

Il futuro dell'Intelligenza artificiale dipende da quello che pensiamo dell'essere umano

di Martina Todaro

Abstract

What can one think about human beings? At an adequate level of abstraction, one can either believe that humans are special or not. According to the resulting irreconcilable metaphysical positions, AI is either a thought experiment that challenges exceptionalism or evidence of human uniqueness. Since the latter half of the 20th century, we have been experiencing the Promethean drift of exceptionalism, which some term transhumanism. This is not a third way of regarding humans but the unforeseen decoupling phenomenon introduced by AI that is separating humanism from anthropocentrism. While expanding the parameters of the AI debate, the resulting philosophical relentlessness risks fuelling political turmoil and jeopardising AI governance, which in turn sets the stage for the future of AI.

Keywords: AI, Future, Personalism, Transhumanism, Post-humanism

Introduzione

Con Intelligenza Artificiale (IA) ci si riferisce a una disciplina e talvolta a un artefatto tecnologico. Ancor prima, però, l'IA rappresenta un'aspirazione (o una paura) umana. Quando dei tre livelli di astrazione ci si limita a usare quello tecnologico, si rinuncia agli strumenti per comprendere l'acceso dibattito sui rischi e le modalità di regolamentazione dell'IA (FLI, 2024).

Non è certo la prima volta nella storia in cui i massimi esponenti di una disciplina mostrano un così profondo disaccordo sulle fondamenta della disciplina¹ che li elegge esperti, guru, *godfather*; tuttavia, è ancora piuttosto raro che un dibattito scientifico arrivi a toccare il tema dell'estinzione umana² e ancor più raro il tema dell'essenza dell'essere umano.

¹ Russell e Godel in logica matematica, Einstein e Bohr in fisica per citare i più famosi senza dover elencare tutti i protagonisti del dibattito millenario sulla filosofia.

² Alcune altre teorie che toccano l'argomento dei rischi esistenziali per l'essere umano riguardano l'olocausto nucleare e il cambiamento climatico, ma anche eventi naturali

L'essenza dell'essere umano a cui facevano riferimento i greci è legata alle caratteristiche peculiari degli esseri mortali. Il sentimento umano nei confronti della morte dipende da vari fattori. Se esistere è una condizione tendenzialmente preferibile a quella di non esistere, ci sono tuttavia delle contingenze che possono far preferire il sacrificio come, in primo luogo, la difesa della propria essenza. Non avendo mai avuto l'opportunità di scelta è difficile dire quanti di noi sceglierebbero l'immortalità e quanti invece il sacrificio in difesa della natura umana. Ciò che conta in questo esperimento mentale è capire che l'essere umano immortale avrebbe caratteristiche così diverse da quello mortale che probabilmente sarebbe più facile chiamarlo in altro modo, dio ad esempio. Ciononostante, esiste oggi una corrente di pensiero che sostiene non solo il desiderio di vincere la morte (vecchio come il mondo, il cui capofila identifico in Dante Alighieri) ma anche l'esistenza di una possibilità tecnologica nella realizzazione di questo desiderio.

Altro fattore determinante nel contesto del rapporto umano con la morte è il passaggio dal concetto di morte dell'individuo a quello dell'estinzione della specie.

Il tema dei rischi catastrofici (X-risks) legati allo sviluppo dell'IA viene sostenuto con più vigore da chi colloca l'estinzione umana all'estremo riservato agli eventi meno desiderabili. Anche in questo caso non c'è largo consenso. Tra i detrattori della teoria troviamo coloro i quali ritengono che l'IA non possa collocarsi tra i rischi esistenziali e coloro i quali accettano la tesi ma non vedono l'estinzione umana come un evento poi così deprecabile. In un recente studio inglese si evince come il concetto di estinzione umana sia interpretato in relazione all'imminenza dell'evento stesso (Shubert, 2019). Il che significa che, scongiurata l'ipotesi di viverlo in prima persona e possibilmente scongiurata l'ipotesi che la vivano i nostri figli e nipoti, da un certo punto in poi una persona su cinque delle intervistate non ritiene che l'estinzione sia in assoluto un evento moralmente negativo.

Un evento è moralmente negativo se incide negativamente sul focus morale. Il focus morale è un riflettore che illumina una porzione di soggetti grande a piacere. Se il focus morale accoglie i conquistatori e non i nativi, gli uomini e non le donne, gli esseri umani e non le altre specie, allora nativi, donne e le altre specie diventano mezzi utili al benessere rispettivamente dei conquistatori, uomini ed esseri umani.

catastrofici come eruzioni vulcaniche, impatti con corpi celesti, pandemie, collasso del bosone di Higgs.

Ad esempio, considerare l'estinzione umana un evento moralmente positivo significa aver posto l'ecosistema al centro del focus morale e giudicare il comportamento dell'essere umano in base all'influenza negativa che ha sul benessere della biodiversità.

Le varie interpretazioni dell'ipotesi dell'IA come X-risk, e più di questo la centralità dell'essere umano rispetto al focus morale, danno luogo a posizioni completamente differenti di governance dell'IA. Non c'è da stupirsi che il dibattito morale e filosofico si trasformi in politico (Calabresi, 1978) ma è proprio la relativa postura normativa che ci impone di rivolgere il pensiero al futuro. Come si legge nel recentissimo International AI Safety Report 2025: «AI does not happen to us: choices made by people determine its future. The future of general-purpose AI technology is uncertain, with a wide range of trajectories appearing to be possible even in the near future, including both very positive and very negative outcomes. This uncertainty can evoke fatalism and make AI appear as something that happens to us. But it will be the decisions of societies and governments on how to navigate this uncertainty that determine which path we will take» (Bengio, 2025).

Ed è proprio il futuro il soggetto del presente documento. Un futuro, quello dell'IA, che una parte degli esperti interpreta come inevitabile e distopico (si pensi alle ragioni che hanno portato all'open letter³ per chiedere una pausa di sei mesi sullo sviluppo dell'IA, (FLI, 2023; OpenAI, 2023), una parte come inevitabile e positivo o al più indifferente (Taylor, 2023), e un'altra parte come un qualcosa che va scritto, o anticipato, e di cui dobbiamo prendercene la responsabilità (Fuller, 2024). Il tema della inevitabilità è un'arma a cui chi si inserisce nel dibattito sull'IA inizia ad abituarsi. Anche sostenere che assisteremo presto a un nuovo inverno dell'IA (Floridi, 2020) rientra in questo schema.

La voce di chi fa ricorso al tema dell'inevitabilità sembra più forte nel buzz che riguarda l'IA. Al punto che varrebbe la pena domandarsi se gli esperti abbiano la responsabilità di darci la loro visione di futuro o se sia la caratteristica di inesorabilità della previsione di una persona qualunque a donare un alone di competenza a chi la esterna. Nel caos dei vari proclami, il sospetto è che i "doomeristi", o profeti della

³ L'iniziativa e lo studio alle spalle sono stati condotti da Future of Life Institute con sede a Boston. Tra i personaggi di spicco dell'associazione troviamo Max Tegmark, autore di Life 3.0, e Nick Bostrom, autore di Superintelligence.

venuta delle macchine, abbiano conflitti di interessi non trascurabili (fear-based marketing e panic as a business), mentre chi più si agita nello sbandierare i limiti invalicabili delle macchine non sia spaventato tanto dai rischi che comporterebbe condividere il mondo con un nuovo compagno di viaggio come lo è stato il Neandertal in passato ma, semmai, dalla catastrofe esistenziale che subirebbe nello scoprire di non essere padrone del proprio destino.

Con l'analisi che segue intendo chiarire le radici del dibattito sull'IA descrivendo le varie posizioni in relazione a come si collocano nei confronti dell'antropocentrismo e dell'umanesimo. Nel contesto della filosofia della mente e dell'ontologia differenziale, sosterrò che il dibattito sull'IA non rivela nulla di interessante sull'IA in sé per sé ma è acceso da opinioni differenti sul rapporto tra IA ed essere umano. Riportare il discorso alla diatriba sull'essenza della natura umana consente di ricollocare l'IA al suo ruolo di elezione, ovvero il soggetto di un esperimento mentale il cui scopo è quello di indagare il confine tra immanente e trascendente.

Sistemi di IA

Sia in ambiente accademico che giuridico ci sono innumerevoli definizioni di sistema di IA (livello di astrazione tecnologico) in grado di rappresentare almeno in parte l'inconciliabilità delle posizioni nel dibattito sull'IA.

Queste definizioni possono essere suddivise in quelle che pongono l'accento su un qualche grado di autonomia della macchina (nei confronti dell'essere umano) e quelle che sottolineano la loro natura di artefatti umani. La contrapposizione filosofica tra i due gruppi è data dalla volontà di sottolineare, o meno, che qualsivoglia comportamento di tali strumenti sia attribuibile a obiettivi definiti dagli esseri umani, siano essi gli utilizzatori o i produttori.

Nel Paragrafo 3: Definizioni del Capitolo 1, dell'AI Act si legge:

(1) 'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; (AI Act, 2024).

Una definizione molto simile a quella che ne dà l'OCSE nel 2024 (OECD, 2024).

Tuttavia, l'OCSE ha aggiornato la sua definizione solo recentemente. Prima del marzo del 2024 l'OCSE sosteneva che:

An AI system is a machine-based system that can, *for a given set of human-defined objectives* [enfasi aggiunta], make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (OECD, 2019).

La rimozione della dicitura “for a given set of human-defined objectives” richiede l’aggiornamento della definizione di autonomia. Nell’interazione umano-macchina, la macchina viene considerata autonoma se è in grado di operare senza alcun coinvolgimento o intervento umano⁴, in contrapposizione a una macchina a funzionamento manuale (Annex AI Act, 2025). Si noti che un certo grado di automazione è già sufficiente per rispettare la definizione di macchina autonoma che ne dà l’AI Act, pertanto è necessario specificare che, almeno per l’OCSE, un sistema di IA è tale se è in grado di definire obiettivi non specificati dagli esseri umani. Questi obiettivi possono essere estrapolati da segnali o valori di riferimento esterni, come quelli ricavabili dall’interazione con l’ambiente (a cui la macchina può mostrare un certa capacità di adattamento).

A complicare il dibattito, OpenAI sostiene che algoritmi di tipo Multi Agent Reinforcement Learning (MARL) siano in grado di manifestare la motivazione intrinseca necessaria per perseguire scopi auto-selezionati (Baker, 2019), e Stuart Russell, che nel mettere in guardia dagli effetti collaterali potenzialmente catastrofici dovuti all’incapacità dell’essere umano di specificare gli obiettivi, finisce per suggerire di rallentare o fermare lo sviluppo di sistemi di IA particolarmente pervasivi (Russell, 2019).

Tuttavia, è bene chiarire che i sistemi di IA presi in considerazione da OpenAI e Russell, siano essi particolarmente affini alla rigenerazione spontanea del codice (apprendimento per rinforzo) o meno (sistemi di raccomandazione), non fanno altro che rispettare la richiesta esplicita, fornita dagli esseri umani, di massimizzare una determinata variabile o funzione. In matematica, massimizzare una variabile significa agire su tutte le altre che si hanno a disposizione. Dunque, nel

⁴ Adriano Fabris sostiene che le macchine dotate di un certo grado di autonomia non sono autonome nel senso che possano “darsi da sé la propria norma”, ma piuttosto perché possono promuovere determinate azioni e raggiungere obiettivi con una crescente indipendenza e imprevedibilità. Fino ad oggi solo gli eventi della natura potevano essere qualificati come imprevedibili (Fabris, 2022).

pieno rispetto del concetto di *pharmakon* caro a Platone e Derrida, dovremmo essere abituati a riflettere sugli effetti collaterali portati dalla tecnologia, specie quella in grado di massimizzare variabili su cui non abbiamo imposto vincoli per la fallacia del buon senso⁵.

In questa sezione non si discute che tali obiettivi inaspettati siano da considerarsi meramente strumentali al raggiungimento di obiettivi umani, siano essi intrinseci, ovvero specificati nel codice stesso, o estrinseci, ovvero esplicitati dall'essere umano in una seconda fase (si pensi all'auto a guida autonoma che decide autonomamente il percorso affinché venga rispettato l'obiettivo umano di giungere a una determinata destinazione).

Pertanto, rimuovere la dicitura “for a given set of human-defined objectives” sembra un tentativo di generalizzazione atto ad anticipare (con l'intento di scongiurare) eventuale ulteriore ritardo dell'attività giuridica nei confronti dell'innovazione tecnologica piuttosto che a giustificare l'imposizione normativa della responsabilità (come può essere quella nei confronti dei danni causati da animali da compagnia) e tanto meno a chiarire la posizione filosofica del legislatore nei confronti del rapporto tra essere umano e macchina.

Se non altro, le diverse posizioni sembrano concordare sul fatto che i sistemi di IA siano legati all'hardware e alla capacità computazionale di quest'ultimo:

(12) All AI systems are machine-based, since they require machines to enable their functioning, such as model training, data processing, predictive modelling and large-scale automated decision making. The entire lifecycle of advanced AI systems relies on machines that can include many hardware or software components. The element of ‘machine-based’ in the definition of AI system underlines the fact that AI systems must be computationally driven and based on machine operations. (Annex AI Act, 2025).

⁵ Vengono alla mente esempi di algoritmi evolutivi che alla richiesta di massimizzare la velocità di locomozione hanno oviato “evolvendo” una gamba in altezza affinché l'obiettivo fosse raggiunto semplicemente lasciandosi cadere. Oppure il robot che avrebbe dovuto imparare a fare i pancake (la simulazione di un braccio che tiene una padella) che ha invece imparato a tirarli più in alto possibile per massimizzare la variabile che misura il tempo in cui il pancake rimane lontano da terra. Più inquietanti quelli che hanno imparato a mettere in pausa Tetris o a crashare la simulazione per non perdere la partita. Per visionare un'ampia lista di esempi di cattiva specificazione degli obiettivi: <https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLR-qJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml>

E questo basta. La specificità dell'IA è nel vincolo della computazione. Questa caratteristica è sufficiente per sostenere tutto il corpo dell'argomentazione che segue. Dunque, d'ora in poi, IA è sinonimo di macchina(e) ed è definita come set di particolari configurazioni della materia.

Formalizzazione della tesi

Premesse

- a) Il futuro è aperto.
- b) Gli esseri umani sono speciali nella loro capacità di sfuggire alla determinatezza della causalità (ipotesi nulla).

Conclusioni

- c) I processi mentali non sono processi algoritmici.
- d) La mente non è computabile.
- e) Gli esseri umani esibiscono la caratteristica di autonomia e la capacità di controllo.
- f) Il futuro dell'IA dipende da quello che gli esseri umani pensano di loro stessi.

Argomentazione

Il futuro è aperto (a) se l'universo non è deterministico e non esclude l'ipotesi di libero arbitrio.

L'essere umano è speciale (b) se è dotato di libero arbitrio.

Dalla (b) seguono direttamente la (c), la (d) e la (e). Riguardo alla (e) si definisca l'autonomia umana come diretta conseguenza del libero arbitrio o di darsi da sé la propria norma (Fabris, 2022) e la capacità di controllo come un'istanza particolare della caratteristica di autonomia legata alla capacità tecnica.

L'algoritmo è un insieme di regole o istruzioni. Solitamente tali istruzioni sono utili per eseguire una sequenza di calcoli (o computazione)⁶. In un calcolo del tipo $5 + 1 = 6$; 5 e 1 sono gli input, 6 è

⁶ Laddove è consentito supporre una correlazione biunivoca tra software (algoritmo)

l'output, e l'operatore somma è l'algoritmo che identifica unicamente tutte le istruzioni che consentono a un elaboratore di eseguire tutte e solo le operazioni ad esso associate (operatori booleani). Tutti gli algoritmi sono deterministici. Un algoritmo è deterministico quando a input identici restituisce output identici. Esistono degli algoritmi, ad esempio quelli di apprendimento automatico e gli algoritmi evolutivi, in grado di eseguire istruzioni condizionali, ovvero in grado di adattare il proprio codice in base agli input⁷ che ricevono⁸ ma anche questi algoritmi sono deterministici perché l'evoluzione delle istruzioni è condotta in maniera conforme alle istruzioni che esso stesso impartisce. Il codice di partenza contiene già ogni possibile combinazione tra input e output così come ogni funzione matematica del tipo $y=f(x)$ esiste⁹ a prescindere dal valore di x . L'essenza della computazione è infatti la relazione tra input e output, è il flipper entro cui viene inserita una pallina. Il flipper esiste¹⁰ a prescindere dalla pallina e, a maggior ragione, dal punto di ingresso e velocità con cui essa fa il suo ingresso nel flipper.

Esistono tre casi particolari con cui il determinismo computazionale deve fare i conti: le forme indeterminate, le forme indefinite e la randomicità.

In tutti e tre i casi non ci si può aspettare output identici a input identici ma se le indeterminate (del tipo 0/0) denotano assenza di informazioni e le indefinite (del tipo 1/0) denotano limitatezza degli strumenti matematici a disposizione¹¹, la randomicità denota pura casualità o impossibilità di previsione. Affinché qualcosa sia impossibile da prevedere non deve manifestare nessuna influenza da parte di

e hardware (capacità computazionale) sarà usata la notazione di "macchine" in luogo di hardware che computano algoritmi di IA.

⁷ Non crei confusione il fatto che in fase di training l'algoritmo di apprendimento supervisionato si adatti agli output desiderati dal programmatore. Dal punto di vista dell'algoritmo questi rimangono input.

⁸ Questo talvolta porta a pensare che gli output di alcuni di questi particolari algoritmi siano imprevedibili e, pertanto, che tali algoritmi siano in grado di scegliere autonomamente il proprio output ma sono entrambi falsi miti giacché dato un tempo sufficiente è possibile risalire all'output di qualsiasi algoritmo anche con carta e penna.

⁹ È possibile pensare al grafico della funzione, senza scomodare il platonismo. Che il rapporto tra cerchio e il suo diametro sia π a prescindere dall'esistenza dell'essere umano e del suo pensiero o meno non aggiunge nulla al concetto di computazione.

¹⁰ Senza dover necessariamente scomodare l'intuizionismo.

¹¹ $y=(x)^{1/2}$ è indefinita per $x<0$ ma torna ad assumere significato con la matematica dei numeri complessi.

ciò che la circonda¹², ovvero non deve essere causata da qualcos'altro. Non c'è output di nessuna delle tre forme a cui riusciremmo ad assegnare un significato ma nei primi due casi questo si presenta come un problema epistemico e nel terzo come ontologico.

Se le premesse (a) e (b) sono valide¹³ allora è possibile ritenere che l'evoluzione dell'IA non sia predeterminata e dipenda dalle scelte degli esseri umani (che sono altresì in grado di assegnar loro un significato).

Si noti che il desiderio o la paura per un certo tipo di futuro non certifica alcuna delle premesse e che la mutua influenza tra essere umano e IA consente di analizzare la premessa (b) elevandola così a ipotesi nulla da verificare o confutare attraverso la statistica (il numero di Turing Test superati da una macchina che consideriamo sufficienti per ammettere che essa pensi è pari al numero di cigni bianchi osservati che consideriamo sufficienti affinché si possa ritenere che tutti i cigni siano bianchi).

Futuro

Il concetto di futuro è strettamente legato a quello del divenire, ovvero del mutare dell'essere (Severino, 2000).

È possibile, in primo luogo, sostenere che un ente possa diventare altro da sé?

Se così non fosse ogni ente sarebbe un eterno immutabile. Con questo non si intende che ogni istante sia uguale al precedente: si intende che causa ed effetto coesistono, ovvero che ogni cosa, così come si manifesta in un determinato istante, non avrebbe potuto essere diversamente da ciò che è (filosofia della necessità, determinismo).

¹² Al tempo stesso, però, è sicuramente più difficile prevedere stati futuri di un sistema che sia influenzato dall'esterno; impossibile se la sorgente d'influenza esterna è randomica o indeterministica. Gli algoritmi che utilizzano input che si basano sul rumore sono pseudo randomici poiché il rumore, incluso quello Browniano, è un fenomeno deterministicamente e/o algoritmicamente causato. Il rumore Browniano, in particolare, è trattabile stocasticamente ma rimane un fenomeno descrivibile con le catene di Markov. Queste sono definite come quei processi (come il lancio di una moneta) il cui risultato non ha memoria del passato (lanci precedenti), ovvero quei processi la cui incertezza epistemica non può essere ridotta traendo informazioni dal passato. Ciò non toglie che non si possano fare previsioni esatte sull'esito del lancio della moneta partendo dalla misurazione della forza e dal momento impressi alla moneta (Vulpiani, 2014). L'unica obiezione possibile alla pseudo randomicità del rumore riguarda quello causato intenzionalmente dalle attività umane, riportando così la diatriba alla definizione di autonomia umana.

¹³ La premessa (a) da sola non è sufficiente in quanto non è detto che in un universo indeterministico l'essere umano sia dotato di libero arbitrio.

Qualora invece si intenda sostenere l'ipotesi di un divenire libero dalle necessità dell'essere, è opportuno considerare l'idea che nel processo del divenire qualcosa sarebbe potuta andare diversamente (filosofia della libertà).

Avere la capacità di predire con un grado di precisione grande a piacere il divenire può dimostrare il determinismo ma ciò non implica che valga il contrario: non riuscire a predire alcuno stato dell'universo non dice nulla sulla metafisica del divenire ma relega, semplicemente, l'osservatore umano a una condizione di incertezza epistemica.

Come ricorda il fisico Angelo Vulpiani: «siamo condannati a lavorare con una precisione finita» (Vulpiani, 2014) e, pertanto, la diatriba tra Einstein e Bohr su determinismo e probabilità non potrà essere decisa su basi empiriche.

Se Dio gioca a dadi con l'universo o meno è il dilemma che è emerso con la meccanica quantistica. L'interpretazione di Copenhagen è l'ultima speranza dell'indeterminista che reclama il caso come legge di natura pur non riuscendo ancora a riportarlo sistematicamente nel macroscopico.

Qualora le particelle elementari mostrino comportamenti puramente randomici, il collasso della funzione d'onda sarebbe un fenomeno non computabile¹⁴. Non a caso, Sir Roger Penrose e Stuart Hameroff sono corsi a verificare la presenza di fenomeni quantistici in determinati neurotrasmettitori (microtubuli). Se per Penrose questa scoperta sia sufficiente a spiegare lo stato di coscienza, per Hameroff e Federico Faggin questa è addirittura la dimostrazione del libero arbitrio (Hameroff, 2014; Mauro D'Ariano, 2020).

Anche nell'ipotesi in cui i fenomeni quantistici siano puramente randomici e che questi abbiano un ruolo nella formulazione dei pensieri o autocoscienza, come può essere libera la volontà se legata a fenomeni, quelli randomici, su cui per definizione non è possibile ambire ad alcun tipo di controllo? E questo eventuale *quantum brain*¹⁵ in che modo si può ritenere speciale rispetto a un'IA che faccia uso di quantum computing? Si noti tra l'altro che ogni problema risolvibile da un quantum computer è risolvibile da un classico computer^{16 17}

¹⁴ Non fa che spostare il problema dal concetto di determinismo a quello di capacità di simulare l'universo al computer.

¹⁵ Quel cervello in grado di manifestare pensieri sensibili a fenomeni quantistici e al tempo stesso mantenere la proprietà di libero arbitrio.

¹⁶ O addirittura con carta e penna dopo un tempo sufficiente.

¹⁷ Come chiarito anche nell'art. (13) dell'Annesso all'AI Act del 6 febbraio del 2025: (13) The term 'machine-based' covers a wide variety of computational systems. For example, the

(Nielsen & Chuang, 2010).

In questo senso Faggin fa bene a specificare che il libero arbitrio è un postulato (Essentia Foundation, 2024) ma così facendo cade nella fallacia dell'argomento circolare nel quale si ostina a dimostrare la premessa assiomatica.

L'essere umano non sembra vicino a risolvere in maniera incontrovertibile il problema del determinismo poiché al momento può aspirare al più a interpretazioni (dall'interpretazione di Copenhagen alla many-worlds interpretation, passando per l'interpretazione di Penrose stesso secondo cui l'osservatore non ha alcun ruolo ma la gravità sì, fino a quella del superdeterminismo). Anche nel caso volessimo trascurare soluzioni analitiche e ricavare sperimentalmente le regole dell'universo dovremmo misurarci col fatto che ad antecedenti simili corrispondono spesso conseguenze molto diverse.

Va sottolineato che il regime caotico, così come descritto da (Maxwell, 1965), Poincaré e Lorentz (il cui articolo era intitolato *Deterministic Nonperiodic Flow*), è imprevedibile ma deterministico (Atmanspacher, 2002).

Mundus e il trasumanar Dantesco

Il concetto di mundus emerge nelle parole di Alfano di Salerno intorno all'anno 1000. A un livello di astrazione adeguato per guardare al rapporto tra essere umano e macchina, il mundus è sostanzialmente un insieme. Tale insieme può essere grande a piacere, può contenere la specie umana o meno a seconda di ciò che si crede sia l'essere umano senza però che questo cambi la natura sostanziale del mundus.

Alfano scrive al Signore:

- g) Mundus ab opposito nomen habere potem.
- h) Mundus erat mundus,
- i) Mundus cum munda creares.

currently most advanced emerging quantum computing systems, which represent a significant departure from traditional computing systems, constitute machine-based systems, despite their unique operational principles and use of quantum-mechanical phenomena, as do biological or organic systems so long as they provide computational capacity (Annex AI Act, 2025).

In (g) Alfano allude alla possibilità di definire il mundus per confronto con ciò che mundus non è. L'approccio ontologico differenziale, di cui Alfano sembra proporsi come precursore, è di vitale importanza per comprendere la natura dell'antropocentrismo e del biocentrismo i cui sostenitori, appunto, definiscono essere umano e vita per antitesi a ciò che umano e vita non sembra essere o non si vuole che sia.

Nei versi successivi Alfano usa una forma tanto criptica quanto geniale nella sua semplicità. Dire che "qualcosa era qualcosa", ovvero che non lo è più, sembrerebbe trasgredire il principio di identità e non contraddizione a meno che "qualcosa" non abbia un duplice significato. E così è. Il primo mundus è riferito al mondo, al cosmo, al creato. Il secondo ha un significato più ampio: pulito, puro, adornato. Almeno limitatamente a questa analisi tornerebbe utile considerare la visione cristiana dell'onniscienza che troverà nuova linfa circa duecento anni dopo con Tommaso D'Aquino. Per quest'ultimo, la conoscenza di Dio è causale, ovvero da riferirsi a un certo ordine causale. Adornato dunque come sinonimo di ordinato, che ubbidiva a un particolare set di regole. Il creato era ordinato (h), quest'ordine è stato creato con un piano preciso, razionale (i).

Quindi, tornando a (g), il mundus può essere definito per differenza con ciò che era prima. E per l'autore *in principio era il caos*, inteso come assenza di ordine (o meglio ancora come nulla, nihilo). Il caos deterministico discusso nel paragrafo dedicato al futuro rientra ovviamente nel mundus. Esclusi dal mundus sono il Creatore stesso e un'altra entità ineffabile: la sorgente randomica, poiché entrambi sono immuni alle regole e per definizione non possono essere né previsti né controllati.

Per comprendere l'essere umano è quindi necessario (non è detto sia sufficiente) assegnargli una collocazione rispetto al mundus poiché, così facendo, possiamo definirlo per confronto con l'immanente (le regole) o col trascendente (ciò che trascende le regole).

Per Alfano di Salerno il mundus è rimasto ordinato fino alla creazione dell'essere umano che grazie al libero arbitrio mina il concetto di onniscienza di Dio.

Ritroviamo questo concetto nella Divina Commedia, nella quale Dante stesso si fa cruccio di testimoniare l'esistenza del libero arbitrio¹⁸. Senza di esso non sarebbe possibile ritenerci responsabili delle

¹⁸ Non a caso Dante riprende il discorso del libero arbitrio con Virgilio, con Marco Lombardo e con Beatrice.

nostre azioni o giudicabili per il nostro comportamento e Dante non potrebbe giustificare l'assegnazione retributiva delle anime al paradiso, purgatorio e inferno.

Ciononostante, e proprio per questa secondaria funzione del libero arbitrio, nella Divina Commedia appare evidente la costituzione di un set di regole di ordine superiore: se commetti il peccato A otterrai la punizione B. Questo getta nuova luce sulla posizione dell'essere umano nel mundus secondo Dante: non siamo in grado di sfuggire alle regole poiché la norma è creata da Dio, è quindi eterna e immutabile, indifferente al divenire.

Eppure, Dante ci racconta del suo desiderio di superare il limite e ricongiungersi con Dio o, per meglio dire, con Beatrice.

Dante vuole arrogantemente contravvenire alla norma divina e ingannare la morte, il destino, e arrivare dove avrebbe potuto solo alla conclusione di una vita degna del paradiso. Egli chiama la riuscita del suo piano: "trasumanar" e questo sarà il primo e unico termine in tutta la Divina Commedia che Dante ammetterà di non saper spiegare: «Trasumanar significar per verba non si poria» (Dante, Paradiso I, vv. 70-71). Egli è nell'aldilà, oltre la condizione umana, non sarebbe capace di spiegarlo a parole perché in quella condizione trascende il linguaggio stesso.

Trasumanar, dunque, per Dante significa uscire dal mundus, benché questo comporti la rinuncia all'essere (essere umani).

Dante non ne fa una questione di pensiero, il trasumanar per lui è fisico, corporeo¹⁹, poiché il limite dell'essere umano è il perimetro della sua esistenza, ovvero ciò che ci separa dal nulla, sia esso quello da cui proveniamo al momento della nascita o quello in cui ripiombare nel momento della morte.

Benché il pensiero che sostiene la trascendenza sia andato raffinandosi con la corrente esistenzialista, ancora oggi è vivo l'ideale di tracotanza che i greci identificavano con *hybris*.

Nell'ambito di questa analisi, i limiti dell'essere umano sono iscritti nel mundus. L'uscita dal mundus comporta l'infrazione di una qualsiasi regola, sia essa di carattere descrittivo-funzionale della natura (leggi della natura) o di carattere normativo (giuridico, sociale, morale, religioso). Come accennato all'inizio di questo paragrafo, la capienza del mundus è definita da chi lo osserva. Il mundus si può interpretare

¹⁹ «Allora incominciai: "Con quella fascia che la morte dissolve men vo suso, e venni qui per l'infernale ambascia» Dante, Purgatorio Canto XVI, vv. 37-39.

come l'insieme che contiene tutti gli insiemi (determinismo), l'insieme vuoto (tutto è randomico o governato dal trascendente) o qualsiasi altra dimensione tra questi due estremi anche a seconda della volontà di prendere in considerazione regole puramente arbitrarie.

Antropocentrismo, umanesimo, post-umanesimo

L'antropocentrismo, benché già dibattuto dai filosofi Greci, assurge a fondamento filosofico con la religione giudeo-cristiana, la quale impone un solo Dio e nel creato un essere (e uno solo) a sua immagine e somiglianza (imago dei).

Nell'arco dei millenni, l'antropocentrismo si è andato raffinando, in grossa parte per l'evoluzione dei detrattori stessi della teoria. Da un antropocentrismo di tipo puramente teologico nel quale l'essere umano segue il destino imposto dal creatore, si arriva a un antropocentrismo di tipo secolare con l'esistenzialismo di Sartre che poneva l'essere umano su di un piedistallo in quanto libero e padrone del proprio destino. È l'era dell'illuminismo il cui valore è nel trionfo della ragione. Il concetto principale è che la conoscenza è positiva per il benessere dell'essere umano (salute, bisogni, felicità). Come ricorda Floridi, la scienza ha rivoluzionato l'antropocentrismo gettando nuova luce sulle caratteristiche dell'essere umano che ora non può più vantarsi di un primato in termini di collocazione nel cosmo (Copernico), primato biologico sugli animali (Darwin), primato della razionalità sulle pulsioni (Freud) e primato sull'interpretazione delle informazioni (Turing) (Floridi, 2017) portando poi al concetto della morte di Dio. Da un antropocentrismo che esalta l'essere umano nei confronti dell'universo tutto o, almeno, nei confronti del regno animale, si è passati a una sorta di antropocentrismo biocentrico nel quale la contrapposizione ontologica avviene tra organismi viventi e materia inorganica (la cui capacità di auto-organizzazione e pseudo-autonomia è rappresentata dal mix di hardware e software che in questo testo ho chiamato macchina).

Come accennato in precedenza, l'unica speranza rimasta all'antropocentrismo è che la mente umana non sia computabile.

Almeno in occidente, inoltre, dal Rinascimento fino agli anni '50 del secolo scorso, l'antropocentrismo ha sempre giustificato l'umanesimo e l'umanesimo ha sempre implicato l'antropocentrismo. Se l'antropocentrismo è la postura descrittiva del pensiero: *l'essere umano è al centro*, l'umanesimo ne è senza dubbio l'istanza normativa: *l'essere*

*umano deve essere al centro, dovrebbe essere al centro*²⁰. Da qui il mantra europeo delle policy *human-centric*. In sostanza si può essere d'accordo con policy *human-centric* o no, dipendentemente dal fatto che si creda che l'essere umano sia destinato al controllo e ad assumersene la responsabilità, o meno. Del resto è sempre stata una dicotomia, platonismo e intuizionismo, essenzialismo ed esistenzialismo, monismo e dualismo, immanenza e trascendenza, antropocentrismo e post-umanesimo (che denota appunto, l'andare oltre l'umanesimo).

Oggi esiste una terza via di pensiero.

È con Kurt Gödel, Alan Turing e Julian Huxley che questo matrimonio tra antropocentrismo e umanesimo inizia a vacillare. L'idea di un umanesimo digitale, introdotto in risposta al nuovo urbanesimo digitale²¹, e la diffusione del transumanismo, hanno irrimediabilmente disaccoppiato l'umanesimo dall'antropocentrismo.

Per i transumanisti è valida l'ambizione al controllo sulla natura anche se non si possiede alcun titolo per reclamarlo (Figura 1).

	Personalismo	Transumanismo	Post-umanesimo
Antropocentrismo	✓	✗	✗
Umanesimo	✓	✓	✗

Figura 1: schema “cosa si può pensare dell'essere umano?”

Nel transumanismo, infatti, si crede possibile il mind uploading (Olson, 2022), ovvero il processo di trasposizione della mente su supporti hardware esterni. Questo ha varie implicazioni tra cui la possibilità di copiare, creare un backup, riavvolgere, condividere, modificare stati mentali come ricordi, replicare, fondere, argomentare le capacità

²⁰ In questo caso essere al centro non implica necessariamente dominio ma controllo.

²¹ Nuovo ambiente, quello digitale, che produce nuove capacità di interazione tra esseri umani.

cognitive. Tutto questo non è possibile nel caso in cui la mente non sia computabile, dunque per i transumanisti l'essere umano non è speciale poiché condivide anche questa proprietà con le macchine e dunque con la materia inanimata.

Il concetto di trascendenza è legato al superamento dei limiti fisici e biologici dell'essere umano, meramente funzionali, tra cui quelli cognitivi e quelli legati alla salute e alla longevità (tipici del mundus). Il trasumanar non è più, come intende Dante un modo per riconciliarsi a Dio ma è un modo per renderci Dio, entità soprannaturali, per esempio creando la vita noi stessi.

A tal proposito, si consideri che in ottica antropocentrica l'essere umano [H] è considerato speciale [H*]. H è speciale (=H*) se possiede almeno una caratteristica [x] che non condivide con nessuno [x*] (l):

(l) $\exists x \mid x = x^*$ caratteristica unicamente umana $\leftrightarrow H=H^*$ essere umano speciale

Si definisca ora l'IA [AI²²] come la disciplina il cui fine è quello di simulare²³ l'intelligenza umana [HI] (Stryker, 2024; Turing, 1950; Gebru, 2024) (m).

(m) AI = Simulazione HI

Del resto, si può facilmente ammettere che la simulazione di un sistema sia una riconfigurazione artificiale del sistema preso in considerazione.

Allora quello che viene considerato da Stuart Russell un successo della specie umana (Russell, 2022), ossia il farci artefici della nascita di una specie che esibisca un tipo di intelligenza paragonabile al nostro,

²² Limitatamente al ragionamento che segue è conveniente la notazione inglese.

²³ Si noti che tale procedimento si basa su una definizione generica di simulazione. Essa può essere intesa come imitazione o come emulazione. Se volessimo imitare il volo degli uccelli avremmo velivoli piumati che sbattono le ali (approccio bottom-up). Avendo optato per l'emulazione del volo degli uccelli ci siamo liberati dai vincoli di forma (hardware) e ci siamo concentrati sulla funzione (software): raggiungere il punto B partendo dal punto A sfruttando una tecnica che ricorda il volo degli uccelli (approccio top-down). Con l'IA, nel perseguimento del fine di simulare l'intelligenza umana, si tentano entrambi gli approcci top-down (orientato ai risultati) e bottom-up (sensibile alla struttura, ad. es. neural network).

non sembra essere un evento che gli antropocentristi celebrerebbero perché è uno degli eventi che sancisce che l'essere umano, almeno sul piano dell'intelligenza, speciale non è.

Generalizzando la (m):

(n) $A(x) = \text{simulazione di } H(x) \mid x = \text{caratteristica umana}$

Con (n) possiamo estendere il pensiero precedente a tutte le caratteristiche umane. Se il nostro essere unici proviene dall'abilità di replicare noi stessi, allora tale replica, in quanto fedele, contraddice l'ipotesi di unicità dell'essere umano (o):

(o) $A(x) = H(x) \rightarrow H \neq H^* \leftrightarrow \nexists x^* \text{ caratteristica unicamente umana}$

Si noti che questo risultato non dimostra che l'essere umano non sia speciale né che non si possa replicare artificialmente l'essere umano, semplicemente perché ciò che pensiamo dell'essere umano non è detto che corrisponda a realtà. Ad esempio, potremmo condividere il pensiero greco secondo cui l'essenza umana risiede nel suo essere mortale. La caratteristica $x=c$ di caducità [c] è però abbastanza semplice da simulare. Benché, come sostiene Ferraris, l'essere umano non possa essere riacceso dopo che si è spento così come è possibile riaccendere una macchina, il problema di simulazione si risolve creando macchine che non possano essere riaccese una volta spente così come accade per gli esseri umani e non viceversa. Rimane molto più difficile simulare la non caducità, ovvero l'immortalità, se non altro perché è la materia stessa, organica o inorganica, ad essere destinata al decadimento. Proprio per questo principio si considera che la questione della longevità sia legata ai limiti del supporto materiale, organico o inorganico, su cui insiste il fenomeno della vita.

h^+ , come si identifica il movimento filosofico transumanista, ambisce al superamento dei limiti umani (super capacità cognitive, super salute, super longevità) grazie alla tecnologia o, più nello specifico, al *merging* (fusione) con le macchine. Il controllo sulla natura si esplica attraverso la capacità di sfuggire al destino che ci vede soccombere per mano della natura stessa, intesa anche come inesorabilità dello scorrere del tempo o caducità.

Il transumanesimo, nasce e viene riproposto nei paesi di lingua anglosassone e spesso viene confuso con il post-umanesimo per via del fatto che in inglese *transhumanism* e *post-humanism* hanno lo stesso

suffisso -ism. In Italiano è bene distinguere tra transumanesimo, inteso come il movimento che aspira ad andare oltre i limiti umani e post-umanesimo, inteso come il movimento critico nei confronti dell'umanesimo.

Il personalismo è una delle concrete manifestazioni dell'antropocentrismo. Nasce con Tommaso d'Aquino, la cui dottrina guarda l'essere umano come a un'essenza ontologica irriducibile al mondo naturale, la cui dignità emerge con la capacità di autodeterminazione (Wojtyła, 1993). Il soggetto persona è identificato dalla dignità e il destinatario dei diritti umani. Per i personalisti tra gli elementi dell'insieme delle persone e gli elementi dell'insieme degli esseri umani esiste una corrispondenza biunivoca, non a caso la Dichiarazione Universale dei Diritti Umani ha una forte connotazione personalista.

Diversamente dal transumanesimo, il personalismo fa affidamento al principio terapeutico secondo il quale scienza e tecnologia sono indispensabili ma la cieca fiducia nella razionalità tecnica è un errore. Secondo il principio terapeutico ogni atto è moralmente giustificabile se ha un fine terapeutico (Petrini, 2017), ovvero una funzione di ripristino delle "normali" condizioni di funzionamento o di riequilibrio verso ciò che si crede rappresenti l'essenza umana.

Allora la differenza sostanziale tra personalismo e transumanesimo sta nel definire l'essere umano nei confronti della tecnica. Per i primi, ispirati da Rousseau, esiste una condizione della natura umana assoluta che è rintracciabile solo spogliando l'essere umano di qualsiasi stampella tecnologica. L'essere umano è ontologicamente determinato, dunque isolabile da ogni contesto, sociale o naturale che sia, e la sua essenza emerge proprio quando è preso in considerazione fuori dell'universo o mundus, proprio come sarebbe possibile aspettarsi per un'entità sovrannaturale o una sorgente randomica. Per Rousseau, infatti, l'essere umano è corrotto dalla società, dal progresso, financo dalla tecnica come sostiene Ferraris (Ferraris, 2024), ma per i personalisti l'essere umano è corrotto persino dalla natura stessa che attraverso quelle che Freud definisce pulsioni (non lontane dagli istinti animali) mette a repentaglio la purezza della ragione e della moralità umana.

Quella condizione trascendentale è quella a cui, in teoria, il personalista aspira grazie al principio terapeutico.

Per i transumanisti, invece, l'essere umano è tecnica. Non semplicemente perché basterebbe il pollice opponibile o la posizione eretta per reclamare il titolo di cyborg (Todaro, 2022), ma perché gli esseri umani semplicemente non esistono "prima" della tecnica (Ferraris, 2021). In quest'ottica non c'è nessuna ragione per i transumanisti di

titubare della natura positiva della tecnica e la volontà di potenza è interpretata come massima razionalità (laddove invece per gli antropocentristi la volontà è superiore alla ragione). Non solo. Per i transumanisti è ammesso pensare che l'IA, o un *merging* con essa, sia la "naturale" evoluzione dell'essere umano se non addirittura l'unico modo per garantire la sopravvivenza di ciò che, pur ostinandosi a chiamare specie umana, perderebbe l'essenza dell'essere umano. Qualora l'abbia mai avuta. Non è chiaro invece se per i personalisti l'essenza superi l'esistenza, ovvero se esistano contesti in cui l'essere umano sia da ritenersi sacrificabile per salvaguardarne l'essenza, ovvero se il mutamento dell'essenza umana sia da considerarsi esso stesso un evento che determina la cessazione dell'esistenza umana²⁴. Da una parte il principio terapeutico che pone il focus sull'essenza, dall'altra il diritto alla vita come bene da difendere a ogni costo.

Infine, il pensiero post-umanista sulla razionalità tecnica è opposto a quello transumanista. La tecnica non è volontà poiché la volontà non è di questo *mundus*. Rifiutando di elevarsi a dominatori, i post-umanisti hanno una visione privilegiata della tecnologia, scevra di ogni vana promessa. Dunque, la tecnologia non è semplicemente il mezzo attraverso il quale si perpetra la devastazione dell'ecosistema ma ne è la causa stessa. Oltre questo, la tecnologia non è solo il mezzo attraverso il quale si abilitano le disuguaglianze sociali ed economiche e si realizzano l'oppressione e la sorveglianza, ma ne è la causa prima. Del resto è difficile che la voce post-umanista (quella vera, non quella transumanista camuffata) emerga nel dibattito per via del relativismo morale e dell'atteggiamento nichilista, più spesso di tipo passivo, che ne caratterizza i sostenitori. Il motivo è che i post-umanisti non credono nel libero arbitrio e da questa premessa la condizione umana assume tutto un altro peso. Ciononostante, Robert Sapolsky sostiene che la consapevolezza di essere determinati possa avere lati estremamente positivi come, ad esempio, il superamento della cultura della colpa e quella della vergogna. Senza libero arbitrio non avrebbero senso né il giudizio né il controllo sociale che derivano dalla deterrenza. L'umano pericoloso, così come l'animale pericoloso, benché non sia possibile assegnargli responsabilità, può lo stesso essere messo in cattività ma come forma di quarantena, non per espiare le proprie colpe, con tutte le ricadute psicologiche che questo comporta (Sapolsky, 2023).

²⁴ La clonazione umana, ad esempio, rimarrebbe un'azione moralmente deprecabile se fosse l'unica soluzione per scongiurare l'estinzione?

Conclusioni

Qualora il futuro sia aperto (a) e l'essere umano sia padrone del proprio destino (b), il futuro dell'IA dipende dalla governance, ovvero dall'evoluzione del dibattito politico che è a sua volta legato alle sfumature del pensiero filosofico trattato nelle sezioni precedenti.

Ciò che emerge dall'analisi del dibattito filosofico e politico è una certa polarizzazione delle credenze. Per alcuni la differenza tra essere umano e macchina sta nel fatto che il primo abbia la capacità di pensare e il secondo no, mentre per altri anche le macchine possono esibire caratteristiche di coscienza. Ciò che caratterizza i due gruppi sono premesse arbitrarie, come ad esempio quella secondo la quale il pensiero sia (o meno) computabile. Questa premessa influenza il concetto di autonomia delle decisioni, rendendo ambigue le definizioni di sistemi di IA e approssimativi i tentativi di attribuzione di responsabilità. Le posizioni sono inconciliabili e difficilmente potremo aspettarci una mediazione tra le parti poiché non sembra esistere un corpus minimo di elementi condivisi e condivisibili su cui avviare un dialogo (Palazzani, 2023). L'incertezza è ontologica ma si rende necessario lavorare su intesa e cooperazione per favorire una governance efficace e ridurre almeno quell'incertezza epistemica creata dal passaggio senza filtri del discorso filosofico a quello politico. In questo contesto, i proclami sull'inevitabilità di alcuni eventi reclutano invece di istruire ed educare alla pluralità di pensiero e favoriscono l'incertezza normativa.

Nel lungo periodo, il futuro dell'IA è legato alla verifica dell'ipotesi nulla (b), del resto il numero di Turing Test falliti da una macchina che consideriamo sufficienti per ammettere che essa non sia in grado di pensare appare essere ancora assolutamente arbitrario.

L'idea che è possibile avere sul futuro di breve periodo dell'IA è in gran parte orientata a capire se l'hype tenderà ancora ad aumentare o diminuirà (permettendo ipotesi sull'inverno dell'IA).

Pertanto, visti i recenti sviluppi geopolitici che vedono gli USA indietreggiare sulla regolamentazione, la Cina incalzare nella corsa alle armi all'IA e l'Europa interrogarsi su come recuperare il gap, chi si interessa dello sviluppo degli scenari sul futuro dell'IA dovrebbe tener conto del fatto che le politiche human-centric europee stiano diventando sempre meno antropocentriche. Questa particolare deriva prometeica è dovuta in parte allo strapotere tecnico transumanista e in parte dalla crescente esigenza di un anti-antropocentrismo biocentrico ed ecologista. Se questo trend si conferma è più facile immaginare

ancora un incremento di hype piuttosto che una diminuzione, almeno nel breve, il che induce anche a pensare ai rischi legati all'archetipo di sistema denominato *success to the successful*.

Lo sviluppo e la diffusione dell'IA rappresentano un costo non indifferente in termini ambientali (energetici, di consumo delle risorse, dell'inquinamento), sociali (disinformazione, depressione, digital divide e digital health) e di governance (aumento della complessità e derive ideologiche) ma l'IA è anche uno degli strumenti che usiamo per risolvere tali problematiche (Vinuesa, 2020, AI for SDGs). Qualora si debba puntare su ulteriori investimenti o frenare lo sviluppo e la diffusione dei sistemi di IA dipende dalle fondamenta del nostro pensiero nei confronti dell'essere umano, ovvero se crediamo che l'IA sia problema o soluzione²⁵.

La complessificazione del mondo, in parte proprio grazie allo sviluppo e alla diffusione dei sistemi di IA in tutti i settori (Harris, 2022), al feroce dibattito che coinvolge gli esperti e alla conseguente crescente incertezza, ha pure la conseguenza di far percepire gli altri player come più minacciosi (più intelligenti e più malvagi) di quel che sono realmente. Questo, tra l'altro, induce a pensare che anche l'IA rappresenti una minaccia più grande di quello che sia realmente²⁶.

Considerazioni finali:

L'antropocentrismo non definisce l'IA in base alla nostra posizione ma la nostra posizione rispetto all'IA. Come quando il dorsista spinge con i piedi contro il bordo della piscina, così l'accanimento nel mostrare la macchina quanto più stupida e inutile possibile ha il solo scopo di spingerci in direzione opposta. Di elevarci metaforicamente.

I tecno-ottimisti fanno lo stesso in verso opposto. Ciò non cambia l'essenza dell'essere umano, né tantomeno quella della macchina, ma il dibattito sempre più aspro e inutilmente tragico aumenta l'incertezza.

La deriva Prometeica che coinvolge gli antropocentristi è in parte causata dal fatto che anche i transumanisti sono umanisti. Del resto è difficile immaginare correnti politiche strutturate al punto da giustificare una valida opposizione al mantra human-centric.

La deriva prometeica si osserva anche nel pensiero di (Floridi, 2016). Il suo *antropo-eccentrismo* che si esplica nella frase «siamo un beautiful glitch», è un tentativo come un altro atto ad allargare il focus

²⁵ Usare l'IA come soluzione scopre il fianco al rebound effect.

²⁶ In effetti, non siamo nuovi alla paura che qualcuno ci rubi il lavoro.

morale figlio di un'ontologia ancora fortemente umanista ma disillusa nei confronti dell'antropocentrismo. Secondo Floridi «Our exceptionalism lies in a special and perhaps irreproducible way of being successfully dysfunctional. We are nature's beautiful glitch, in a universe-system that has fortuitously and probably uniquely generated a form of life most unlikely to occur again, and certainly anomalous and strange. We are endowed with consciousness, intelligence, mental life, and self-determination.» Le informazioni importanti che ci raccontano il pensiero di Floridi sull'essere umano sono:

siamo un fenomeno naturale,

e fortuito (probabilmente tanto improbabile da rivelarsi unico).

Il glitch, del resto, è una forma di rumore e, come argomentato nella nota 12: tolto l'essere umano dalle sorgenti di rumore (nature's beautiful glitch), esso è un fenomeno deterministico. Determinati e padroni del nostro destino al tempo stesso dunque. Ovviamente questo è un modo come un altro per partire da premesse arbitrarie e nella fattispecie particolarmente inconsistenti. Il problema sorge nel momento in cui da questa premessa si arriva a giustificare la dignità umana e il diritto alla privacy.

Un potenziale prosieguo della ricerca potrebbe essere quello di analizzare una quarta prospettiva sull'essere umano, quella che nella Figura 1 avrebbe un segno verde in corrispondenza dell'antropocentrismo e una x rossa in corrispondenza dell'umanesimo. È possibile che qualcuno sostenga l'eccezionalismo umano ma anche che questo non debba essere ostentato. L'unico occidentale che potrebbe in qualche modo avvicinarsi a questa posizione è Francesco d'Assisi ma troverei più promettente guardare a oriente anche per rendere la ricerca meno occidental-centrica.

Bibliografia

- FLI, Should we slow down AI research? | Debate with Meta, IBM, FHI, FLI, YouTube Video, 7 Maggio 2024: <https://www.youtube.com/watch?v=pYPfUsuJwk4>
- Schubert S., Caviola L., Faber N.S., *The Psychology of Existential Risk: Moral Judgments about Human Extinction*. Sci Rep 9, 15100, 2019. <https://doi.org/10.1038/s41598-019-50145-9>
- Calabresi G., Bobbitt P., *Tragic Choices*, New York, W. W. Norton & Company, Political Science Quarterly, Volume 93, Issue 3, pp. 506–507, 1978: <https://doi.org/10.2307/2149545>, 1978.
- Bengio Y. et. al., “*International AI Safety Report*” (DSIT 2025/001, p. 14, 2025; <https://www.gov.uk/government/publications/international-ai-safety-report-2025>
- Future of Life Institute (FLI), Pause Giant AI Experiments: An Open Letter, 2023. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- OpenAI, Planning for AGI and beyond, 2023. <https://openai.com/index/planning-for-agi-and-beyond/>
- Taylor J., Rise of artificial intelligence is inevitable but should not be feared, ‘father of AI’ says, The Guardian, 2023 <https://www.theguardian.com/technology/2023/may/07/rise-of-artificial-intelligence-is-inevitable-but-should-not-be-feared-father-of-ai-says>
- Fuller T. et al., *Responsible futures*, Chapters, in: Roberto Poli (ed.), Handbook of Futures Studies, chapter 19, pages 259-279, Edward Elgar Publishing, 2024
- Floridi, L. AI and Its New Winter: from Myths to Realities. Philos. Technol. 33, 1–3, 2020. <https://doi.org/10.1007/s13347-020-00396-6>
- The Artificial Intelligence Act - Regulation (EU) 2024/1689, Chapter I: General Provisions, Article 3: Definitions, 2024.
- OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, 2019: <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>.
- OECD, *Explanatory Memorandum On The Updated Oecd Definition Of An AI System*, OECD Artificial Intelligence Papers March 2024 No. 8, 2024.
- Fabris, A., *Ripensare l'agente morale. Al di là di analitici e continentali*. Filosofia Morale/Moral Philosophy, (1), 155-164, 2022. Retrieved from <https://mimesisjournals.com/ojs/index.php/MF/article/view/1747>
- ANNEX to the Communication to the Commission, Approval of the content of the draft Communication from the Commission - Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act), 6 febbraio 2025.
- Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mor-datch, L., *Emergent Tool Use From Multi-Agent Autocurricula*, 2019: ArXiv, abs/1909.07528.

- Russell S., *Human Compatible: Artificial Intelligence and the Problem of Control*. United States: Viking, 2019: ISBN 978-0-525-55861-3.
- Stryker C., Kavlakoglu E., *What is AI?*, “IBM”, 2024: <https://www.ibm.com/think/topics/artificial-intelligence>
- Turing A., *Computing Machinery and Intelligence*, *Mind*. 59 (236): 433–460. doi:10.1093/mind/LIX.236.433, 1950.
- Gebru, T., Torres, É. P., *The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence*. *First Monday*, 29(4), 2024: <https://doi.org/10.5210/fm.v29i4.13636>
- Floridi L., *What the Near Future of Artificial Intelligence Could Be*. *Philosophy and Technology*, 2020: Available at SSRN: <https://ssrn.com/abstract=3570424> or <http://dx.doi.org/10.2139/ssrn.3570424>
- Russell S., *If We Succeed*. *Daedalus* 2022; 151 (2): 43–57, 2022: doi: https://doi.org/10.1162/daed_a_01899
- Vulpiani A., *Caso, probabilità e complessità*, Ediesse, p. 106, 2014.
- Severino E., *La legna e la cenere*, Rizzoli, pp. 18-19, 2000
- Maxwell J. C., *Illustrations of the dynamical theory of gases*, *Philosophical Magazine* 19 (1860); in *The Scientific Papers of James Clerk Maxwell*, W. D. Niven. ed., Dover, New York, p. 442, 1965.
- Atmanspacher H., *Determinism is ontic, determinability is epistemic*, In Harald Atmanspacher & Robert Bishop (eds.), *Between Chance and Choice: Interdisciplinary Perspectives on Determinism*. Thorverton UK: Imprint Academic. pp. 49--74, 2002.
- Hameroff S., Penrose R., “Reply to seven commentaries on “Consciousness in the universe: Review of the ‘Orch OR’ theory”, *Physics of Life Reviews*. 11 (1): 94–100, 2014
- Mauro D’Ariano G., Faggin F., “*Hard Problem and Free Will: an information-theoretical approach*”, 2020: arXiv e-prints, Art. no. arXiv:2012.06580, 2020. doi:10.48550/arXiv.2012.06580.
- Nielsen, M. A., & Chuang, I. L., *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge: Cambridge University Press, 2010.
- Essentia Foundation, *Quantum Consciousness Debate: Does the Wave Function Actually Exist? | Penrose, Faggin & Kastrup*, Youtube Video, 25 Agosto 2024, <https://www.youtube.com/watch?v=0nOtLj8UYCw>
- Salerno A., *Da Carmina, XXXIII, Item oratio, seu confessio metrica*, citato in Antonio Viscardi, *La Letteratura italiana: Storia e testi*, vol. 1, *Le origini*, Ricciardi, p. 387, 1951
- Floridi L., *La quarta rivoluzione*, Raffaello Cortina Editore, Milano, p. 106, 2017.
- Olson E. T., *The Metaphysics of Transhumanism*. In Karolina Hübner (ed.), *Human: A History (Oxford Philosophical Concepts)*. New York, NY: Oxford University Press. pp. 381-403, 2022.
- Wojtyła K., *Thomistic Personalism*, Theresa Sandok (trans.), in *Person and Community: Selected Essays (Catholic Thought from Lublin: Volume 4)*, Andrew

- N. Woznicki (ed.), New York: Peter Lang, pp. 165–75, 1993.
- Petrini C. Bioethics of Clinical Applications of Stem Cells. *Int J Mol Sci.*;18(4):814, 2017.
doi: 10.3390/ijms18040814. PMID: 28417921; PMCID: PMC5412398.
- Ferraris M., *Intelligence as a human life form*, *Journal of Responsible Technology*, Volume 18, 100081, 2024.
- Todaro M., *Cyborgs Outsmarting Attention and Philosophical Implications*, Michałowska, M., *Crossing the Border of Humanity: Cyborgs in Ethics, Law, and Art*. Proceedings of the International Online Conference December 14–15, 2021, Medical University of Łódź, Poland, pp. 101-106, 2022:
DOI: <https://doi.org/10.6084/m9.figshare.18093383.v1>
- Ferraris M., *In Praise of the Anthropocene*, contenuto in Luisetti F., “The Speculative Migrants of the Anthropocene. Human Flows in the Neoliberal Planet.” *Itinerari* 59, no. 2, 2021: doi:10.7413/2036-9484023.
- Sapolsky R., *Determined: a science of life without free will*. New York: Penguin Press, 2023.
- Palazzani L., *Bioethics on dialogue*. *Medicina E Morale*, 72(2), p.145. 2023. <https://doi.org/10.4081/mem.2023.1232>
- Vinuesa R., Azizpour H., Leite I. et al. *The role of artificial intelligence in achieving the Sustainable Development Goals*. *Nat Commun* 11, 233. 2020: <https://doi.org/10.1038/s41467-019-14108-y>
<https://ai-for-sdgs.academy/observatory>
- Harris T., *The wisdom gap*, Center for Humane Technology, 2022: <https://www.humanetech.com/insights/the-wisdom-gap>
- Floridi L., *On Human Dignity as a Foundation for the Right to Privacy*, *Philos. Technol.* 29, 307–312, 2016: <https://doi.org/10.1007/s13347-016-0220-8>