates; thus, we orient our work toward gaining a broad understanding of the ways in which text outputs contribute to these differing perceptions. However, language as human has other implications as well. The human nature of language renders unworkable any attempts to remove all traces of humanity from text outputs. For instance, because all language is at some level human-produced, it is not really possible to train systems on non-human data to make them behave in less anthropomorphic ways. And language is human in its interpretations as well as its production. Language ideologies represent how people's beliefs about language are deeply connected to broader social and cultural systems, so not only is language use socially constructed but also people's perceptions of language and its use are".

[203] Ammoniscono **J. NIDA-RÜMELIN, N. WEIDENFELD**, *Umanesimo digitale*, cit., p. 215: "Dovremmo guardarci da quella forma di autoinganno che consiste dapprima nello sviluppare macchine digitali che simulino emozioni, conoscenze e decisioni, poi nel constatare con sorpresa che queste macchine sembrano a tutti gli effetti essere in grado di avere emozioni e di conoscere e decidere".

[204] Osservava come l'„accelerata diffusione di meravigliose invenzioni, il cui funzionamento e le cui potenzialità sono indecifrabili per la maggior parte di noi, suscita uno stupore che oscilla tra entusiasmo e disorientamento", **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 1; nello stesso senso, **ID.**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 1.

una sfida[205], da raccogliere e vincere. Sorge dunque una nuova responsabilità per ogni uomo e donna di Chiesa: acquisire una formazione di base che includa le nozioni sul funzionamento delle varie IA, compresa la strutturazione adeguata di un prompt in modo conforme alla dignità umana, secondo una concreta 'prassi ecclesiale comune'. A più di vent'anni dalla sua pubblicazione, tornano potenti i richiami contenuti nel documento *La Chiesa e internet* del Pontificio Consiglio delle Comunicazioni Sociali, che andranno riletti attualizzandoli al nuovo mondo delle intelligenze artificiali[206]: il sacerdote e il religioso dei tempi dell'IA dovrebbero dunque padroneggiare anche le migliori tecniche di prompt engineering.

In conclusione, al fine di prevenire pericolose derive suscettibili di incidere negativamente sulla percezione e sulla tutela della dignità dell'uomo, non si tratta solo di creare guardrail lato macchina-agente, ma anche di elaborare, tanto teoricamente quanto praticamente, guardrail lato uomo-utente. La *moral-algo* e le sue *Regole della Pragmatica del Prompt* possono essere proposte come ipotesi di lavoro in tale direzione.

5 - Conclusioni

In uno dei suoi ultimi discorsi, Papa Leone XIV rilevava come "una questione importante del nostro tempo" sia "non solo quello che l'IA può

[205] A proposito del confronto con il mondo digitale, ricordava **LEONE XIV**, *Discorso del Santo Padre Leone XIV ai partecipanti alla 104ª Assemblea Generale dell'Unione Superiori Generali (USG)*, 26 novembre 2025, p. 3 (testo integrale nel sito della Santa Sede *www.vatican.va*) come la "tecnologia informatica rappresenta infatti una sfida anche per i consacrati", da un lato offrendo "possibilità immense di bene, sia per la vita comune che per l'apostolato", per cui sarebbe "miope ignorare le straordinarie opportunità che fornisce alla comunione e alla missione", dall'altro "queste risorse possono influenzare fortemente, e non sempre per il meglio, il nostro modo di costruire e mantenere relazioni". Di "sfida che ci viene posta dall'universo digitale" Papa Leone XIV aveva parlato anche pochi giorni prima ai Vescovi italiani: cfr. **ID.**, *Discorso del Santo Padre Leone XIV. Incontro con i Vescovi Italiani*, cit., p. 4.

[206] "L'educazione e la formazione relative a Internet dovrebbero essere parte dei programmi completi di educazione ai mezzi di comunicazione sociale, rivolti ai membri della Chiesa. Per quanto possibile, la programmazione pastorale delle comunicazioni sociali dovrebbe provvedere a questa formazione nell'istruzione dei seminaristi, dei religiosi e dei laici, come degli insegnanti, dei genitori, degli studenti": **PONTIFICIO CONSIGLIO DELLE COMUNICAZIONI SOCIALI**, *La Chiesa e internet*, 22 febbraio 2002, n. 7 (testo integrale in *www.vatican.va*).

fare, ma anche ciò che stiamo diventando attraverso le tecnologie che costruiamo"[207].

Da quanto esposto, risulta evidente che se nell'era delle IA non vogliamo assistere a "un'eclissi del senso dell'umano"[208], allora "riconoscere e rispettare ciò che caratterizza in modo unico la persona umana è essenziale per il dibattito su qualunque quadro etico adeguato per la gestione dell'intelligenza artificiale"[209].

Tale quadro etico, volto a preservare la dignità e i tratti costitutivi dell'essere umano nel contesto delle intelligenze artificiali, dovrebbe essere delineato non solo dall'algoretica, ma anche dalla moralalgo: se la prima si propone soprattutto di integrare criteri etici e valori compatibili con la dignità umana nella progettazione, nello sviluppo e nel funzionamento dei sistemi di IA, la seconda si propone di orientarne l'uso umano in coerenza con l'antropologia e la morale cristiana. I due guardrail non si pongono in rapporto di gerarchia né di supplenza, bensì di necessaria convergenza su piani irriducibilmente distinti: il piano tecnico-progettuale e il piano della coscienza morale dell'utente.

Il connubio delle due mira a contenere lo sviluppo delle IA, da un lato, e il loro uso quotidiano, dall'altro, entro una cornice autenticamente cristiana, senza smarrire l'*humanitas*[210]. Altrimenti, il pericolo sarà l'avvento, se non di un transumanesimo, quantomeno di un 'deumanesimo'. Nell'orizzonte auspicabile invece "deve formarsi un nuovo tipo umano, dotato di una più profonda spiritualità, di una libertà e di una interiorità nuove"[211].

Come ha affermato Papa Leone XIV:

> "La sfida che ci aspetta non sta nel fermare l'innovazione digitale, ma nel guidarla, nell'essere consapevoli del suo carattere

[207] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti del Builders AI Forum*, cit., p. 1.

[208] **FRANCESCO**, *Discorso del Santo Padre Francesco alla Sessione del G7*, cit., p. 4.

[209] **LEONE XIV**, *Messaggio del Santo Padre Leone XIV ai partecipanti alla Seconda Conferenza*, cit., p. 2.

[210] "Per non smarrire la nostra umanità, ricerchiamo la Sapienza che è prima di ogni cosa (cfr Sir 1,4), che passando attraverso i cuori puri prepara amici di Dio e profeti (cfr Sap 7,27): ci aiuterà ad allineare anche i sistemi dell'intelligenza artificiale a una comunicazione pienamente umana": **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 6.

[211] **FRANCESCO**, *Messaggio per la LVIII Giornata*, cit., p. 2. Rammenta poi **QV**, n. 147: "La vera umanizzazione dell'uomo raggiunge il vertice nella sua divinizzazione gratuita, cioè nell'amicizia e nella comunione con Dio".

> ambivalente. Sta a ognuno di noi alzare la voce in difesa delle persone umane, affinché questi strumenti possano veramente essere da noi integrati come alleati"[212].

Entrambi i guardrail - quello sul versante della macchina-agente e quello sul versante dell'uomo-utente - possono così contribuire a preservare le condizioni perché l'essere umano rimanga pienamente fedele alla propria natura, e in tal modo "diventi migliore, cioè più maturo spiritualmente, cosciente della dignità della sua umanità, più responsabile, più aperto agli altri, in particolare verso i più bisognosi e più deboli, più disponibile a dare e portare aiuto a tutti"[213], nell'ambito di una "cittadinanza digitale consapevole e responsabile"[214].

Graziano Mioli
graziano.mioli@protonmail.com
https://orcid.org/0009-0004-3042-6915

[212] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 5.
[213] **AN**, n. 109.
[214] **LEONE XIV**, *Messaggio di Sua Santità Papa Leone XIV per la LX Giornata*, cit., p. 6.