La sfida etica dell'intelligenza artificiale e il ruolo della filosofia

Gianfranco Basti

Elenco acronimi usati nell'articolo

IA	Intelligenza Artificiale
LOS	Logica di Ordine Superiore
LPO	Logica del Primo Ordine
ME	*Machine Ethics* (Eica delle Macchine)
ML	*Machine Learning* (Appredimento Automatico)
MT	Macchina di Turing
MTU	Macchina di Turing Universale
AWS	*Autonomous Weapon Systems* (Sistemi d'Arma Autonomi)
RNA	Reti Neurali Artificiali

1. *Introduzione: le sfide etiche dell'IA*

Nell'attuale *Era della Comunicazione*, gli esseri umani e le macchine interagiscono e dipendono gli uni dalle altre sempre più fortemente e inscindibilmente, come altrettanti *agenti di comunicazione "consci" e "inconsci"*[1]. Ciò apre la strada ad una discussione sempre più vasta ed articolata sulle *implicazioni etiche e legali* che l'uso di sistemi di IA in ogni campo del quotidiano della nostra vita sociale, economica e culturale comporta.

Come ricordato nell'Introduzione al saggio su *Etica dell'Intelligenza Artificiale e della Robotica* nella *Stanford Encyclopedia of Philosophy* di Vincent C. Müller – che invito a consultare per avere un quadro approfondito, ampio e aggiornato al 2021 sull'attuale dibattito – la discussione sull'etica nell'IA si può suddividere in due grandi filoni.

1 Cfr. G. Basti, *The Post-Modern Transcendental of Language in Science and Philosophy*, in Z. Delic (ed.), *Epistemology and Transformation of Knowledge in a Global Age,* InTech, London (UK) 2017, pp. 35-62.

> «Questioni etiche che sorgono con i sistemi di IA come *oggetti*, cioè strumenti realizzati e utilizzati dagli esseri umani. Questo include questioni di privacy e manipolazione, opacità e distorsioni, l'interazione uomo-robot, i problemi occupazionali e gli effetti dei sistemi di IA "autonomi". In secondo luogo, questioni che riguardano i sistemi di IA come *soggetti*, cioè l'etica per i sistemi stessi dell'IA considerati come "agenti morali artificiali" nella cosiddetta *Machine Ethics*»[2].

Ognuna di tali questioni costituisce così una sfida per la riflessione e la pratica filosofica contemporanee. In una parola, il pregio di questo lungo saggio è di fare un po' d'ordine in una letteratura sull'argomento che è letteralmente esplosa in questi ultimi anni e di cui si fa fatica a individuare delle linee-guida con cui orientarsi. Ciascuno dei temi elencati nella citazione sopra riportata e che è contenuta nell'*Introduzione* al saggio, costituisce un tema di una delle sottosezioni dell'articolo, là dove viene anche citata dall'Autore la principale letteratura al riguardo disponibile al 2021.

In questo mio contributo mi concentrerò su alcune questioni legate al secondo filone delle problematiche etiche dell'IA, quelle legate ai sistemi di IA come, *soggetti*, come *agenti morali artificiali*. Non solo perché sono le più nuove ed intriganti, ma perché strettamente legate ad un vero e proprio salto di qualità che si richiede ai programmi di studio delle nostre facoltà e dipartimenti di filosofia, di diritto e più in generale di scienze umane. Esse, infatti, troppo lentamente si stanno adeguando a questa sfida da cui dipende il presente e il futuro della professione e dell'insegnamento filosofico, giuridico e umanistico in generale, e quindi il suo posto nell'Accademia. Vi torneremo nelle Conclusioni di questo articolo.

In ogni caso, per motivi di chiarezza, spiego brevemente a cosa allude Müller e la letteratura al riguardo per ciascun gruppo delle questioni elencate, anche se le suddividerò in un diverso ordine logico.

1. Questioni di *privacy e di manipolazione dei dati.* Sono le più evidenti e facili a comprendersi. I sistemi di IA si applicano ovunque ci siano delle grandi basi di dati la cui gestione è impossibile agli umani, e che ormai con la progressiva informatizzazione di qualsiasi aspetto della vita personale, sociale ed economica, riguardano i dati sensibili e non di tutti noi. Ciò che forse sfugge ai più ed è paradossale ma vero, è che questi sistemi, profilandoci e incrociando i dati che ci riguardano ogni volta che usiamo internet o lo smartphone, accediamo

2 Cfr. V. C. Müller, *Ethics of Artificial Intelligence and Robotics*, in E.N. Zalta (ed.), «Stanford Encyclopedia of Philosophy», (June 1, 2021) https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/. Data di accesso 6 novembre 2022.

a una banca dati, richiediamo un documento online, o semplicemente usiamo un motore di ricerca su internet, ormai conoscono le nostre abitudini, attitudini e preferenze molto meglio di quanto noi conosciamo noi stessi. E che queste profilazioni siano accessibili ad altri e non a noi stessi crea un grosso problema etico-legale che dovremmo prima o poi affrontare. Soprattutto perché fin d'ora queste profilazioni sono usate sistematicamente nella creazione di *fakes* per influenzare determinati gruppi di persone, con gravi problemi sull'autonomia delle scelte non solo in campo economico-commerciale, ma innanzitutto in campo politico-sociale. Un vero problema per le democrazie occidentali in quella "guerra non-dichiarata" ma effettiva ormai da diversi anni fra regimi democratici e regimi autocratici in cui siamo tutti più o meno dolorosamente coinvolti, soprattutto dove quella guerra è dichiarata (vedi Ucraina).

2. *Problemi di opacità e distorsione nel trattamento dei dati.* Mentre i classici *sistemi esperti* nel trattamento e nella classificazione automatica dei dati legati al cosiddetto *approccio simbolico* all'IA non soffrono di questo genere dei problemi, i molto più potenti sistemi di IA che includono algoritmi di ML basati su architetture multistrato di reti neurali (il cosiddetto *deep-learning*) soffrono sistematicamente di un problema di *opacità* allo stesso programmatore nel trattamento dei dati. Nei sistemi esperti, infatti, gli alberi inferenziali per la classificazione dei dati sono definiti dal programmatore e quindi il percorso seguito dal sistema per arrivare alla decisione finale è sempre ricostruibile e perciò controllabile. Questo invece è sistematicamente impossibile nei sistemi di ML basati sulle reti neurali multistrato, che per questo motivo necessariamente enfatizzano anche preconcetti (*biases*) o più esattamente "propensioni negative" verso determinati gruppi o tipologie di individui presenti nelle basi-dati statistiche su cui il ML si esercita in maniera sistematicamente "non-trasparente" o "opaco". Tutto ciò con gravi conseguenze quando i dati, le predizioni e le decisioni prese o suggerite all'operatore umano dal sistema riguardano la vita, il lavoro, la salute, la giustizia, o il benessere economico delle persone che invece hanno diritto ad una piena "trasparenza" sulle motivazioni che hanno motivato le scelte che li riguardano. Ma su queste fondamentali limitazioni degli algoritmi di ML torneremo nella Sez. 3 dopo che nella Sez. 2 ci saremo resi conto meglio come filosofi in cosa consista l'opacità nel processo decisionale dei sistemi autonomi di IA dotati di ML.

3. *Interazione uomo-robot.* Anche se ancora non troppo evidente ai molti rispetto alla precedente, si tratta di una problematica emergente, che diventerà sempre più rilevante, quando i robot e i sistemi di IA che li controllano saranno diffusi su vasta scala. Essi sono infatti destinati a sostituire o affiancare gli umani oltre che nell'industria, nella chirurgia, nelle operazioni di sicurezza e

salvataggio ad alto rischio e sempre di più in operazioni militari dove già sono molto diffusi, anche in tantissime altre applicazioni che riguardano la vita di tutti noi. Anche i più fragili, a cominciare dall'assistenza infermieristica, domestica e in molte altre relazioni di cura delle persone finanche educative.

4. *Problemi occupazionali.* È ovvio che i sistemi di IA e di robotica, nella misura in cui sostituiscono gli umani in compiti legati non solo a lavori manuali e di fatica com'era all'inizio, ma anche legati ai servizi alle persone creano problemi occupazionali e di riqualificazione della forza lavoro su vasta scala – anche dei filosofi e dei giuristi! –, compresa la necessità di compensazioni per lavoratori non più riqualificabili. E questo ha inevitabili implicazioni etico-sociali.

5. *Autonomia decisionale dei sistemi di IA*. È certamente *il problema eticamente più delicato dei sistemi di IA* robotizzati o meno, definiti, appunto *sistemi autonomi di IA*. Sistemi, cioè, le cui decisioni/azioni con alta rilevanza etico/legale perché riguardano la vita, la salute e il benessere socioeconomico delle persone *non dipendono* direttamente dal programma implementato nella macchina, proprio perché dotati di ML. Esempi di sistema autonomo di IA sono *i veicoli a guida autonoma* sia terrestri che aerei, in particolare quelli usati come armi, oppure, più in generale i robot con applicazioni militari (*autonomous weapon systems AWS)*, ma anche robot medicali e di cura, oppure, su scala applicativa ancora più ampia, come i *sistemi di domotica* per la "casa intelligente". Infine, ciò che sfugge – ma si può intuire il perché questa problematica non sia divulgata visti gli interessi economici in gioco – esiste un campo in cui l'autonomia decisionale dei sistemi di IA si applica su vasta scala da più di un decennio. Quello delle *transazioni veloci automatizzate sui mercati finanziari* che ormai coprono oltre il 50% delle transazioni stesse sui mercati mondiali. È in particolare con questa classe di problematiche etico-legali dei sistemi autonomi che entriamo nella necessità di considerare i sistemi autonomi di IA come *soggetti* dell'azione morale come cioè *agenti morali artificiali.*

6. *La Machine Ethics.* È chiaro, afferma Müller, che con i sistemi autonomi siamo di fronte alla considerazione dei sistemi di IA come *soggetti decisionali inconsapevoli.* Cosicché, quando queste decisioni hanno una rilevanza etico-legale allora, per un semplice sillogismo, abbiamo bisogno di "un'etica per le macchine o *Machine Ethics"*.

«L'etica per le macchine si occupa di garantire che il comportamento delle macchine nei confronti degli utenti umani, e forse anche di altre macchine, sia eticamente accettabile»[3].

Più propriamente, nota ancora Müller, citando V. Dignum,

«Il ragionamento dell'IA dovrebbe essere in grado di tenere conto dei valori sociali e delle considerazioni morali ed etiche; soppesare le rispettive priorità dei valori detenuti dai diversi stakeholder in vari contesti multiculturali; spiegare il suo ragionamento; e garantire la trasparenza»[4].

In altri termini, afferma ancora Müller,

«Vi è un ampio consenso sul fatto che "il dover render conto delle" *(accountability)*, e "il dover rispondere alle" (*liability*) regole morali e legali siano requisiti fondamentali che devono essere rispettati dalle nuove tecnologie ma la questione nel caso dei robot e dei sistemi di IA in particolare quelli autonomi è come ciò possa essere fatto e come possa essere assegnata la responsabilità morale e legale»[5].

Il vero problema, ripetiamo, è che di per sé è ambiguo parlare di "*responsabilità (responsibility)* etico/legale" di un sistema autonomo di IA *inconscio*, mentre è loro richiesto che soddisfino criteri di "*rendicontabilità (accountability)* etico-legale" delle loro decisioni/azioni come accade per gli umani. Soprattutto quando queste decisioni hanno evidente rilevanza etiche e morali per la società e/o singoli gruppi o individui. È questa sostanziale questione che va risolta per poter parlare in maniera significativamente cogente di *Machine Ethics* e/o di sistemi di IA come "agenti morali artificiali".

3 M. Anderson - S. Leigh Anderson (eds.), *Machine Ethics*, Cambridge UP, New York 2011, p. 15.
4 V. Dignum, *Ethics in Artificial Intelligence: Introduction to the Special Issue*, in «Ethics and Inform. Techn.», 20,1(February 13, 2018), pp. 1-2.
5 Cfr. European Commission, Directorate-General for Research and Innovation, *Unit RTD.01,* 2018, p. 18.

2. *"Responsabilità distribuita" uomo-macchina e l'etica delle macchine*

2.1 *Il problema dello statuto ontologico di sistemi di IA come agenti morali artificiali*

In un significativo articolo del 2016 Luciano Floridi e Mariarosaria Taddeo hanno introdotto la preziosa nozione di *responsabilità distribuita* uomo-macchina riguardo la nostra problematica:

> «Gli effetti di decisioni o azioni basate sull'intelligenza artificiale sono spesso il risultato di innumerevoli interazioni tra molti attori, tra cui progettisti, sviluppatori, utenti, software e hardware. Da una siffatta *azione distribuita* (*distributed agency*) deriva una *responsabilità distribuita* (*distributed responsibility*)»[6].

Come filosofi ma anche come logici ed informatici ciò che richiede un attento approfondimento teoretico è l'attribuzione di *responsabilità* morale ad agenti o sistemi *impersonali* non dotati di *consapevolezza* (*awareness*) come le macchine e/o il loro software-hardware. D'altra parte, quando parliamo di "persone" stiamo evidenziando che non si tratta semplicemente di "individui", ma, appunto, di *persone*, ovvero di *individui-consapevoli-in-relazione comunicativa* con altri soggetti umani all'interno di una determinata comunità socioculturale e dunque linguistica. In questo senso – come appena ricordato e ho discusso altrove citando anche la bibliografia al riguardo[7] –, occorre ricordare che già da oggi e sempre più nell'immediato futuro la nostra società della comunicazione è composta di *agenti comunicativi consapevoli,* gli umani, e *agenti comunicativi inconsapevoli,* le macchine, sempre di più interconnessi e interdipendenti gli uni dagli altri. Ed è in questo quadro teoretico che va inserita la relazione di "responsabilità distribuita" fra *agenti morali umani* e *agenti morali artificiali,* oggetto di questo contributo.

Quanto detto è perfettamente congruente con l'evidenza che, etimologicamente, "responsabilità" (*responsibility*) significa "rispondere-ad-altri", "render-conto-ad-altri" (*accountability*) delle nostre decisioni/azioni rendendo esplicito il fatto che esse "soddisfano" regole etiche condivise dalla comunità di appartenenza sull'agire in senso moralmente/legalmente "buono". Per co-

6 L. Floridi - M. Taddeo, *What Is Data Ethics*, in «Phil. Trans. Roy. Soc. A: Math., Physic., Engineer. Sc.», 374,2083(2016).

7 Cfr. G. Basti, *The Post-Modern Transcendental of Language in Science and Philosophy*, cit.; Id., *Ethical Responsibility vs. Ethical Responsiveness in Conscious and Unconscious Communication Agents*, in «Proceedings», 47,68(2020), pp. 1-7.

modità nel resto dell'articolo, continuerò perciò a parlare di "responsabilità morale condivisa" uomo-macchine per rimanere nel *trend* della discussione attuale. Tuttavia, siccome nella scienza come nella filosofia rigorosamente intesa la precisione è tutto, "responsabilità" va intesa nel senso – questo sì comune ad agenti comunicativi consci e inconsci – di "rendicontabilità trasparente" al resto della società *dell'eticità/legalità delle proprie decisioni* come ciò che caratterizza un agente morale sia esso umano o artificiale. E questo malgrado in inglese *responsibility* e *accountability* siano due termini sostanzialmente sinonimi.

2.2 *L'etica delle macchine e la revisione del ruolo della filosofia nell'accademia e nella società*

In un altro articolo in corso di pubblicazione scritto con il mio amico, il professore di fisica teorica Giuseppe Vitiello[8], abbiamo cercato di definire rigorosamente lo statuto ontologico dei *sistemi autonomi di IA* come *agenti morali artificiali* che giustifica l'esistenza della nuova disciplina della ME. Essa dovrebbe essere parte dell'insegnamento di *filosofia morale e antropologia* dei nostri Dipartimenti di Filosofia proprio perché la nostra società è sempre di più destinata ad essere costituita di agenti morali umani e artificiali, strettamente interagenti. Qui mi limiterò a discutere nella Sezione Terza di questo contributo un aspetto solo, il più fondamentale, di quelli elencati e discussi nel suddetto articolo con maggiore dovizia di particolari. Quello che riguarda le ultime due condizioni fondamentali da soddisfare per attribuire *responsabilità (rendicontabilità) etico-legale ai sistemi autonomi dell'IA* in comparazione a quelle che definiscono la *responsabilità (rendicontabilità) etico-legale degli agenti morali umani.*

D'altra parte, la serie completa delle condizioni suddette e non solo delle ultime due che suppongono tutte le precedenti, costituisce un'elencazione sintetica dei contenuti di base del *programma di studi* per formare dei filosofi e dei giuristi con le competenze necessarie per interloquire *professionalmente* e *costruttivamente* – e non semplicemente in forma esortativa o peggio omiletica – con ingegneri informatici riguardo le sfide etiche dell'IA e, nello specifico della ME. Non si può infatti chiedere agli ingegneri la competenza etico-legale *professionale* che può essere solo di filosofi e giuristi per la *progettazione*,

8 G. Basti - G. Vitiello, *Deep-learning opacity and the ethical accountability of AI systems,* in R. Giovagnoli - R. Lowe (eds.), *The Logic of Social Practices,* vol. II*,* Springer, Berlin-New York 2024 (in stampa).

sviluppo e applicazione di sistemi autonomi di IA le cui decisioni/azioni soddisfino specifici e ben definiti *criteri etico-legali* nei diversi campi. Ma soprattutto soddisfino il *diritto alla trasparenza* che la società deve rivendicare sia dagli uomini che dalle macchine su come quelle decisioni/azioni *effettivamente* soddisfino quei criteri quando quelle decisioni/azioni riguardano la vita e il benessere delle persone.

Il problema è che i sistemi autonomi di IA più efficienti per il trattamento di basi di dati talmente grandi da risultare intrattabili agli umani sono proprio quelli, come già ricordato nell'Introduzione, che manifestano un *irriducibile opacità* del loro processo decisionale perché basati sul cosiddetto *deep-learning.* Cioè, su modelli di ML basati su *Reti Neurali Artificiali* (RNA)[9] *multistrato* che, come nei cervelli naturali, hanno fra lo strato dei neuroni di input (cortecce sensorie) e dei neuroni di output (cortecce motorie, comportamento linguistico incluso) molteplici *strati profondi* di neuroni che elaborano l'informazione (strati di cortecce associative)[10]. Come uscire dal vicolo cieco?

9 Matematicamente, una RNA corrisponde ad una *matrice di probabilità transitive* con cui modellizzare il meccanismo statistico (elettro-chimico) delle *sinapsi* mediante cui i neuroni sono connessi nel sistema nervoso e nel cervello. Secondo questo modello statistico, ogni cella della matrice definisce un valore della probabilità di attivazione, o di transizione "acceso-spento" 1/0, di un singolo neurone. Si tratta di probabilità *condizionali* "transitive" perché il valore definito nella singola cella dipende dai valori delle altre celle (neuroni) connesse nella matrice, e dove quindi ogni connessione ha un "peso statistico" diverso nel determinare la probabilità di attivazione (transizione 1/0) del neurone dipendente. Non si tratta dunque di probabilità *incondizionate* e quindi puramente *casuali* come nel classico caso del lancio dei dadi. I diversi modelli di reti neurali probabilistiche così definite dipendono dunque dalle diverse funzioni statistiche *lineari* o *non-lineari* di aggiornamento dei pesi statistici. Chiaramente si possono così modellizzare nelle RNA le diverse "porte logiche booleane" di un computer, ognuna corrispondente a un "connettivo logico", innanzitutto i principali ("e", "o", "se…allora", "se e solo se… allora"), del calcolo logico proposizionale espresso in forma booleana. Il paradigma computazionale di riferimento sarà così la *Macchina di Turing* (MT) *statistica,* non quella *deterministica,* esattamente come nel caso dei *computer quantistici* anche se con ben altre proprietà rispetto a calcolatori basati su dinamiche di tipo classico, non-quantistico.

10 Storicamente, il primo modello di ML basato su RNA è quello del cosiddetto "perceptrone" (cfr. F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Cornell UP., Buffalo NY 1961). Esso non aveva "strati profondi" di neuroni fra quelli di input e output e, soprattutto, la funzione di attivazione per l'aggiornamento dei pesi statistici era *lineare.* Questo significa che non era possibile implementare in questa architettura di rete la porta logica booleana dello XOR, ovvero della "disgiunzione esclusiva" (*aut, aut*) fondamentale perché la rete potesse eseguire automaticamente *compiti di classificazione* dei dati.

3. *Le due condizioni da soddisfare per attribuire responsabilità morale a uomini e macchine malgrado l'opacità dei processi decisionali in ambedue*

3.1 *Il Test di Turing come fondamento del programma di ricerca dell'IA e la distinzione fra IA simbolica e non-simbolica (reti neurali)*

Il fatto di usare un approccio comparativo fra uomini e macchine per la soluzione del nostro problema è *costitutivo* del programma di ricerca dell'IA in quanto basato sul cosiddetto *Test di Turing* o *Test dell'Imitazione* definito da Turing in un famoso articolo del 1950[11], fra la capacità dell'uomo e della macchina di eseguire *calcoli logici*, anche se non necessariamente di *logica matematica* ma anche di *logica modale* come è il nostro caso. Infatti, per implementare nella macchina *competenze etico-legali* è necessario che essa esegua anche calcoli di *logica deontica*[12]. Calcoli cioè dove cioè l'operatore di "obbligo/permesso" che caratterizza i calcoli modali deontici è una delle possibili interpretazioni *intensionali* dell'operatore di "necessità/possibilità" che caratterizza la logica modale in generale come *logica filosofica* distinta da quella matematica[13]. Nello specifico, parleremo così di sistemi di IA che

11 Cfr. A.M. Turing, *Computing machinery and intelligence*, in «Mind» 59(1950), pp. 433-460.

12 Cfr. il primo trattato moderno di logica modale assiomatizzata: C.I. Lewis - C.H. Langford, *Symbolic Logic,* Century Company, New York 1932 e la sintesi di esso nei testi ormai classici di M.J. Cresswell - G.E. Huges, *A New Introduction to Modal Logic*, Routledge, London 1996 e, in italiano, di S. Galvan, *Logiche intensionali. Sistemi proposizionali di logica modale, deontica, epistemica*, Franco Angeli, Milano 1991. Di fatto, essa consiste nell'aggiunta di opportuni *assiomi modali* all'ordinario calcolo logico proposizionale, che regolino l'uso degli operatori di *necessità* □ e *possibilità* ◊ nei calcoli modali, nelle loro varie interpretazioni *intensionali*. Storicamente è bene ricordare che la logica modale era stata "tagliata via" dall'orizzonte culturale e accademico moderno a partire dal XV secolo, per un uso miope del "rasoio di Occam" (sebbene risalente ad Occam stesso e quindi al XIII sec. con la sua negazione della *distinzione reale* (modale) "essenza-esistenza"), conseguente all'abbandono del pensiero scolastico. Nella Scolastica, invece, la logica modale si era sviluppata almeno fino al XIV secolo, come in particolare l'opera di Tommaso d'Aquino e soprattutto di Duns Scoto attestano, nelle quali – pur con ben note differenze dottrinali – tanta parte ha l'analisi delle diverse modalità della necessitazione in logica, fisica, metafisica e teologia. La logica modale era infatti la logica del *sillogismo modale* aristotelico che, come ha dimostrato il grande logico polacco J. Łukasiewicz (J. Łukasiewicz, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*. *Second Enlarged Edition*, Oxford UP, Oxford 1957) non è formalizzabile nell'interpretazione *estensionale* del sillogismo che è quella dominante nella modernità da Leibniz in poi (cfr. nota 13).

13 Infatti, gli operatori modali possono essere interpretati in senso *aletico* (distinguendo fra necessità/possibilità, "logica", "fisica" e "metafisica"), in senso *epistemico* (distinguendo fra necessità/possibilità della credenza fondata o "scientifica" e non-fondata o "opinione"), in senso *deontico* (distinguendo fra necessità/possibilità etica dell'"obbligo" e del "permesso" in senso morale o legale) Cfr. S. Galvan, *Logiche intensionali,* cit., e G. Basti - F. Panizzoli, *Istituzioni di Filosofia Formale*, LUP-Lateran University Press, Roma 2018. Ora, ciò che distingue la *logica intensionale* (con la "s") del *pensiero intenzionale* (con la "z") della filosofia e delle discipline umanistiche in quanto inscindibili dagli autori umani singoli o collettivi e dai loro contesti storico-culturali che le esprimono,

eseguono calcoli di *logica booleana modale*[14], sebbene resti aperta in questo caso la questione teoretica se calcoli di questo tipo possano ancora avere nella Macchina di Turing Universale (MTU), modello astratto di computer programmabile, il loro paradigma di riferimento, senza che per questo venga meno il Test dell'Imitazione come fondamento dell'IA[15]. Esso è infatti all'origine del programma di ricerca dell'IA, fin dalle sue origini, databili all'estate del 1956. Infatti, nel famoso workshop tenuto presso l'Università di Dartmouth fra i maggiori esperti del campo, da Marvin Minsky, a Claude Shannon, al futuro Premio Nobel per l'economia Herbert Simon, si è delineato con incredibile lungimiranza il programma di ricerca dell'IA[16].

Esso si è così fin da subito sviluppato secondo le due direttrici fonda-

dalla *logica estensionale* della matematica pura ed applicata, è che nella prima non vale l'*assioma di estensionalità.* Ovvero, "se due classi sono equivalenti (hanno la medesima estensione) allora sono identiche": (e i loro predicati (p.es., "essere acqua" e "essere H_2O") possono essere reciprocamente sostituiti senza cambiare il senso della proposizione, che invece caratterizza la logica matematica. P.es., quando in aritmetica scrivo "5 = 3+2" di fatto sto affermando l'equivalenza fra l'insieme di 5 elementi e l'unione (somma) dell'insieme di 3 e quello di 2 elelementi (cfr. E. Zalta, *Intensional Logic and the Metaphysics of Intentionality*, MIT Press, Cambridge MA 1998). Infatti, se, per esempio nel linguaggio religioso sostituissi "acqua" con "H_2O" come in "Signore, benedici questa H_2O", la proposizione diventerebbe immediatamente insensata, tenuto conto che l'acqua come simbolo religioso ha diverse interpretazioni nelle varie comunità linguistiche religiose. In questo senso si dice che il linguaggio dei resoconti soggettivi degli stati intenzionali della mente è un "linguaggio in prima persona" singolare/plurale di individui e di gruppi (Cfr. W. Sellars, *Intentionality and the Mental*, in H. Feigl - M. Scriven - G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Mind. Concepts, Theories and the Mind-Body Problem*, vol. II, Minnesota UP, Minneapolis 1958, pp. 507-524). Dal che si intuisce il carattere necessariamente *relazionale* della logica modale e delle sue interpretazioni (semantiche) intensionali ("aletiche", "epistemiche", "deontiche"), senza perdere di *universalità logica* in quanto formalizzabile e formalizzata, e dunque anche *algoritmicamente implementabile* in sistemi di IA, p.es., nella cosiddetta "ingegneria della conoscenza" (*knowledge engineering*) per lo sviluppo di database semantici per i vari gruppi/usi linguistici, ma anche, come nel nostro caso, nella ME.

14 Infatti, l'iniziale approccio assiomatico di Lewis alla logica modale è stato poi sviluppato nella *logica modale relazionale* di S. Kripke, e quindi *algebrizzato* in forma di *logica modale booleana,* basata sull'*algebra delle relazioni* in topologia*.* In tal modo si fornisce una versione direttamente algoritmizzabile di logica modale (Cfr. R.I. Goldblatt, *Mathematical modal logic: a view of its evolution*, in «Journal of Applied Logic», 1(2003), pp. 309-392; e V. Goranko - M. Otto, *Model theory of modal logic*, in P. Blackburn - F. J. van Benthem - F. Wolter (eds.), *Handbook of Modal Logic*, Elsevier, Amsterdam 2006, pp. 252-331).

15 Su questa base, Walter Freeman, John Searle, Hubert Dreyfus, tutti e tre professori e ricercatori all'Università della California a Berkeley, hanno fortemente criticato la teoria che un IA basata sul paradigma della MT possa essere un modello delle operazioni logiche eseguite da una mente intenzionale. In modo ingenuo ma evocativo, Searle con la sua famosa metafora della "stanza cinese", in contrapposizione alla "stanza del Test di Turing", ha affermato che una MT non è un modello delle computazioni del cervello umano perché essi, in quanto base neurale di una mente intenzionale, sono computi di una logica intensionale e non estensionale come quelli di una MT (Cfr. J.R. Searle, *Minds, brains, and programs*, in «The Behav. and Brain Sc.», 3(1983), pp. 128-135).

16 Cfr. J. McCarthy - M. Minsky - N. Rochester - C. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, in «AI Magazine», 27/4(August 1, 1955). http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

mentali che ne definiscono tuttora il campo: a) i sistemi dell'*IA simbolica* o "sistemi esperti" perché simulano nell'albero inferenziale esplicito di un programma, da cui dipendono le decisioni del sistema, l'abilità di un esperto umano; e b) i sistemi dell'*IA pre-simbolica* basato sulle RNA e quindi dotati di un algoritmo di ML, "supervisionato" o "non-supervisionato"[17] da cui dipendono le decisioni del sistema. Infatti, la vastità delle basi di dati (*big-data*) su cui questi sistemi si applicano esclude la loro trattabilità da parte di un qualsiasi agente umano, rendendo questi sistemi *indispensabili* alla nostra società. A questa classe appartengono anche i cosiddetti *sistemi autonomi,* destinati a sostituire gli umani in tantissime funzioni.

I modelli di ML più recenti ed efficienti di ML sono quelli delle cosiddette "reti neurali convolutive"[18]. Essi sono basati su un'architettura di RNA multistrato costituita da "neuroni" la cui "funzione di attivazione" della risposta 1/0 del neurone è di tipo *statistico* e *non-lineare* proprio come nei neuroni

17 I più diffusi sono gli algoritmi "supervisionati" di ML, così definiti perché le classi in cui suddividere automaticamente i dati sono già conosciute. Sostanzialmente, sono algoritmi di ottimizzazione. Durante la *fase di addestramento* (*training*) del modello di ML, si presenta al sistema un campione dei dati di cui già si conosce la classificazione per studiare come il modello "impara" automaticamente a classificarli. Si calcola poi con un'opportuna funzione di errore la distanza fra la classificazione ottenuta dal modello e quella desiderata. Quindi, questa misura viene "retro-propagata" (*back-propagation*) sui neuroni dei diversi strati della rete per l'aggiornamento automatico dei pesi statistici delle connessioni fra neuroni. Il processo viene reiterato fino ad ottenere il risultato ottimale (= minimizzazione dell'errore). In una seconda *fase di prova* (*testing*) si ripete il processo per un altro campione del database diverso dal precedente, ma di cui si conosce sempre la classificazione corretta, per provare la *capacità di generalizzazione* del modello appreso ad altri dati da classificare. Se il risultato è positivo, il sistema si applica al resto del database non ancora classificato. Gli algoritmi di ML "non-supervisionati" sono invece quelli in cui è la rete stessa a definire le classi mediante tecniche statistiche di *raggruppamento delle variabili significative* per definire ciascuna classe. Molto spesso i modelli "non-supervisionati" sono usati come tecnica di inizializzazione dei pesi dell'algortitmo supervisionato per incrementare la velocità e l'accuratezza dell'apprendimento.

18 Il primo modello di rete neurale convolutiva fu pubblicato nel 2012 (A. Krizhevsky - I. Sutskever - G.E. Hinton, *ImageNet classification with deep convolutional neural networks*, in *NIPS'12*, Proceedings of the 25th International Conference on Neural Information Processing Systems, vol. 1, ACM Publ., 2012, pp. 1097-1105), e divenne immediatamente famoso perché applicato con successo al riconoscimento/classificazione secondo un migliaio di classi diverse, di una base-dati costituita da oltre un milione di volti, con miliardi di parametri, tratti dal database di pubblico dominio *ImageNet* gestito dal Laboratorio di *computer vision* dell'Università di Stanford, concepito per la ricerca di algoritmi efficienti di ML nel settore specifico del riconoscimento di immagini (cfr. https://www.image-net.org), Esso oggi ha raggiunto oltre quattro milioni di immagini perché continuamente alimentato dai sistemi di riconoscimento facciale sparsi ovunque nel mondo. Le reti convolutive sono attualmente il modello di ML supervisionato più diffuso ed efficiente per compiti di classificazione su basi di dati particolarmente ampie, non solo di immagini. Cfr. l'articolo: A. Ghosh - A. Sufian - F. Sultana - A. Chakrabarti - D. Debashis, *Fundamental Concepts of Convolutional Neural Network*, in V.E. Balas - R. Kumar - R. Srivastava (eds.), *Recent Trends and Advances in Artificial Intelligence and Internet of Things*. *Intelligent Systems Reference Library,* vol. 172*,* Springer, Berlin-New York 2020, pp. 519-567, come la sintesi introduttiva più aggiornata e completa di questo modello di ML.

naturali[19]. Ciò li rende capaci in linea di principio di implementare la porta logica booleana della "disgiunzione esclusiva" (lo XOR, *aut…aut*) essenziale per eseguire *operazioni di classificazione* fra i diversi oggetti. In questo senso l'architettura delle RNA convolutive simula l'architettura multistrato delle cortecce associative del cervello umano, dove ogni strato corrisponde matematicamente a un "filtro" mediante cui i neuroni di ciascun strato estraggono progressivamente le caratteristiche comuni dell'input dalle cortecce sensorie. Ogni strato corrisponde così a un livello di "astrazione" o di indipendenza dalle caratteristiche spazio-temporali del singolo input e quindi rendono capaci la rete di operare veri e propri processi di "generalizzazione statistica" su basi di dati talmente vaste che le rendono intrattabili a qualsiasi operatore umano e dove dunque l'utilizzo di questi sistemi di IA risulta indispensabile nella nostra società.

Il rovescio della medaglia è che sono proprio questi modelli di ML che manifestano il problema di un'*irriducibile opacità* del loro processo decisionale *proprio come accade nei processi decisionali umani*. Se capiamo perciò come noi umani possiamo definirci "agenti morali responsabili" malgrado l'opacità irriducibile dei nostri processi decisionali (epistemologicamente, la "formulazione di un giudizio", "speculativo" o "pratico" che sia), possiamo imitare artificialmente questa procedura anche nei sistemi di IA per rendere consistente l'attribuzione ad essi della connotazione di "agenti morali artificiali" malgrado l'opacità dei loro processi decisionali.

19 *Matematicamente*, una RNA corrisponde ad una "matrice di probabilità transitive" con cui si vuole modellizzare su un computer digitale il meccanismo probabilistico di attivazione dei neuroni di una rete, basato neurofisiologicamente sul meccanismo elettro-chimico e dunque probabilistico e non deterministico delle *connessioni sinaptiche* fra i neuroni. In questa modellizzazione, ogni cella della matrice definisce un valore della probabilità di attivazione, o transizione 1/0, di un singolo neurone. Si tratta quinci di probabilità *condizionali* "transitive" perché il valore definito nella singola cella dipende dai valori delle altre celle (neuroni) connesse nella matrice, e dove quindi ogni connessione (sinapsi) ha un "peso statistico" diverso nel determinare la probabilità di attivazione del neurone dipendente. I diversi modelli di reti neurali probabilistiche così definite dipendono dunque dalle diverse funzioni statistiche non-lineari di aggiornamento dei pesi statistici. Di fatto, l'output di una rete neurale consiste nel definire una distribuzione di probabilità delle variabili e delle loro correlazioni dei dati che corrisponda al meglio alla distribuzione di probabilità e alle correlazioni effettivamente presenti nei dati di input. Dove il fatto che si tratti di variabili correlate significa che si tratta di probabilità "condizionate" e non "incondizionate" e dunque non puramente *casuali* come nel caso paradigmatico del lancio dei dadi.

3.2 *Il contributo della neuroetica alla soluzione del problema: mente estesa e dimensione interpersonale della responsabilità morale*

Il progetto di ricerca della *neuroetica* nell'ambito delle neuroscienze cognitive è legato storicamente alla critica del neurofisiologo portoghese Antonio Damasio al dualismo cartesiano mente-corpo, con la conseguente *separazione assoluta di razionalità ed emotività* nelle decisioni morali. Tale critica è basata sulle evidenze neuropsicologiche da lui raccolte a partire dagli anni '90, poi confermate da molteplici altre ricerche. Tali evidenze dimostrano che, per un difetto *nell'area che controlla le emozioni* del loro cervello (l'amigdala e l'ippocampo), alcuni soggetti con quoziente intellettivo medio-alto o addirittura alto, che non mostravano disturbi né dell'attenzione, né del linguaggio, né tantomeno del ragionamento astratto, mostravano però una sistematica difficoltà *a prendere decisioni altruistiche e moralmente corrette* per la loro *vita di relazioni interpersonali* che li rendesse eticamente bene accetti alle loro comunità di appartenenza[20].

Tuttavia, il contributo fondamentale della neuroetica alla soluzione del nostro problema della responsabilità morale malgrado l'irriducibile opacità dei processi decisionali nell'uomo e nella macchina viene da un altro, molto noto, rappresentante della neuroetica, Neil Levy, professore all'Università di Oxford dove è Fondatore e Direttore del *Centre for Neuroethics*, ora denominato, *The Oxford Uehiro Centre for Practical Ethics* che ha nella *Neuroethics* ma anche nell'*AI and Digital Ethics* due dei suoi principali e strettamente interconnessi campi di ricerca[21]. In uno dei suoi primi libri sulla neuroetica, conosciuto e tradotto in tutto il mondo, Levy critica non solo l'idea che *l'autocoscienza* (l'"io cartesiano-kantiano") sia il soggetto delle decisioni cognitive e morali, ma anche la falsa idea che nel cervello esista un "controllore", sede della coscienza e localizzato in qualche parte del cervello (di solito nelle cortecce dei lobi prefrontali)[22]. Infatti, l'agente dei processi decisionali è l'intero insieme dell'individuo umano, il cervello in relazione col corpo e attraverso di esso con l'ambiente fisico e sociale ovvero la *persona,* in quanto "individuo-in-relazio-

20 Cfr. A. Damasio, *Descartes'Error: Emotion, Reason, and the Human Brain,* Putnam Publishing, New York 1994. Come ho discusso nel Cap. V del mio manuale di antropologia, G. Basti, *Filosofia dell'uomo,* Edizioni Studio Domenicano, Bologna 2008, pp. 258ss., questa riscoperta del ruolo dell'emotività umana (la facoltà della *cogitativa*, nella terminologia tommasiana) nel giudizio morale è del tutto congruente con il carattere intenzionale dell'atto morale proprio della Scolastica contro il razionalismo etico cartesiano-kantiano che tanti disastri ha prodotto nella modernità.

21 Cfr. Il portale web del Centro all'indirizzo: https://www.practicalethics.ox.ac.uk/

22 Cfr. N. Levy, *Neuroethics: Challenges for the 21st Century,* Cambridge UP, Cambridge (UK) 2007, spec., pos. 494-498 (Kindle Edition).

ne", dove la componente cosciente del processo decisionale è solo la "punta dell'iceberg", senza che per questo l'atto intenzionale cognitivo o deliberativo sia meno "nostro" in quanto *persone* e non "autocoscienze" ipostatizzate [23].

Per risolvere perciò l'enigma della "responsabilità condivisa" uomo-macchina, diventa essenziale l'estensione ai sistemi di IA della distinzione neuroetica tra la *rapida e incosciente responsività*[24] dei nostri cervelli ai vincoli ambientali e *la lenta responsabilità cosciente* delle nostre menti verso i medesimi vincoli. A tal fine, è utile riportare qui il seguente esempio usato da N. Levy in un altro libro completamente dedicato al rapporto fra coscienza e responsabilità morale. L'esempio riguarda la questione della responsabilità morale di un pilota umano molto ben "addestrato" a guidare come il famoso pilota automobilistico Ayrton Senna.

> «È caratteristico dei processi coscienti che sono molto più lenti di quelli inconsci[25]. La rapida responsività di agenti altamente addestrati come (...) Senna

23 Infatti, Levy è sostenitore della cosiddetta ipotesi della *mente estesa* che colloca la mente non in qualche parte del cervello ma *nelle relazioni di scambio di informazione* fra cervello, corpo e ambiente, ipotesi alla quale dedica un lungo capitolo (Cfr. N. Levy, *Neuroethics,* cit., pos. 494-922). D'altra parte, l'ipotesi della mente estesa è congruente con l'altra neurofisiologica di Damasio che, coerentemente con quanto affermato da Brentano, fonda la *aboutness* della relazione intenzionale, non primariamente nel rapporto cosciente "soggetto-oggetto", ma nella stessa biologia e in particolare nella nozione di *omeostasi* basata sul fatto che l'organismo *vivente*, dalla cellula all'uomo, è un sistema "aperto" che scambia continuamente energia e informazione con l'ambiente (A. Damasio, *The strange order of things. Life, feeling and the making of cultures*, Pantheon Books, New York 2018). Il cervello che non scambia più energia e informazione col suo ambiente interno ed esterno è il cervello del cadavere! D'altra parte, come mostrato al Cap. 6 del mio testo di antropologia (G. Basti, *Filosofia dell'uomo*, cit., pp. 348ss.), la tesi della "mente estesa" è del tutto congruente con la "localizzazione" tommasiana dell'anima come "contenente il corpo" (e non platonicamente come "contenuta" nel cervello), e in genere la localizzazione di tutte le sostanze spirituali (angeli e Dio) come "contenenti" le realtà materiali su cui effettuano la loro funzione di "controllo" (*gubernare et regere*).

24 "Responsività" traduce qui l'inglese *responsiveness* che potrebbe essere tradotta più correntemente con "reattività". Biologicamente, però, con *responsiveness* si intende la capacità di risposta "auto-adattiva" della cellula (e del vivente in generale) alle variazioni dell'ambiente mediante complessi circuiti di autoregolazione *non-lineari* che la caratterizza come "vivente", non riducibili al principio di azione-reazione della meccanica newtoniana. Infatti, come N. Wiener ci ha insegnato, i viventi e le macchine sono sistemi a *controllo attivo,* caratterizzati da circuiti più o meno complessi di "retroazione" (*feedback*) e "auto-regolazione" (*self-regulation*), e quindi capaci di inviare *segnali comunicativi* (energia + informazione) fra i loro sottosistemi. In questo sono distinti dai sistemi meccanici lineari o a *controllo passivo* (basato cioè esclusivamente sul principio di azione-reazione) della meccanica classica. Cfr. N. Wiener, *Cybernetics. Or communication and control in the animals and the machines,* MIT Press, Cambridge (MA) 1962[2].

25 Il riferimento qui è ad un'altra fondamentale evidenza neurofisiologica che anima il dibattito neuroetico soprattutto negli Stati Uniti. Fin dai primi lavori sperimentali di B. Libet (B. Libet et al., *Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act*, in «Brain», 106,3(1983), pp. 623-642) è ben noto che i "potenziali di azione" che precedono l'invio del segnale nei neuroni coinvolti in un atto intenzionale, precedono di alcuni millesimi se non decimi di secondo la componente cosciente dell'atto intenziona-

deve certamente essere guidata da questi ultimi e non dai primi. *Sembra quindi falso che gli agenti debbano essere consapevoli delle informazioni a cui rispondono per essere responsabili di come rispondono ad esse*. (...) La responsabilità morale diretta richiede che un agente cosciente sia consapevole della *rilevanza morale (moral significance)* delle proprie azioni»[26].

Questa distinzione, però, può anche far luce sulla questione dei sistemi autonomi di AI dotati di ML, intesi non come oggetti ma come *soggetti inconsci di decisioni eticamente/legalmente rilevanti*, o più precisamente come *agenti morali artificiali altamente addestrati*, attraverso tecniche di ML. La velocità della loro "responsività" all'ambiente è infatti molto più alta di quella di un cervello umano addestrato![27] Si pensi – per rimanere all'esempio appena citato da Levy – agli *autoveicoli a guida autonoma*, che cominciano a popolare il traffico delle nostre città (a Pechino e a San Francisco sta già succedendo) ma destinati a diventare nel giro di pochi anni una parte significativa della nostra società della comunicazione basata sull'interazione uomo-macchina. Infatti, proprio perché caratterizzati da una velocità di risposta alla strada e a chi la popola (veicoli o pedoni) molto più alta di un guidatore umano, esse garantiscono molta più sicurezza e faranno scendere significativamente il numero degli incidenti stradali.

In altre parole, nel suo libro Levy sottolinea che nella formulazione umana sia di *giudizi cognitivi* che di *giudizi morali* – formalmente, *valutazioni/decisioni logiche "vero/falso"(1/0)*, rispettivamente in contesti modali aletici e deontici (cfr. nota 13) – ciò che è "trasparente", cioè cosciente per noi e infine "trasparente" anche agli altri nella misura in cui linguisticamente comunichiamo "in prima persona" i nostri stati di coscienza, è ciò che è *prima* e *dopo* la formulazione del giudizio/decisione stessi. Il processo di formazione del giudizio è invece *assolutamente inconscio* e quindi *opaco per chiunque*, proprio come accade ai sistemi autonomi di IA dotati di *deep-learning*.

---

le stesso. Ciò rende possibile *di fatto* "leggere le intenzioni nel cervello" degli agenti umani (J. Haynes - G. Roth - M. Stadler - H. Heinze, *Reading intentions in the human brain*, in «Current Biology», 17/4(2007), pp. 323-328). Di qui la falsa conclusione – che sarebbe vera se fosse l'autocoscienza cartesiana e non la persona il soggetto dell'atto libero – dell'illusorietà del libero arbitrio umano, come per esempio si afferma nel libro che ha avuto un'ampia diffusione negli USA di D.M. Wegner, *The Illusion of Conscious Will,* MIT Press, Cambridge (MA) 2002.

26 Cfr. N. Levy, *Consciousness and Moral Responsibility*, Oxford UP, Oxford 2014, pos. 114. 121 (Kindle Edition).

27 Banalmente, ma significativamente, la capacità di risposta di un neurone è dell'ordine dei decimi di secondo (75 msec.), quindi è un oscillatore che lavora a 10Hz (dieci oscillazioni al secondo). I chip delle CPU dei nostri computer lavorano a centinaia di MHz (milioni di oscillazioni al secondo), sono quindi sette ordini di grandezza più veloci di un neurone biologico.

In questo senso, possiamo dire che i sistemi autonomi di IA dotati di ML con la loro intrinseca opacità sono "i vincitori" del "gioco dell'imitazione" del Test di Turing, su cui si basa il programma di ricerca sull'IA fin dalle sue origini nel 1956, molto più dei *sistemi simbolici di IA*, il cui processo decisionale è generalmente "trasparente" proprio perché basato su un albero inferenziale di decisioni definito dal programmatore.

Infatti, quando produciamo un giudizio morale su un'azione e/o una scelta che dobbiamo eseguire (= *giudizio pratico*), ciò che è "cosciente", "trasparente" a noi, ed eventualmente agli altri è *solo ciò che precede e segue* il giudizio/decisione morale in quanto tale.

In effetti, possiamo distinguere *tre passi* in ogni giudizio/decisione morale umani[28]:

1. *In un primo momento*, esaminiamo consapevolmente *le diverse componenti dell'azione/scelta* che valuteremo con un giudizio morale su di essa. Cioè, consideriamo una molteplicità di fattori, principalmente le situazioni simili passate, la situazione concreta attuale, le future conseguenze pratiche della nostra azione/scelta e, naturalmente, anche le *norme morali astratte* che dovrebbero governare la nostra azione (vincoli etico/legali).

2. *In seguito*, combinando attraverso *un processo inconscio* queste e altre componenti non considerate al primo passo (prima di tutte le *emozioni*, come ci ha insegnato la neuroetica), produciamo il nostro *giudizio/decisione morale* (= "giudizio pratico") sull'azione giusta da compiere "qui e ora". Formalmente, produciamo la nostra *valutazione di logica deontica* (= giudizio morale) sull'azione/scelta che vogliamo eseguire, che, formalmente, sarà una valutazione di logica deontica del primo ordine, avendo per argomento variabili *individuali* – la singola azione da eseguire/non-eseguire.

Tuttavia, l'*essere davvero responsabili* verso noi stessi e gli altri *della rilevanza morale* delle nostre azioni richiede che, prima di eseguire la nostra azione/scelta che abbiamo deciso (giudicato) di eseguire, ci "riflettiamo sopra", ovvero,

3. *Come terzo passo*, facciamo riflessivamente un *ragionamento morale,* una sorta di "*auditing* morale a noi stessi" sul *giudizio morale* appena formulato. Formalmente, si tratterà di una valutazione deontica di *logica di ordine superiore* (LOS) sulla decisione deontica di *logica del primo ordine* (LPO) che abbiamo formulato nel giudizio sull'azione giusta da effettuare. Questo al fine di valutare/giustificare a noi stessi prima di tutto e eventualmente anche agli

28 Cfr. per maggiori dettagli il Capitolo V del mio manuale di antropologia, G. Basti, *Filosofia dell'uomo,* cit., spec. pp. 265ss., dedicato alla struttura e ai momenti costitutivi dell'atto morale umano.

altri, se effettivamente questo giudizio/decisione soddisfa tutti i vincoli etico/legali che avevamo posto prima di effettuare il giudizio – e eventualmente altri vincoli/regole morali che non avevamo considerato.

È dunque soprattutto attraverso *il ragionamento pratico "trasparente", susseguente al giudizio pratico* che in sé è completamente "opaco" (inconscio) che la nostra decisione morale diventa "trasparente" a noi stessi e agli altri, e quindi il nostro atto morale soddisfa davvero le condizioni di "responsabilità" (*responsibility*) e "rendicontabilità" (*accountability*) etico/legali soggettive e intersoggettive. Si pensi, per esempio, in campo legale alla distinzione fra il "pronunciamento della sentenza" da parte del giudice e la susseguente pubblicazione del cosiddetto "apparato della sentenza", dove le motivazioni/giustificazioni della sentenza sono spiegate dal giudice attraverso un ragionamento opportuno. Solo con questo terzo passo, insomma, la nostra decisione e conseguente azione diventa *un atto di responsabilità morale pieno* perché "trasparente" in quanto capace di giustificare/spiegare anche agli altri e non solo a noi stessi la moralità/legalità delle nostre decisioni/scelte. Più in generale, il ragionamento pratico che precede l'esecuzione dell'azione, e che ha per oggetto il giudizio/decisione morale sull'azione giusta da eseguire, è "una valutazione della nostra valutazione", cioè, formalmente, consiste in una valutazione metalogica del nostro processo decisionale deontico di LPO sull'azione da eseguire, che richiede necessariamente l'uso di una LOS[29].

3.3 *La doppia condizione da soddisfare per un'attribuzione consistente dell'attributo ontologico di agenti morali artificiali ai sistemi autonomi dell'IA*

Alla luce di quanto detto, la soluzione del problema fondamentale in ME di come attribuire in modo consistente la capacità di agire moralmente ad agenti artificiali e in particolare a sistemi autonomi di IA dotati di ML, malgrado l'ineludibile "opacità" dei loro processi decisionali, richiede che si soddisfino due condizioni:

1. *L'implementazione "trasparente" negli algoritmi di ML supervisionati di vincoli etico/legali*. Cioè implementare algoritmi di minimizzazione degli errori nel processo di addestramento che soddisfino anche *condizioni etico-legali*. In questo senso, abbiamo visto che il cosiddetto approccio di "etica

29 Cfr. C. Benzmüller - X. Parenta - L. van der Torre, *Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support*, in «Artificial Intelligence», 287 (May 24, 2020), pp. 1-50.

dei valori" alla logica deontica[30] rispetto a quello della cosiddetta "etica delle virtù"[31] sembra essere il più adatto per essere implementato direttamente negli algoritmi ML. Infatti, in entrambi i casi, dell'apprendimento automatico e dell'obbligo deontico basato sui valori, c'è comunque una funzione di errore che deve essere minimizzata, o che è la stessa cosa, un *valore desiderato* (matematico o etico che sia) da dover ottimizzare. Ad esempio, nel caso del ML supervisionato per sistemi autonomi di IA usati per il *trading* automatico nei mercati finanziari già ricordati, un algoritmo di ML "eticamente soddisfacente" o "moralmente buono" per il trading significa che esso non si basa solo sulla massimizzazione del profitto, ma anche sulla soddisfazione di determinate clausole etiche (ad esempio, investimenti non derivanti da origini illecite, non basati sullo sfruttamento dei lavoratori, etc.). In ogni caso, almeno a livello di ricerca, diversi modelli di ML "etico" cominciano ad essere studiati e già esiste una diffusa letteratura al riguardo, sempre crescente vista l'attualità e la rilevanza della problematica[32].

2. *L'implementazione in sistemi di IA di un meta-sistema di valutazione etico-legale automatico delle decisioni* per verificare/spiegare in modo trasparente se le decisioni prese dal sistema soddisfano effettivamente i criteri etici stabiliti nell'algoritmo ML e/o, nel caso di sistemi di IA simbolica, i criteri etici implementati nell'albero decisionale del programma. Solo recentemente i ricercatori in IA hanno iniziato a studiare questa componente fondamentale della ME, che formalmente, come nel caso della formalizzazione del ragionamento pratico umano, richiedono algoritmi di logica deontica di ordine superiore al primo. Infatti, il loro scopo è quello di fornire una valutazione metalogica delle decisioni prodotte da un sistema di IA dotato di "competenza etica" o nei termini di clausole etico/legali da soddisfare nel programma ML e/o, nei casi di sistemi dell'IA simbolica, direttamente negli algoritmi deontici che compongono l'albero inferenziale di decisione di sistemi dell'IA simbolica.

Per aiutare i non addetti ai lavori a comprendere, sono già ampiamente diffusi e usati sistemi di IA che usano algoritmi di logiche di ordini superiore al primo per *un meta-controllo sulla consistenza* di dimostrazioni matematiche

30 Cfr. L. Floridi et al., *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, in «Minds and Machines», 28/4(2018), pp. 689-707.

31 Cfr. S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford UP, Oxford 2017.

32 Cfr. per esempio, T. Kehrenberg - Z. Chen - N. Quadrianto, *Tuning Fairness in Balancing Target Labels,* in «Front. Artif. Intell.», 3,33(2020), pp. 1-12, come uno dei più recenti modelli di ML per implementare addirittura *criteri di equità sociale,* e S. Lo Piano, *Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward*, in «Humanit. Soc. Sci. Commun.», 7,9(2020) per una review su vari modelli di ML "etico" proposti.

o logiche molto complesse effettuate da operatori umani, o anche dai cosiddetti "dimostratori automatici di teoremi", sempre nell'ambito della ricerca logico-matematica. Molto più diffusi ancora, tanto da costituire una branca a sé dell'informatica, la cosiddetta *programmazione funzionale*, sono i sistemi di IA per il meta-controllo della *affidabilità e sicurezza* dei programmi di altri sistemi automatici il cui malfunzionamento produrrebbe effetti catastrofici e che quindi non possono essere testati nel loro campo applicativo come qualunque altro software. Si pensi, per esempio, ai sistemi che controllano automaticamente la navigazione area, l'atterraggio degli aerei in tutti i grandi aeroporti, gli scambi ferroviari delle grandi stazioni, le centrali nucleari, le grandi reti di distribuzione dell'energia, le grandi reti di telecomunicazione, etc. Come si vede sono sistemi automatici di controllo e meta-controllo molto poco conosciuti dal grande pubblico, ma da cui già dipende e dipenderà sempre più la vita e la sicurezza di tutti noi.

Non è casuale, dunque, che uno dei maggiori esperti mondiali nello sviluppo e nell'applicazione di dimostratori automatici in logica e matematica, Christoph Benzmüller[33], si sia dedicato in questi ultimi anni col suo gruppo allo sviluppo di uno dei primi "ragionatori etici" (*ethical reasoners*)[34]. Esso è finalizzato alla meta-valutazione e dunque alla *giustificazione/spiegazione* etico/legale automatica usando logiche deontiche di ordine superiore, delle decisioni prese da un sistema di IA (simbolica o no) dotato della sua "competenza etica", prima che le decisioni del sistema stesso si trasformino in azioni sull'ambiente esterno. Siccome il sistema di Benzmüller e dei suoi colleghi è pensato per integrarsi con qualsiasi altro sistema di IA come un'ulteriore (meta)livello del suo processo decisionale, è chiaro che sistemi di questo tipo sono in grado di risolvere in forma sistematica il problema dell'ineludibile opacità dei processi decisionali dei sistemi autonomi dell'IA non-simbolica in ME.

Usando le stesse parole degli autori dell'articolo in cui presentano il loro ragionatore etico, «la motivazione principale è infatti lo sviluppo di strumenti adeguati *per il controllo e la gestione (governance) di sistemi autonomi intelligenti*»[35]. È chiaro che siffatti sistemi di IA che implementano logiche deontiche di ordine superiore al primo, sono tutti sistemi dell'IA *simbolica*. Non

33 Cfr. il suo famoso dimostratore automatico per l'analisi di consistenza di teoremi LEO-II: C. Benzmüller - N. Sultana - L.C. Paulson - F. Theiss, *The higher-order prover LEO-II*, in «J. Aut. Reas.», 55/4(2015), pp. 389-404.

34 Si noti la precisione della denominazione: se un sistema autonomo di IA (simbolico o non-simbolico) dotato di competenze (*skills*) etiche è un "decisore etico", il meta-sistema di IA di "valutazione delle valutazioni" del decisore è un "ragionatore etico".

35 C. Benzmüller - X. Parenta - L. van der Torre, *Designing normative theories*, cit., p. 1.

solo quindi essi sono assolutamente "trasparenti" al controllo umano (non ha senso un ML per sistemi LOS), ma come nel caso del "ragionatore etico" di Benzmüller sono in grado di "giustificare/spiegare" le valutazioni che essi producono sulla congruità delle decisioni "opache" prese da un modello di ML con i criteri etico/legali implementati nell'algoritmo di ML stesso, ed eventualmente con altri criteri utili alla (meta-)valutazione della congruità etica delle loro decisioni. In una parola, un sistema di IA integrato col "ragionatore etico" sono sistemi che in linea di principio sono in grado di soddisfare il "diritto alla trasparenza" che la società e tutti noi pretendiamo da agenti morali, umani o artificiali che siano, quando le loro decisioni, anche se prese in maniera assolutamente e ineludibilmente "opaca", impattano in maniera significativa la nostra vita.

4. *Conclusione: la sfida educativa dell'etica delle macchine per la filosofia*

In questo contributo, abbiamo esaminato brevemente alcune sfide etiche dell'IA e in particolare ci siamo concentrati su quella dell'ineludibile "opacità" di sistemi di IA dotati di ML (*deep-learning*) nei loro processi decisionali che contraddice il fondamentale "diritto alla trasparenza" che la società ha verso decisori, umani o artificiali, quando le loro decisioni – malgrado l'opacità ineludibile del processo decisionale tanto nell'uomo come nella macchina – hanno conseguenze di rilevanza etico-legale. Ovvero conseguenze che condizionano significativamente le vite delle persone o di interi gruppi sociali nei più diversi ambiti della nostra società delle comunicazioni.

Abbiamo visto come questa esigenza di trasparenza etico-legale viene soddisfatta da sempre per gli umani, così da giustificare per essi l'attribuzione dello statuto ontologico di *agenti morali responsabili coscienti* in etica. E come può essere soddisfatta in modo del tutto consistente col Test dell'Imitazione di Turing, anche per le macchine, in particolare i sistemi autonomi di IA, così da poter giustificare per essi l'attribuzione dello statuto ontologico di *agenti morali responsabili artificiali* in etica delle macchine.

In conclusione, si può dire che, se generalmente il futuro professionale dei filosofi (e dei giuristi[36]) è strettamente legato all'acquisizione di competenze

36 È nota la crisi occupazionale che l'utilizzo di sistemi di IA negli studi legali ha prodotto e sempre più produrrà per gli avvocati, innanzitutto in compiti di ricerca e documentazione per la preparazione e l'analisi delle cause e dei processi legali. Solo avvocati con competenze informatiche sopravviveranno, visto anche l'alto costo economico che l'adattamento di questi software alle specifiche esigenze di uno studio legale ha solo utilizzando società esterne di consulenza informatica. E questo proprio per l'incapacità cronica di comunicare efficacemente fra avvocati e informatici, perché educati a modi di

informatiche, ciò è particolarmente vero per tutte le problematiche etiche legate all'IA. Esse richiedono una collaborazione professionale interdisciplinare fra informatici, filosofi e giuristi, dai livelli teoretici più alti a quelli applicativi di più diretto significato (e impatto) pratico sulla vita delle persone e delle società.

Ciò richiede una profonda *revisione dei programmi di studio* delle nostre facoltà e dipartimenti di filosofia e di legge per inserirvi, non solo corsi di logica formale assiomatizzata (matematica, modale e algebrica) adeguati alla sfida, ma anche corsi di informatica di base (intelligenza artificiale inclusa). Essi devono consentire allo studente il passaggio da determinate teorie filosofiche (ontiche, epistemiche, deontiche) formalizzate, alla possibile implementazione algoritmica dei processi inferenziali connessi – il cosiddetto "pensare a schemi" o *computational thinking*. Il suo insegnamento è raccomandato, per esempio, dalla Comunità Europea nelle scuole di qualsiasi ordine e grado, ma è sistematicamente carente proprio nelle facoltà di filosofia, di diritto e di scienze umane in generale. Da questa integrazione sistematica fra filosofia, diritto e informatica, viceversa, dipende in larga parte non solo il futuro professionale di tanti filosofi e giuristi, e, in larga parte, la sopravvivenza delle facoltà e dipartimenti di filosofia e ormai anche di legge, tutte alle prese con un drammatico calo degli studenti. Ma da questa integrazione soprattutto dipende in larga parte la preservazione dei valori umanistici all'interno della nostra società della crescente interdipendenza uomo-macchina.

ragionare del tutto opposti, ma che sono invece facilmente conciliabili.

Aquinas, LXV (2022), 299-321

La sfida etica dell'intelligenza artificiale e il ruolo della filosofia

Gianfranco Basti

Abstract

In questo articolo, discuto la questione etica fondamentale riguardante l'inevitabile "opacità" delle decisioni dei sistemi autonomi di IA dotati di processi di apprendimento automatico (*deep-learning*). Ciò è in palese contrasto con il fondamentale "diritto alla trasparenza" che la società ha nei confronti sia dell'uomo che delle macchine quando le loro decisioni hanno una rilevanza etico/legale, perché riguardano la vita e il benessere delle persone. Parto dall'evidenza che negli esseri umani la "trasparenza" e, quindi, la "rendicontabilità etico/legale" delle nostre azioni come agenti morali responsabili, non è in contraddizione con l'inevitabile "opacità" (inconsapevolezza) dei processi cognitivi attraverso i quali sviluppiamo il nostro giudizio morale (=decisione deontica) sull'azione giusta da eseguire. In effetti, la responsabilità morale delle nostre decisioni/azioni dipende da ciò che è *prima* e *dopo* il nostro "giudizio morale". Vale a dire, dipende dai "vincoli etici" che imponiamo esplicitamente al nostro giudizio morale [formalmente, una decisione deontica di logica del primo ordine (LPO)] prima di eseguirlo. E soprattutto, dipende dalla "meta-valutazione etica" o dal "ragionamento/riflessione morale" [formalmente, una valutazione deontica di logica di ordine superiore (LOS) della decisione di LPO] per giustificare a noi e agli altri la moralità del nostro giudizio. Nel resto dell'articolo discuto, alla luce della ricerca più avanzata, di come è oggi possibile implementare nei sistemi autonomi di IA un simile doppio processo di valutazione deontica di LPO e di LOS delle decisioni. Nelle conclusioni, sottolineo come tutto questo è una palese esemplificazione dell'urgente necessità di un aggiornamento dei programmi di studio dei nostri dipartimenti di Filosofia e Diritto per dare agli studenti competenze effettive in logica e informatica per sviluppare, in collaborazione con gli informatici, sistemi di IA eticamente soddisfacenti.

Parole chiave:
intelligenza artificiale; etica delle macchine; apprendimento automatico; apprendimento profondo; logica deontica.

Abstract
In this paper, I address mainly the ethical fundamental issue concerning the unavoidable "opacity" of AI autonomous systems decisions, when they are endowed with machine learning (deep-learning) processes. This is in patent contrast with the fundamental "right to transparency" that the society has toward both humans and machines when their decisions have an ethical/legal relevance, because concerning the life and the welfare of persons. I start from the evidence that in humans, the "transparency" and then the "ethical/legal accountability" of our actions as responsible moral agents, is not in contradiction with the unavoidable "opacity" (unawareness) of the cognitive processes

by which we perform our moral judgement (=deontic decision) about the right action to execute. In fact, the moral accountability of our decisions/actions depends on what is *before* and *after* our "moral judgement". Namely, it depends on the "ethical constraints" we impose explicitly to our moral judgement [formally, a deontic first order logic (FOL) decision] *before* performing it. And overall, it depends on the "ethical meta-evaluation" or explicit "moral reasoning/reflection" [formally, a deontic higher order logic (HOL) assessment of the FOL decision] for justifying to us and others the morality of our judgement. In the rest of the paper I discuss, in the light of the edge research in this field, how it is today possible to implement in AI autonomous systems a similar double FOL and HOL deontic evaluation process. Finally, I emphasize that the precedent discussion is a patent exemplification of the urgent necessity of an updating of the study programs of our Philosophy and Law departments for giving our students effective competences in logic and computer science for developing, with computer scientists, ethically accountable AI systems.

Keywords:
artificial intelligence; machine ethics; machine learning; deep learning; deontic logic.