GIANLUCA GIANNINI – ANTONIO PESCAPÈ
[LUCA LO SAPIO]

AI E FUTURO DI *SAPIENS* TRA NUOVI ORIZZONTI E ANTICHI TIMORI

ABSTRACT: ***AI AND FUTURE OF SAPIENS BETWEEN NEW HORIZONS AND ANCIENT FEARS***
AI in its various forms and articulations represents a range of new opportunities and strategies that our species can use to improve its condition. Often the scenarios presented in association with AI speak of a humanity threatened by machines or, on the contrary, of a utopian future in which man and machine will be completely hybridised in a harmonious manner.
In this interview, a moral philosopher, Professor Gianluca Giannini, and Antonio Pescapè, professor of computer engineering, will try, each from their own perspective, to answer some of the key questions emerging in the field of artificial intelligence, on the one hand avoiding sensationalist journalistic exaggerations, and on the other hand highlighting that the real challenge posed by intelligent machines lies in the new ways in which humans will narrate themselves and their possibilities of existence.

S&F: Ci sono numerose definizioni di Intelligenza Artificiale. Se rivolgiamo la nostra attenzione a "intelligenza", ad esempio, troviamo almeno una sessantina di definizioni[1]. D'altro canto, anche la definizione di artificiale presenta delle difficoltà, dal momento che spesso viene strutturata in opposizione a naturale. Dunque, ci sono evidenti difficoltà sia a definire che cosa sia l'intelligenza, sia nel definire che cosa sia l'artificiale. Ritenete sia possibile trovare un comun denominatore in grado di mettere d'accordo le diverse prospettive?

A.P.: Sicuramente è più che complesso, se non impossibile, trovare un comune denominatore in grado di "accordare" le diverse prospettive, soprattutto perché risulta piuttosto difficile trovare una definizione unica di Intelligenza Artificiale. Volendo

[1] S. legg, M. Hutter, *Universal intelligence: a definition of machine intelligence*, in «ArXiv: 0712.3329 [cs.AI]», https://doi.org/10.48550/arXiv.0712.3329.

partire da un punto di vista tecnologico, quando si parla di Intelligenza Artificiale ci si riferisce a un ampio insieme di tecnologie e, di conseguenza, a uno scenario altrettanto ampio nel quale trovano spazio applicazioni molto diverse tra loro. Oggi quando discutiamo *lato sensu* di Intelligenza Artificiale parliamo di aree tematiche quali *Computer Vision, Pattern Recognition, Automated Reasoning, Game theory, Logics, Multi-Agents, Fuzzy systems, Knowledge representation, Speech Recognition, Natural Language Processing, Machine Learning, Deep Learning, Cognitive Robotics*. Tutte queste aree tematiche trovano la loro collocazione (fonte Scimago) in più di 165 riviste internazionali e sono state alla base di più di 900 conferenze internazionali di settore; si può facilmente intuire come il termine Intelligenza Artificiale sia, quindi, un termine *"umbrella"* sotto il quale si trovano cose molto diverse tra loro. A quanto finora detto va aggiunta un'altra importante distinzione, quella tra *Strong AI* e *Weak AI*. Con *Strong AI* ci si riferisce a scenari nei quali le macchine sono effettivamente in grado di pensare ed eseguire compiti da sole, essere veri e propri *alias* di un essere umano. In questa categoria sono ricompresi il *Turing Test* (Turing), il *Coffee Test* (Wozniak), il *Robot College Student Test* (Goertzel), l'*Employment Test* (Nilsson), il *Flat Pack Furniture test* (Tony Severyns), il *Mirror Test* (Tanvir Zawad). Con *Weak AI* ci si riferisce invece a scenari in cui ci si concentra su un compito ristretto, riferendosi al fenomeno per il quale le macchine che non sono troppo intelligenti per svolgere il proprio lavoro possono essere costruite in modo tale da *sembrare* intelligenti, simulando solo la funzione cognitiva umana. Questo tipo di AI "debole" agisce in maniera "semplice" ed è vincolata dalle regole che le vengono imposte senza mai poter agire oltre tali regole. Al di là di queste due definizioni, *Strong AI* e *Weak AI*, la tassonomia dell'Intelligenza Artificiale si specializza ulteriormente in altri quattro tipi:

· AI di Tipo 1: si riferisce a una AI reattiva e specializzata in una singola area. Ad esempio, la redazione e la revisione di contratti di finanziamento di natura commerciale.

· AI di Tipo 2: si riferisce a una AI che ha memoria o "esperienza" appena sufficiente per prendere decisioni adeguate ed eseguire azioni appropriate in situazioni o contesti specifici.

· AI di Tipo 3: si riferisce a una AI che ha la capacità di comprendere pensieri ed emozioni che influenzano il comportamento umano.

· AI di Tipo 4: si riferisce a una AI come quella tipicamente rappresentata in TV (nelle rappresentazioni cinematografiche/nella fantascienza). Le macchine che utilizzano questo tipo di intelligenza artificiale sono autocoscienti, super-intelligenti, senzienti e coscienti.

G.G.: Anzitutto consentitemi di fare due considerazioni di "contesto". Una apparentemente più complessiva e che parte da una

domanda di fondo: "perché tanto interesse per l'AI?". La risposta è sotto gli occhi di tutti: come ci è stato appena ricordato da Antonio, il termine *umbrella* cela letteralmente un pluriverso. Lo si ritrova, il senso complessivo di questo pluriverso, persino in approcci meramente illustrativi e/o divulgativi *a buon mercato*. È ben descritto, ad esempio, dalla docuserie *The Age of A.I.* andata in onda su YouTube, che nel concentrarsi su tutta una progressione di dettagliate ricostruzioni di situazioni e contesti (dall'occasione domestica all'ambito delle fabbriche di automobili, passando per le sale operatorie, arrivando alla *driverless car*), mostra chiaramente come la tecnologia *smart* abbia già plasmato il mondo in cui ci muoviamo attimo per attimo. Va infatti detto che l'AI ha una potenzialità di interazione, trasformazione e rideterminazione del contesto che può essere, potrà essere, unico nella storia di *sapiens*. I programmi di AI stanno infatti già mutando e trasformando ampi campi dell'agire umano: dall'economia a quelli più *banalmente* del quotidiano. Si pensi agli assistenti vocali (sempre più diffusi nei dispositivi elettronici); alle ANN (*Artificial Neural Network*) ovvero le reti neurali artificiali che setacciano i Big Data in cerca di modelli che serviranno a prevedere tendenze, gusti, desideri… E così via. Anzi forse si può dire, con buona dose di certezza, che l'ombrello omnicomprensivo che sottende l'acronimo AI, al momento e in linea di principio, concerne forme applicative d'*intelligenza* automatica limitate a compiti ben definiti in ambiti ben precisi che procedono, però, in direzioni plurime al punto da investire tutta la sfera dell'agire umano.

La seconda considerazione è legata proprio ai termini in questione. AI, *Intelligenza Artificiale*, infatti, è una sorta di ossimoro: intelligenza, cioè un qualcosa di naturale coniugato ad artificiale, che letteralmente è: non-naturale. Che intendiamo allora con AI? In realtà se ci riferiamo agli albori[2] «far sì che una macchina agisca con modalità che sarebbero definite intelligenti se un essere umano si comportasse allo stesso modo», ci troviamo già in un incavo problematico. È infatti un'espressione *controfattuale*, perché anzitutto è del tutto evidente che quello cui si mira non ha nulla a che vedere con il pensiero bensì con il comportamento. *Se l'essere umano si comportasse… sarebbe definito intelligente*: non significa che la macchina sia intelligente o che addirittura stia pensando. In questo allora, forse, e me ne rendo conto in via del tutto esemplificativa, soccorrono le definizioni per restringere il perimetro. In linea di massima, io mi riferisco a due direttrici definitorie per poi arrivare a una terza che potrebbe fungere in qualche modo da "collante". E quindi una definizione *teorico-disciplinare*, per cui l'AI è quella «disciplina, appartenente all'informatica che studia i fondamenti teorici, le metodologie e

[2] A. Turing, *Computing machinery and intelligence*, in «Mind», LIX, 236, pp. 433-460.

le tecniche che permettono di progettare sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico delle *prestazioni* che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana», da cui il termine chiave in cui si risolve il "comportamento" è il rassicurante (?) "prestazione". Poi, in seconda battuta, una definizione *applicativa*, per cui «l'obiettivo non è quello di replicare o simulare l'intelligenza umana» bensì «quello di *riprodurre* o *emulare* l'intelligenza umana, in quanto non vi è alcun motivo a priori che impedisca che *talune prestazioni* dell'intelligenza umana possano essere fornite anche da una macchina». È del tutto evidente che nel caso dell'emulazione le prestazioni 'intelligenti' sono ottenute utilizzando dei meccanismi propri della macchina, eventualmente differenti da quelli ipotizzabili per l'uomo, ma appunto tali da fornire *funzioni/risultati* qualitativamente equivalenti e quantitativamente superiori a quelli umani. Qui nell'intreccio riproduzione-prestazione, emulazione-funzioni-risultati si gioca la ricaduta di ordine applicativo giacché l'obiettivo dell'AI è costruire entità intelligenti *capaci-di*. Cioè macchine in grado di calcolare come agire in modo efficace e sicuro in un'ampia varietà di situazioni. Ovviamente, in maniera diretta e/o indiretta, avendo l'uomo come fine. Tuttavia qui, mi permetto, è possibile inserire quella che considero una definizione *filosofica* (e spero che Antonio non se ne abbia a male) di AI. Se quello che possiamo circoscrivere come modello standard di AI riguarda l'agire razionale, in altri termini il fine è quello di metter su un'entità, un soggetto agente intelligente, cioè un soggetto agente intelligente *ideale* che intraprenda in ogni situazione la migliore (l'aggettivo è meramente descrittivo) delle azioni possibili, abbiamo già posto le basi per un qualcosa di assolutamente inedito.
Ovvero di un peculiare e diverso *artificium* che fondi la sua singolarità, in qualche modo emulativa di quella dell'uomo, in una sorta di autonoma creatività sorretta da un paradigma costitutivo *ingaggio-esperienza-apprendimento-fare* tipica del vivente. Questo peculiare *artificium* sarebbe in tutto e per tutto un'alterità assoluta che, sfuggendo alle dinamiche interpretative passate rispetto ai canoni tradizionali di 'macchina', si staglierà per l'uomo come un vero e proprio interlocutore-altro. E questo lo dico senza cedere alle tentazioni dei tecnofobici e/o, al contrario, dei tecnoentusiasti.

S&F: Una delle dicotomie più influenti nella storia dell'intelligenza artificiale è, come ci ha appena ricordato il Prof. Pescapè, quella tra una Strong AI e una Weak AI o, in altri termini, un'intelligenza artificiale cognitivo-produttiva e un'intelligenza artificiale ingegneristico-riproduttiva. Le linee

di ricerca più recenti sembrano aver puntato più su quest'ultima che sull'intelligenza artificiale forte. Ha forse ragione, dunque, Floridi[3] quando sostiene che per comprendere l'intelligenza artificiale bisogna partire dalla constatazione del divorzio tra intelligenza e capacità di agire? In tal senso l'AI sarebbe più una nuova capacità di agire che una riproduzione dell'intelligenza umana.

A.P.: Concordo con le linee di ricerca recenti: anche io nel mio laboratorio sto portando avanti approcci tipici della *Weak AI*, applicati ai dati di rete così come ad altri domini applicativi come quelli della salute, ambito nel quale l'AI può fare molto. E provo a dire qui, in parole semplici, il principio-base che guida gli approcci di *Weak AI* e il perché oggi gli sforzi maggiori, in termini sia di ricerca sia di innovazione, si profondono qui e non nella *Strong AI*. L'intuizione di fondo è quella che, come nel caso di un bambino, si può fare (o provare a fare) qualcosa non perché la si sappia fare bensì perché si è osservato qualcuno (un adulto, ad esempio) fare quella o una cosa simile (che è il darwininsmo della macchina?). Questa intuizione negli approcci di *Weak AI* si sostanzia nella fase di apprendimento (training) – durante la quale si costruisce il modello addestrando una rete – e in quella di uso (testing) – durante la quale si valuta la capacità del sistema addestrato di operare su dati diversi da quelli utilizzati nella fase di addestramento. Il *Machine Learning* (ML) – e il più recente *Deep Learning* (DL) – fa parte di questa categoria. Sebbene approcci di *machine learning* fossero stati già proposti alla fine degli anni '50, l'*hype* a cui abbiamo assistito e a cui stiamo assistendo è legato alla combinazione di tre importanti condizioni: disponibilità di considerevole potenza di calcolo a costi contenuti, disponibilità di considerevole spazio di memorizzazione a costi contenuti (entrambi abilitati anche dal paradigma cloud), infine disponibilità di quantità di dati mai viste prime. Grazie a queste tre condizioni, gli algoritmi di ML e di DL oggi cominciano ad avere performance tali da poter prevedere un loro utilizzo massivo in ogni attività quotidianamente.

G.G.: Mi sembra di aver cominciato a dire qualcosa in tal senso, laddove ho ricostruito il perimetro entro cui calare questo macro-fenomeno che porta le insegne "AI" e ho poi accennato al paradigma interpretativo cui in qualche modo riferirsi: quello di *alterità assoluta*. La capacità di agire è indefettibilmente la prerogativa dell'*artificium* che è progettato e realizzato quale punto di caduta di un insieme di tecniche computazionali ispirate dal modo in cui gli esseri umani utilizzano il proprio sistema nervoso e il

[3] Cfr. L. Floridi, *Etica dell'intelligenza artificiale. Sviluppi, opportunità sfide*, Raffaello Cortina Editore, Milano 2022. Dello stesso autore cfr. anche il meno recente *Infosfera. Etica e filosofia nell'età dell'informazione*, Giappichelli Editore, Torino 2009.

proprio corpo per sentire, imparare, ragionare e agire. È del tutto evidente, e proprio attraverso l'uso di modelli computazionali, che l'*artificium*, a partire dal proprio "linguaggio costitutivo" e dalla propria materiale difformità, opererà – e non a caso utilizzo questo verbo per sottolineare ancora una volta che trattasi di questione legata all'*agere* e non all'*intelligere* – in maniera del tutto diversa dall'uomo. È questo, in maniera del tutto avalutativa, che mi fa dire che si apre lo spazio di un altro correlazionale. Il richiamo alla *Strong AI*, che prevede un automa cosciente della sua intelligenza e capace quindi di poter agire in modo indipendente, e la *Weak AI*, che prevede invece un automa limitato a un'attività specifica per cui è progettato senza alcuna possibilità di espansione autonoma – e sappiamo che al giorno d'oggi per la difficoltà di costruire delle vere *Strong AI*, con totale autonomia e capacità di adattamento, si è dato maggior credito all'approccio *Weak*, che rappresenta la maggior parte delle AI con cui abbiamo a che fare nel quotidiano – non mi sembra dirimente segnatamente a questo aspetto che pongo in rilievo. Del resto, mi sembra chiaramente emerso già dalla prima risposta.

S&F: ***Alcuni studiosi (ma non solo) impegnati nel campo dell'AI – da Bostrom a Stephen Hawking passando per Elon Musk e Max Tegmark – mettono in guardia dai suoi potenziali pericoli. Credete che si tratti di meri scenari distopici o immaginate il futuro della nostra specie realmente minacciato dalla diffusione dell'AI?***

A.P.: L'utilizzo massivo di algoritmi di *Weak AI* apre un tema importante legato a una serie di questioni cruciali quando si utilizzano approcci di *Machine/Deep Learning*; questioni che oggi rappresentano la frontiera della ricerca in questo settore. In particolare, l'AI deve (quanto più possibile) essere etica (provando a capire cosa significa), trasparente, *trusted* (affidabile), antropocentrica, inclusiva, responsabile e neutrale (nel senso che non operi distinzioni, ad esempio di sesso o di razza o di religione). A tal fine vi sono due questioni di estrema importanza: da un lato, abbiamo il *bias* introdotto dai dati di addestramento nella decisione presa dall'approccio di AI, dall'altro l'interpretabilità del comportamento, del risultato e della predicibilità del comportamento. Per comprenderci meglio possiamo prendere come esempio il caso della corte americana che ha utilizzato un algoritmo di AI – addestrato su dati relativi a reati di una particolare area di Los Angeles – e nel fare screening si è scoperto che l'algoritmo soffriva di un "bias di razza". Oppure possiamo richiamare il caso della AI che si rivolge in maniera discriminatoria nei confronti delle donne perché "addestrata" con dati (quelli prodotti dagli uomini) affetti da bias di genere; è chiaro, quindi, che ci potremmo trovare davanti

a un complesso di decisioni che vengono prese in una certa maniera perché condizionate da dati polarizzati dal punto di vista, ad esempio, etnico o sociale. Mentre per quanto concerne la questione dell'interpretabilità fino a quando non saremo in grado di comprendere i meccanismi e le motivazioni per le quali un algoritmo di AI prende una decisione non potremo mai utilizzarlo in contesti dove il risultato (responso, verdetto) richiede una "certificazione" frutto di una interpretabilità del processo di decisione.

Molte sono le preoccupazioni nei confronti dell'AI, tra cui quelle di Elon Musk, Bill Gates, Steve Wozniak, e del compianto Stephen Hawking (in alcuni casi le posizioni arrivano a sfiorare forme di tecno-luddismo). Ma sono molti anche i tecnoentusiasti. Io non mi iscrivo a nessuna delle due categorie. Il CEO di Google Sundar Pichai – sollecitato a rispondere riguardo l'impatto dell'Intelligenza Artificiale sul lavoro e sulla paura che la società nutre nei confronti di questa tecnologia – l'ha paragonata in termini di portata e importanza a scoperte come il fuoco; ed è proprio così. La scoperta del fuoco ha cambiato la storia dell'umanità: la paura, l'ingegno, la fantasia e l'incoscienza hanno reso possibile l'utilizzo del fuoco, che da iniziale pericolo si è rivelato il più prezioso strumento per il progresso dell'umanità. Dal fuoco per illuminare, per riscaldare, per difenderci. L'evoluzione e il progresso portano all'oggi e alla gestione del fuoco con le sue regole, le prescrizioni, e le misure di sicurezza, con professionalità specializzate al controllo e alla gestione del fuoco. E, come è stato per il fuoco, così oggi dobbiamo imparare a gestire l'Intelligenza Artificiale Generale. A tal proposito, il Regolamento Europeo sull'AI dell'aprile 2021 (a mio avviso tardivo) propone, tra le varie, una classificazione delle applicazioni di AI: Vietate, Alto Rischio, Rischio Limitato, Rischio Minimo. E, per capirci, nella classe delle applicazioni Vietate troviamo:

- l'uso di sistemi di IA che distorcono il comportamento di una persona attraverso tecniche subliminali;
- l'uso di sistemi di IA che sfruttano qualsiasi vulnerabilità in modo da causare o essere suscettibili di causare danni fisici o psicologici;
- l'uso di sistemi di IA che consentono la valutazione/classificazione dell'affidabilità di persone fisiche mediante l'attribuzione di un punteggio sociale (*social score*)

L'economia predittiva abilitata dall'IA che influenza gli esseri umani cambiando il loro processo decisionale e, di conseguenza, il loro comportamento è accettabile? (seti è piaciuto questo vino potrebbe piacerti anche quest'altro, e lo si acquista). Ma l'economia predittiva diviene poi informazione predittiva e ciò permette di moltiplicare la capacità di fare propaganda (abbiamo visto cosa è successo durante il COVID-19, stiamo vedendo che cosa

sta accadendo nella discussione sulla guerra Russa/Ucraina): la novità è che si può mirare agli individui, si genera una informazione personalizzata: questa è probabilmente la più grande minaccia per la stabilità della società e della democrazia, per come la conosciamo oggi.

G.G.: Per ora, almeno, preferirei non entrare nel merito di formule ammalianti come "l'AI deve essere etica… etc.". Al di là dalle seduzioni di cui dicevo prima tra tecnofobia e tecnoentusiasmo che animano solitamente i dibattiti gazettieri, mi limito a fare una considerazione che ritengo preliminare se si ragiona nell'alveo del rapporto tra "scienza e filosofia" anche segnatamente al macro-fenomeno AI: ogni nuova tecnologia, sin dai tempi del Cannocchiale di Galileo Galilei ha sollevato e solleva dubbi, preoccupazioni, resistenze, incredulità e proiezioni d'ogni tipo non tanto per l'*ente* in sé ma, ovviamente, segnatamente alle ricadute che può avere. Sovente, invece di mettere in forma e affrontare le questioni che potrebbero/dovrebbero individuarsi, ci siamo lasciati (e ci lasciamo ancora) influenzare da una sorta di congenita diffidenza, specie per cambiamenti che potrebbero apparire troppo rapidi ma di cui, in fondo, siamo pur sempre gli attori sintagmatici, anche se tendiamo a dimenticarcene. Non v'è dubbio che, in particolar modo quelle in atto da oltre un quarto di secolo (ma in realtà potremmo pre-datare di qualche decennio, se si pensa alle tecnologie del Bios), le rivoluzioni tecniche e tecnologiche presentino risvolti complessi e controversi, prefigurando scenari prossimi e non troppo a venire di radicali modificazioni dell'antropo-sfera. Tuttavia, una cosa va anzitutto rilevata e sottolineata, altrimenti ogni discorso, già dalle sue premesse, ricade in una chiacchiera di senso comune: le rivoluzioni tecniche e tecnologiche non sono mai estranee al contesto antropico in cui si determinano e si sviluppano. Non solo ne sono il portato ma, finanche, vi interagiscono in maniera profonda, al punto da trasfigurare e rideterminare il contesto stesso e, quindi, come sempre e come da sempre, l'umano stesso che non solo non è mai stato un qualcosa di dato una volta e per tutte ma, anzi, ha rinvenuto la possibilità persistentiva più propria nella sua costante trasfigurazione. Di fatto, si può dire che le rivoluzioni/evoluzioni tecniche e tecnologiche non sono mai neutre, indifferenti e imparziali. Finanche nei termini di accessibilità. Per cui, rispetto al taglio della sua domanda, mi preoccuperei più della detenzione/disponibilità e, quindi, dell'uso potenziale, di determinate "risorse". Mi sembra, in quest'ottica, che imprenditori come Musk, Zucherberg etc. vadano ascoltati alla stessa stregua di tutti gli altri "addetti ai lavori" ricordando, però, che non sono estranei al contesto ma, anzi, per certi versi *sono* essi stessi il contesto.

S&F: ***L'AI rappresenta una manipolazione di terzo livello, vale a dire una tecnologia che riesce, in maniera autonoma, a mettere in comunicazione artefatti prodotti da sapiens, senza che quest'ultimo debba essere coinvolto, in qualche misura, nel loro funzionamento e nella loro supervisione. Ritenete che questo possa determinare una marginalizzazione di sapiens e un rifluire di quest'ultimo all'interno di una dimensione costituita da agenti artificiali che non hanno più sapiens stesso come terminus ad quem?***

A.P.: Siamo ancora lontani da scenari di questo tipo e dobbiamo anche evitare che i tecnoentusiati di turno o gli appassionati di film ambientati nel futuro tecnologico confondano la realtà con una plausibile finzione, sebbene le finzioni sembrano sempre più vicine alla realtà che stiamo vivendo. Sicuramente c'è però anche l'idea dell'umano intrappolato nella macchina e tutta la lunga narrazione sul preoccupante ruolo del progresso (spesso solo sviluppo o confuso come tale), demonizzato in funzione di una centralità dell'uomo. C'è poi il rapporto tra *homo sapiens* e *homo technologicus*, un rapporto che sembra sostanzialmente destinato a essere sempre più simbiotico. Nel 1973 un'azienda tedesca, la Kuka, entrò nel mercato della robotica con una macchina pionieristica e la chiamò Famulus. La scelta del nome è quantomai interessante perché all'epoca, l'idea dominante era di una macchina a completo servizio dell'uomo; oggi, invece, i software di AI – in particolare quelli delle applicazioni più avanzate del *Machine Learning* – richiedono una "mediazione" dell'umano, ossia della conoscenza e dell'esperienza dei *sapiens*, in base al "principio" dell'uomo nel loop, dove l'esperienza umana è centrale nei processi di apprendimento dell'AI. La guerra tra Russia e Ucraina ci ha mostrato come la tecnologia e in particolare l'intelligenza artificiale sia centrale sia nella guerra "materiale" sia in quella immateriale della "cyber guerra". E a oggi in campo militare le intelligenze artificiali prevedono che sia sempre l'uomo a prendere l'ultima decisione: non solo quindi "in the loop", ma "at the end of the loop".

G.G.: Visto che in qualche modo insiste sul punto…
Non v'è dubbio che la macro-questione etica che sottende ogni tipo di riflessione (anche e soprattutto filosofica) ruota attorno all'interrogativo: "AI così potenti potrebbero a loro volta progettare macchine ancor più intelligenti di loro stesse e di *sapiens* ragion per cui *sapiens* si troverebbe a condividere il pianeta con una nuova *specie* (artificiale) che a tal punto sarebbe dominante?".
Ora, al di là di retro-pensieri più o meno fantascientifici, la questione non è banale tant'è che numerose personalità del mondo

della scienza, della tecnologia e della ricerca a tutto tondo (Hawking, Wilczek, Bill Gates, Wozniak, lo stesso Musk come veniva ricordato prima) hanno espresso cautele e si sono fatti promotori di una lettera aperta (*Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter*) in cui, tra le altre cose, è sostenuta da un lato e con forza la ricerca sull'AI avvertendo, però e contestualmente, sulla necessità che l'AI *faccia quel che le chiediamo di fare*. Ovvero: «a causa del grande potenziale dell'IA, è importante ricercare come sfruttarne i vantaggi evitando potenziali insidie». Sostanzialmente, e come rilanciato in sede di UE, la ricerca e lo sviluppo dell'AI dev'essere comunque "umanocentrica". Posta così la questione sembra essere risolta, e in fondo la "pezza a colori" di stampo etico sembra fungere da buon palliativo. In realtà a me la questione, che tra le altre cose ruota attorno al posto dell'uomo *nel Loop*, qui e ora ma anche e soprattutto domani, mi pare si presenti come epocale, al punto da far deflagrare gli approcci etici tradizionali. E la formula "umano-centrismo" sembra risolvere tutto ma, in realtà, è assolutamente vuota.
Ripartiamo allora dal problema che ci si pone in maniera ritornante e che, appunto, il presunto fine eticamente assoluto dell'umanocentrismo non risolve in premessa, se non è ovviamente riempito di contenuti. La questione che, in quanto tale è capitale, è quella di raggiungere un accordo tra le nostre reali preferenze e l'obiettivo posto alla/nella macchina. Cioè, e in maniera stringente, il problema che ci si pone non è banalmente mezzi-fine bensì, e più radicalmente, dell'allineamento dei valori. I valori e dunque gli obiettivi affidati alla macchina devono essere allineati a quelli dell'uomo. E questa è, fondamentalmente, questione filosofica, perché in gioco vi sono valori e obiettivi di ànthropos. Presente e prossimo-venturo. Il *piccolo* problema è che rispetto a questo, ancor prima di fantasticare sulle *potenzialità* dell'*artificium*, dovremmo esser consapevoli che i valori sono storici e non assoluti e metafisici. Dov'è il punto allora… Trovare una soluzione al cospetto di questioni non lineari, ovvero autentici dilemmi – che è letteralmente un problema decisionale tra due imperativi morali, nessuno dei quali è del tutto preferibile – segnatamente alla coniugazione tra pratica e valore. La questione dell'allineamento dei valori cui accennavo. L'espressione che mi piace molto, *Human in the Loop*, fa riferimento a questo snodo qui e non tanto, e semplicisticamente, a che tipo di postazione operativa l'uomo ha e avrà nella sua relazione-coniugazione con la macchina. Rispetto a questo, approcci tradizionali – ovviamente parlo dell'etica e della filosofia morale – non reggono più. Sono, a mio modo di vedere, vetusti e assolutamente non all'altezza della sfida.

S&F: ***Solitamente le problematiche etiche dell'AI sono esaminate a partire da due modelli normativi: l'etica consequenzialista (in***

particolare in una delle versioni dell'utilitarismo) o l'etica deontologica, nella versione kantiana o nella versione pluralista, ad esempio il principialismo di Ross, Beauchamp e Childress (vale a dire più principi da usare in situazione e attraverso l'applicazione di criteri di bilanciamento). Ritenete che questi approcci siano sufficienti? O credete piuttosto che di fronte ai rapidi avanzamenti delle tecnologie digitali e dell'AI siamo costretti ad adottare forme, più o meno sofisticate, di una qualche '"etica della situazione" non sussumibile in principi pre-stabiliti?

A.P.: *The Moral Machine Survey* (pubblicato su Nature nel 2018)[4] ha mostrato – nel caso delle auto a guida autonoma – come sia praticamente impossibile pre-programmare con un approccio *Top-Down* una AI di un veicolo autonomo con una disciplina morale. Qualunque essa sia. Sia essa una disciplina di natura consequenzialista/utilitarista, fare cioè ciò che provoca la minore quantità di danni aggregati (e la maggiore quantità di benefici aggregati); sia essa una disciplina costruita intorno a una qualche forma di imperativo categorico, dove si fissa *ex ante* una regola pratica, indipendentemente dagli scopi e dalle circostanze specifiche. 40 milioni di decisioni in dieci lingue raccolte da milioni di persone in 233 paesi hanno mostrato come la morale non è un concetto universale: ad es., i Paesi orientali considerano meno accettabile di noi Europei – in un incidente causato da un'auto a guida autonoma – "sacrificare" un anziano per salvare un bambino, mentre quelli di cultura ispanica e francese sono più orientati a salvare le donne. Ne viene comunque fuori un paradosso: nella maggior parte dei casi emerge che gli utenti preferirebbero che un'auto a guida autonoma salvasse i pedoni a scapito del suo equipaggio, ma anche che non comprerebbero mai un'auto a guida autonoma programmata in questo modo. E quindi mi viene da concludere citando Megatron Transformer di Nvidia, una intelligenza artificiale di Oxford a cui è stato posto il quesito: "L'intelligenza artificiale sarà mai etica?" La risposta della "macchina", di Megatron Transformer, è stata: "l'AI è uno strumento e, come ogni strumento, viene utilizzato nel bene e nel male. Non esiste una buona intelligenza artificiale. Solo umani buoni e cattivi."

G.G.: È quello a cui accennavo in precedenza: gli approcci – diciamo così – tradizionali a cui allude anche nella domanda, sono letteralmente evaporati e perciò decisamente inutilizzabili e forse è arrivato il momento di sbarazzarcene e di liquidare una volta e per tutte le raccapriccianti pretese di ancorare la morale al naturale, facendo finanche della morale qualcosa di naturale…

[4] E. Award, S. Dsouza, R. Kim *et al.*, *The moral machine experiment*, in «Nature», 563, 7729, 2018, pp. 59–64.

Vediamo di capirci: siamo abituati – immagino che ad Antonio sia già capitato di imbattersi in filosofi morali in tal senso orientati decine e decine di volte – a tipologie d'approccio al dilemma etico di stampo *verticale* e/o *orizzontale*. Lo sperimentiamo ancora nel solco di quel sapere senza statuto epistemologico qual è la Bioetica. Ci si riempie la bocca, si monopolizzano dibattiti pubblici d'ogni tipo e sorta, riproponendo questo schema bipolare verticale-orizzontale di fatto non sortendo effetto alcuno nei termini di incidenza e incisività. Se da un lato, infatti, l'approccio verticale – fuori spazio-tempo da circa cinque secoli – si riferisce a un codice di valori metafisicamente fondato, quindi prescrittivo e indiscutibile, in cui il valore trova la sua legittimazione in un'istanza extra-fisica (Dio, mondo delle idee etc.), dall'altro, quello orizzontale, sembra presentare aspetti di maggiore efficacia ed efficienza. Non mira infatti a imporre valori eterni e si dimostra solitamente attento alle esigenze umane che tengano conto delle condizioni e delle trasformazioni storiche. In quest'ottica rientra, ad esempio, la cosiddetta etica ingegneristica, oppure l'etica *delle* nuove tecnologie e, comunque, ogni piattaforma riflessiva che accompagna un complemento di specificazione al sostantivo "etica". In questo solco, che potremmo definire a-religioso e che, indefettibilmente, è quello in cui si gioca il dibattito etico più serio e problematico come pure si accennava, Bentham e Kant sono i due *campioni* di riferimento. Ovvero come sintetizzato dal Prof. Pescapè, da un lato, il *conseguenzialismo/utilitarismo* in ragione del quale è utile ciò che ha come conseguenza la più grande felicità del maggior numero di persone e che tende a fare dell'etica persino una scienza esatta alla stregua della matematica. Dall'altro, appunto, il *de-ontologismo* in ordine al quale un atto morale è un atto che sarebbe giusto per qualsiasi tipo di persona, in circostanze simili a quelle nelle quali un soggetto si trova nel momento di eseguirlo. Se poi tutto questo è addirittura sotteso dalle formulazioni dell'imperativo categorico, e cioè dall'*operare in modo che la massima della volontà possa sempre valere in ogni tempo come principio di una legislazione universale* fino all'*agisci in modo da trattare l'umanità sia nella tua persona sia in quella di ogni altro sempre come fine e mai come mezzo*, sembra proprio che abbiamo dato un contenuto all'umano-centrismo di cui prima. Questo approccio sembra infatti mettere l'uomo al centro del mondo morale perché lo considera come noumeno e lo esorta a non contemplare mai soltanto la parte fenomenica di se stesso ma, anche, quella razionale. Solo così l'uomo può evitare di far danno a se stesso e agli altri.
Questo è l'orizzonte etico-filosofico nel quale siamo abituati a muoverci. Primo problema: nessuna filosofia, o teoria etica, è assumibile in tutte le circostanze, al massimo possiamo parlare di criteri orientativi. Secondo problema: come ho già detto, i valori sono storici (e storicizzabili). Il valore è l'attività della

valorazione che è sempre in una collocazione spazio-temporale. È una tensione a far-valori, ma il valore è ciò che qui e ora vale attraverso l'azione. Quello che oggi per noi è valore, domani potrebbe essere disvalore e, soprattutto, disgiunto dall'azione è niente più che una dichiarazione d'intenti. Terzo problema: da quanto ci siamo detti sino a ora, e come accennava nella domanda, l'etica è sempre un'etica della situazione. Qualcuno in tal senso ha parlato di *relativismo etico*… va bene. Purché non nel senso di abbandono del valore e dell'agire etico. Bensì nel senso del valore e dell'agire etico sempre e indefettibilmente relativamente-*a* quale, dunque, permanente etica situazionale.
Perciò, se relativamente alla macro-questione AI, si cerca una risposta "chiusa"… beh è sbagliata l'aspettativa riposta nella domanda. La domanda di per sé è sempre aperta, e non potrebbe essere altrimenti. La risposta "chiusa"… che è quello che si cerca in vista di una pre-programmazione dell'*artificium*, cade miseramente al cospetto dell'imponderabile.
Si è detto: è l'aspettativa riposta nella domanda a esser fuori fuoco; alla fine tutto rifluisce solo in un problema di *Manuale d'istruzioni etiche*, in un'etica da etichetta, che giocherà e danzerà sempre, in maniera del tutto inconcludente, sugli assetti variabili dei due criteri di cui prima in seno all'approccio orizzontale. Ovvero: o l'etica precede oppure corriamo ai ripari dopo nei termini del dispositivo costi-benefici. O, ancora: il computabile dispositivo costi-benifici è addirittura prodromo in maniera tale da dare all'etica una fondazione sincretistica tra *conseguenzialismo/utilitarismo* e *de-ontologismo*.
Sotto questo profilo, scienziati, ingegneri informatici, biotecnologi e filosofi ed eticisti parleranno sempre lingue estremamente distanti, avendo finalità distinte che sembreranno trovare soluzioni a un problema solo quando, per mero caso, gli interessi s'incrociassero… Possiamo permettercelo?
Direi di no. Direi proprio di no.
Per me la questione è questa e si gioca al di qua del bene e del male: oggi, nell'era dell'AI, l'uomo è sottoposto a una pressione decisiva che lo proietta verso un proprio, ulteriore (e forse definitivo) trascendimento. Come conseguenza delle rivoluzioni biotecnologiche e informatiche, si impone che l'uomo si proietti oltre se stesso, verso qualcosa d'altro. Una trasfigurazione complessiva che ingiunge di ripensare qual è la condizione umana passata, presente e futura.
Il problema della scelta, della scelta etica, se c'è, è purtroppo dopo…. Propongo per questo, in ultima battuta, un approccio circolare. La domanda sull'AI è anzitutto domanda sull'uomo. Sull'uomo che ha messo mano all'AI; sull'uomo che la utilizza e utilizzerà e, soprattutto, sull'uomo del domani.
Come s'è provato a dire sin dalle prime battute, non v'è dubbio che l'Intelligenza Artificiale sia una delle più grandi promesse dell'umanità; grazie ai suoi sviluppi, attuali e dietro l'angolo,

saremo probabilmente in grado di fare cose che oggi sarebbero impensabili. Vivremo meglio, e magari più a lungo e più felici. E tuttavia non è possibile non cogliere sino in fondo anche quelli che potrebbero essere, ma che *in nuce* già sono proprio come segnalato, i risvolti connessi a questo tipo di tecnologia che «raggiungerà e poi supererà enormemente la finezza e la duttilità di quelli che consideriamo i migliori fra i tratti umani».
Ho detto qual è la questione: parlare di AI implica parlare di tecnologia e filosofia, di macchine e di uomini, *naturale* e *artificiale* in termini del tutto nuovi.
Significa sviluppare sempre più la capacità di conoscere l'essere umano coi suoi bisogni, e sviluppare la capacità di integrarsi con l'innovazione tecnologica senza perdersi in essa. E in questo la coappartenenza in un univoco progetto di filosofi e scienziati è decisivo.
Motivo per il quale dobbiamo metter mano a un Lògos flessibile e polifunzionale e strutturare una lingua comune come base di confronto che esca sia dai gerghi del linguaggio scientifico sia da alcune complessità del linguaggio filosofico.
Questo costituisce lo scaffolding euristico per l'edificazione di un nuovo umanesimo, una nuova condizione dell'umano che riparta da una ritrovata intimità scientifico-filosofica che è ormai ineludibile e vitale proprio per l'umano a venire.
In fondo dobbiamo ripartire dalla consapevolezza che, ogni volta, le nostre storie hanno inizio da una fine. Dalla fine di un'idea di noi stessi su noi stessi che ci ha accompagnati produttivamente anche per secoli.
Nietzsche le chiamava *metafore vitali*, anche nei termini delle nostre auto-narrazioni che ci re-insediano nel nostro rapporto con il mondo esterno, anche per dare un senso. Da quando abbiamo preso a costruirle e raccontarle, però, il tema è sempre stato il desiderio di emanciparci da noi stessi per divenire qualcosa di diverso e d'altro per continuare a essere.
In quanto esseri umani, viviamo tra le rovine di uno splendore immaginato, mai raggiunto e mai raggiungibile. Abbiamo sempre avuto un'idea ben più alta del nostro destino. Siamo e continuiamo a essere, né più né meno che alla stessa maniera degli altri viventi, a partire dai connotati della nostra strategia persistentiva di specie.
Certo complessa, complessissima…
Ma questo non è che fa di noi un vivente speciale e/o che gode di uno statuto ontologico privilegiato.
Ne dice semplicemente del fatto che, in una circolarità che non conosce soluzione di continuità, nel farci in relazione con l'ente, anche l'ente artificiale creato da noi, troviamo il modo di continuare a essere e, perciò, ridirci, nei termini anche di ipotesi auto-narrative che si superano costantemente.

E alla fine, e giusto per chiudere – mi accorgo di essere andato un po' troppo oltre –, rispetto all'AI la vera domanda è: "ne vale la pena?".
A me sembra che abbiamo già risposto, e da tempo.

GIANLUCA GIANNINI insegna Filosofia Morale presso l'Università Federico II di Napoli e Coordinatore del Corso di Laurea Triennale in Filosofia
gianning@unina.it

ANTONIO PESCAPÈ insegna Sistemi Informatici presso l'Università Federico II di Napoli e Delegato del Rettore all'Innovazione e alla Terza Missione. È inoltre direttore della DIGITA Academy
antonio.pescape@unina.it

LUCA LO SAPIO è Ricercatore di Filosofia morale presso l'Università degli Studi di Torino e Docente esterno incaricato di Percezione ed Etica delle Biotecnologie Industriali presso l'Università degli Studi di Napoli Federico II
luca.losapio@unina.it