Gianfranco Basti

ETICA DEL MACHINE LEARNING

PARTE III:
ETICA E MACHINE LEARNING

Roma 2025

SOMMARIO

Schema del Corso

PARTE PRIMA: LE ORIGINI DELL'IA E DEL MACHINE LEARNING

1. Dalla Logica all'Informatica: il Calcolatore Universale
2. Il Test di Turing e il Programma di Ricerca dell'IA Simbolica
3. La Nascita delle Scienze e Neuroscienze Cognitive
4. L'IA Pre-Simbolica e le Reti Neurali
    a. IA Discriminativa e IA Generativa
    b. Reti Neurali Naturali
    c. Reti Neurali Artificiali
    d. Il Machine Learning

PARTE SECONDA: ALGORITIMI DI MACHINE LEARNING.
IA DISCRIMINATIVA E IA GENERATIVA

1. Machine Learning Supervisionato, Con Rinforzo, Non-Supervisionato,
2. Algoritmi di Machine Learning nell'IA Discriminativa: il Deep-Learning
    a. Il Perceptrone Multistrato

b. Machine Learning Supervisionato: la Back-Propagation
c. Machine Learning Supervisionato: Reti Convolutive
d. Machine Learning Non-Supervisionato
e. Machine Learning Quantistico

3. IA Generativa: la Rivoluzione dei Transformers nel Machine Learning

PARTE TERZA: ETICA E MACHINE LEARNING

1. Etica nell'IA ed Etica nel Machine Learning
2. Le "Ingiustizie Algoritmiche" nel Machine Learning
3. Problemi Etici dell'IA Generativa
4. Neuroetica nelle Neuroscienze Cognitive ed Etica nel Machine Learning
5. Logica Deontica e la Nozione di "Algoritmo Buono" nel Machine Learning

CONCLUSIONI

- Responsabilità etica condivisa uomo-macchina nei sistemi di IA autonomi e non-autonomi

7. ETICA NELL'IA E NEL MACHINE LEARNING

7.1. Le sfide etiche dell'IA

- Nell'attuale **Era della Comunicazione**", gli esseri umani e le macchine interagiscono e dipendono l'uno dall'altro sempre più fortemente e inscindibilmente, come altrettanti **agenti di comunicazione "consci" e "inconsci"** (Basti, 2017).
- Ciò apre la strada ad una discussione sempre più vasta ed articolata sulle **implicazioni etiche e legali** che l'uso sempre più diffuso di sistemi di IA in ogni campo del quotidiano della nostra vita sociale, economica e culturale comporta.
- Come sintetizzato molto bene nel saggio su *Etica dell'Intelligenza Artificiale e della Robotica* nella *Stanford Encyclopedia of Philosophy* di Vincent C. Müller (Müller, 2025) citato nella Bibliografia Generale – che invito a consultare per avere un quadro approfondito, ampio e aggiornato al 2025 sull'attuale dibattito – la discussione sull'etica nell'IA si può suddividere in due grandi filoni:

1. I sistemi di IA intesi come **oggetti,** cioè strumenti utilizzati dagli umani;
2. I sistemi di IA intesi come **soggetti,** cioè come agenti morali aritificiali.

- Citando direttamente da questo articolo, la discussione sull' etica nell'IA riguardano attualmente:

«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***» (Parentesi mia).

- 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 2025 e che quindi non cito qui.
- In ogni caso, spiego brevemente a cosa allude l'Autore e **la vasta letteratura citata al riguardo per ciascun gruppo delle questioni elencate**, anche se le suddividerò in un diverso ordine logico.
- Vediamo innanzitutto i problemi etici e legali che riguardano i sistemi di IA come **oggetti,** ovvero come strumenti usati dagli esseri umani (singoli, compagnie, istituzioni pubbliche e private, governi, …).

1. **Questioni di *privacy e di manipolazione dei dati.*** Sono le più evidenti e facile 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 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, facciamo un acquisto online, accediamo 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 come singoli e come governi.
- 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**.
- Una democrazia rappresentativa come sono le nostre oggi **non funziona più** se le scelte sono sistematicamente condizionate in maniera subdola ma reale. Nascondere la testa sotto la sabbia come stiamo facendo **non risolve il problema ma lo acuisce.**

- Un vero problema per **le democrazie occidentali** in quella **"guerra non-dichiarata" ma effettiva ormai da diversi anni fra regimi democratici e monocratici** in cui siamo tutti più o meno dolorosamente coinvolti.

2. **Problemi di opacità e distorsione nel trattamento dei dati.** Come abbiamo studiato nella Prima e nella Seconda Parte, 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 necessariamente enfatizzano preconcetti o più esattamente **"propensioni negative" verso determinati gruppi o tipologie di individui presenti nelle basi-dati statistiche** su cui si effettua l'allenamento del

sistema (*training*) – generalmente le minoranze – su cui il ML si esercita in maniera sistematicamente "non-trasparente" o "opaco".

- Infatti, come sappiamo, **l'aggiornamento dei pesi statistici delle variabili** mediante **retro-propagazione dell'errore** nelle RNR avviene sempre in maniera **cieca,** senza tener conto della **rilevanza etica** della singola variabile coinvolta, visto che il sistema senza le opportune precauzioni, **minimizza l'errore globale** rispetto alla distribuzione statistica da riconoscere/corrispondere in input, enfatizzandone anche le **intrinseche distorsioni**.
- Gli stessi problemi di opacità **aggravati dalla complessità della struttura** li hanno i modelli **multi-head di attenzione** e le **reti multistrato FFN** dei Transformer nell'IA generativa.
- Tutto ciò con gravi conseguenze quando i dati, le predizioni e le decisioni prese o suggerite all'operatore umano dal sistema di IA **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 e le loro possibili soluzioni** che riguardano i sistemi di IA come **soggetti,** ovvero come **agenti morali artificiali** torneremo nelle prossime sezioni perché riguardano la ***Machine Ethics*** **(ME)**.

3. **Interazione uomo-robot.** Anche se ancora non troppo evidente ai molti rispetto ai problemi precedenti, si tratta di una **problematica etica-legale emergente**, che diventerà sempre più rilevante, quando i robot e i sistemi di IA che li controllano saranno diffusi su **vastissima scala**.
   - I robot – inclusi i veicoli aerei e terrestri **a guida autonoma** – sono infatti **destinati a affiancare o sostituire gli umani** oltre che nell'**industria**, nelle **comunicazioni** (p.es., call-center automatici), nella **chirurgia** (robot chirurgici), nelle **operazioni di salvataggio** ad alto rischio e sempre di più in **operazioni militari** (droni armati, *robot-soldiers,* artiglieria robotizzata, etc.), tutti campi specifici dove già sono molto diffusi, anche in tantissime **altre applicazioni che riguardano la vita di tutti noi** (si pensi alle auto a guida autonoma), **anche i più fragili.** A cominciare dall'**assistenza infermieristica, domestica** e in molte altre

relazioni di cura delle persone, finanche **educative** (sistemi di insegnamento a distanza "intelligenti" in grado di adattarsi al singolo alunno**)**.

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, creano **problemi occupazionali** e di **riqualificazione della forza lavoro** su vasta scala, compresa la necessità di **compensazioni per lavoratori non più riqualificabili**. E questo ha inevitabili implicazioni etico-sociali-politiche su cui i governi devono intervenire.

♦ Con tutto ciò ci siamo introducendo nel secondo gruppo di problematiche che riguardano l'etica nell'IA, quello che riguarda i sistemi di IA come **soggetti o agenti morali artificiali** oggetto della ME.

5. **Autonomia decisionale dei sistemi di IA**. È certamente il problema eticamente più delicato dei sistemi di IA robotizzati o meno, soprattutto in quei sistemi come i veicoli a guida autonoma sia terrestri che aerei, in particolare quelli usati come armi, o come i robot con applicazioni militari (*autonomous weapon systems AWS)*, medicali e di cura, o infine su scala applicativa ancora più ampia come i sistemi di domotica.

- Infine, ciò che sfugge all'opinione pubblica – ma si può intuire il perché questa problematica non sia divulgata **visti gli interessi economici in gioco** – esiste un campo in cui l'autonomia dei sistemi di IA si applica su vasta scala, almeno dal 2008 in poi, all'indomani della grande crisi dei mercati finanziari su scala mondiale.
- Quello delle **transazioni automatizzate sui mercati** che ormai coprono oltre il **50% delle transazioni stesse** sui mercati mondiali azionari, finanziari e delle materie prime. Gli algoritmi di ML che si usano in queste applicazioni dell'IA seguono infatti una **logica di massimizzazione dei profitti** senza usare quasi mai **alcun vincolo etico.**

6. **La *Machine Ethics.*** È chiaro, afferma Müller, che con questa classe di problematiche siamo di fronte alla considerazione dei sistemi di IA come **soggetti** in base al semplice sillogismo che "se le macchine agiscono in modi eticamente rilevanti, allora abbiamo bisogno di un'etica per le macchine o ***Machine Ethics (ME)"***.

♦ In generale,

«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.» (Anderson & Leigh Anderson, 2007, p. 15)

♦ Più propriamente, nota ancora Müller, citando Virginia Dignum

«Il "ragionamento" (processo decisionale, NdR) dei sistemi autonomi di 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 soggetti compartecipanti (*stakeholders*) in vari contesti multiculturali; spiegare il processo decisionale; e garantire la trasparenza (Dignum, 2018, pp. 1-2)».

♦ In altri termini, afferma ancora Müller,

«Vi è un ampio consenso sul fatto che "il dover tener conto delle" *(accountability)*, e "il dover rispondere alle" (*liability*) regole morali e legali siano requisiti fondamentali che devono essere rispettati dalle nuove tecnologie (cfr. (European Commission, Directorate-General for Research and Innovation, Unit RTD.01, 2018, p. 18)) 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».

- Il vero problema è che **ha ben poco senso definire "responsabile"** nello stesso senso di un soggetto umano consapevole **un sistema di AI e/o un robot** benché le loro azioni e decisioni abbiano evidenti conseguenze etiche e morali.
- È questa sostanziale ambiguità che va risolta per poter parlare in maniera significativamente cogente di *Machine Ethics* e/o di sistemi di IA come "agenti morali artificiali" soprattutto se siamo di fronte ad un "autonomia decisionale" di questi sistemi.

7.2. Le "Ingiustizie Algoritmiche" nel Machine Learning

- Prima però di affrontare questa fondamentale questione concentriamoci sul concetto di **ingiustizia algoritmica** nei sistemi di IA basati sul ML.
- Fra i vari contributi pubblicati in questi ultimi anni riguardo una **IA eticamente responsabile** particolare rilevanza nell'ambito degli informatici ha avuto l'articolo provocante fin dal titolo *Algorithmic Injustice: Towards a Relational Ethics*

presentato nel 2019 al *Black in AI workshop* durante il più prestigioso dei Congressi Annuali sulle RNA *il NIPS 2019 - Neural Information Processing System Conference* (Birhane & Cummins, 2019), ampliato poi in un articolo pubblicato nel 2021 (Birhane, 2021). **Ingiustizie algoritmiche** non intenzionali ma reali.

- Evitare questi problemi richiede, infatti, secondo gli Autori del contributo "sviluppare e dispiegare sistemi algoritmici etici", nel contesto di un **approccio relazionale all'etica**. In effetti, affermano gli autori, quando

«I sistemi di apprendimento automatico che deducono e prevedono il comportamento e l'azione individuale, sulla base di estrapolazioni statistiche superficiali, vengono implementati nel mondo sociale, **sorgono vari problemi non voluti ma reali**. Questi sistemi **enfatizzano e perpetuano stereotipi sociali e storici piuttosto che profonde spiegazioni causali di questi stereotipi**. Nel processo di ML, individui e gruppi, spesso ai margini della società che non riescono a rientrare in “scatole stereotipate”, **subiscono conseguenze indesiderabili**. Vari risultati lo illustrano: distorsione nel rilevare il colore della pelle nei pedoni; pregiudizi nei sistemi di polizia predittiva sulla criminalità; pregiudizi di genere e discriminazioni nelle inserzioni economiche per carriere di tipo STEM; pregiudizi razziali negli algoritmi di recidiva

criminale; pregiudizi nei motori di ricerca; pregiudizi e discriminazioni in medicina; e pregiudizi nelle assunzioni, per menzionarne solo alcuni» (Birhane & Cummins, 2019, p. 1).

- Ciò significa che gli algoritmi di apprendimento automatico dell'IA, quando applicati a supporti automatizzati per i processi decisionali in diverse sfere sociali, politiche ed economiche, **non sono affatto *value-free* o *a-morali***.
- Ora, sebbene gli Autori in questi loro contributi non pretendono di offrire soluzioni quanto piuttosto di enfatizzare un problema, non è casuale che la necessità di **un approccio statistico dinamico e non statico ad un'etica relazionale** è ciò che caratterizza anche la teoria rivoluzionaria di Amartya Sen della **giustizia distributiva comparativa** basato su un **principio di equità (*fairness*)** nell'ambito delle teorie sociali ed economiche, per la quale è stato insignito del premio Nobel nel 1988, e che Sen ha formalizzato nel quadro della cosiddetta **Teoria delle Scelte Sociali (*Social Choice Theory*)** che egli stesso ha contribuito a creare (Sen, 2017).
- Infatti, ciò di cui abbiamo bisogno perché possano essere sviluppati algoritmi etici sia nell'IA simbolica che, soprattutto, nel ML, basati su un principio di equità che

superi le discriminazioni fra persone e gruppi in un determinato contesto sociale è proprio di una **teoria formalizzata delle scelte sociali** come vedremo in §7.4.

- Ora, ciò che gli Autori intendono per **approccio etico relazionale** necessario per sviluppare un'IA eticamente responsabile si basa sull'evidenza che:

«**Né le persone, né l'ambiente, sono statici**; ciò che la società ritiene giusto ed etico **cambia nel tempo**. Il concetto di correttezza e comportamento etico è, quindi, un obbiettivo che cambia e **non qualcosa che può avere una risposta definitiva o può essere "risolto" una volta per tutte**. È possibile che ciò che è considerato etico attualmente e all'interno di determinati domini per determinate società **non sarà valutato allo stesso modo in un momento diverso**, in un altro dominio o per una società diversa».

- Ora, come abbiamo discusso altrove (Basti, Capolupo, & Vitiello, 2020), la sfida computazionale fondamentale che impedisce l'uso estensivo della teoria dell'equità di ispirazione **personalistica e quindi relazionale** quale quella di Sen nella

modellistica sociale ed economica è esattamente la stessa affrontata nel documento NISP sopra citato.

- **Vale a dire la necessità di una *ponderazione statistica*** di tipo ***dinamico*** **delle variabili in gioco sulle modificazioni dei contesti,** necessaria sia nell'etica relazionale, come nel caso del data streaming (Rutten, 2000; Basti & Vitiello, 2022).
- Come abbiamo visto, i **modelli statistici di attenzione multi-head** dei Transformer possono fornire una prima risposta al problema della necessaria molteplice contestualizzazione anche **in campo etico**.
- Tuttavia occorre qui delineare quali sono **le principali sfide etico-legali** che pongono i siatemi di IA generativa basata sui Transformer, perché oggetto attualmente di un **accesso dibattito** data la novità e la potenza di queste architetture di IA generativa in gran parte ancora da scoprire e da siluppare.

7.3. Le sfide etico-legali dell'IA generativa

- L'IA generativa **affronta grandi sfide etiche come bias, disinformazione e rischi per la privacy, e sfide legali che includono violazioni del copyright, vuoti di responsabilità e incertezza normativa.**

♦ Questi problemi sono profondamente interconnessi, rendendo la governance complessa e urgente.

7.3.1. Principali Sfide Etiche

- **Bias e Discriminazione**
  - I modelli di IA spesso riproducono i bias presenti nei dati di addestramento, portando a risultati ingiusti o discriminatori.
- **Disinformazione e Deepfake**
  - L'IA generativa può produrre contenuti falsi altamente convincenti (immagini, video, testi), minando fiducia, verità e valori democratici.
- **Privacy e Protezione dei Dati**
  - Dati personali sensibili possono essere utilizzati nell'addestramento senza consenso, sollevando preoccupazioni su sorveglianza e uso improprio.
- **Trasparenza e Spiegabilità**

- La natura “black box” dei grandi modelli rende difficile capire come vengano generati i risultati, riducendo l’accountability.

- **Disuguaglianze Sociali**
  - L’accesso diseguale agli strumenti di IA e il loro uso improprio possono ampliare i divari esistenti nella società.

7.3.2. Principali Sfide Legali

- **Violazione del Copyright**
  - L’addestramento dei modelli su opere protette senza permesso o compenso è una disputa legale centrale.
- **Ambiguità sulla Proprietà Intellettuale**
  - Gli attuali quadri normativi (fair use, eccezioni per text/data mining) sono inadeguati per affrontare l’uso di materiale protetto da parte dell’IA.
- **Uso Giudiziario e Responsabilità**

- Se usata nei tribunali o nei processi decisionali legali, l'IA rischia di minare la discrezionalità giudiziaria e lo stato di diritto se i risultati sono distorti o fuorvianti.

- **Frammentazione Normativa**
  - Diverse giurisdizioni (UE, USA, Giappone, Brasile) applicano standard differenti, creando incertezza per lo sviluppo globale dell'IA.
- **Vuoti di Responsabilità**
  - Stabilire chi sia legalmente responsabile per danni causati da output dell'IA (sviluppatori, utilizzatori, o utenti finali) rimane irrisolto.

◦

7.3.3. Tabella Comparativa

♦ TIPI DI SFIDA	♦ QUESTIONI SPECIFICHE	♦ RISCHI/IMPLICAZIONI
♦ **ETICHE**	♦ Bias e discriminazione	♦ Rafforza stereotipi, trattamenti ingiusti
	♦ Disinformazione e deep-fake	♦ Minaccia la democrazia, erode la fiducia
	♦ Privacy e uso improprio dei dati	♦ Viola diritti, abilita sorveglianza
	♦ Mancanza di trasparenza	♦ Indebolisce l'accountability
	♦ Disuguaglianza sociale	♦ Amplia il divario digitale

♦ **LEGALI**	♦ Violazione del copyright	♦ Cause legali, freno alla creatività
	♦ Ambiguità Proprietà Intell.	♦ Protezioni incerte per i creatori
	♦ Uso giudiziario improprio	♦ Minaccia lo stato di diritto
	♦ Frammentazione normativa	♦ Incertezza sulla conformità
	♦ Vuoti di responsabilità	♦ Mancanza di accountability chiara

7.3.4. Rischi Chiave e Prossimi Passi

- **Rischio:** La disinformazione generata dall'IA potrebbe destabilizzare elezioni o l'equità giudiziaria.

- **Rischio:** Le dispute sul copyright possono rallentare l'innovazione e portare a costosi contenziosi.
- **Azioni Necessarie:**
  - Stabilire **obblighi di trasparenza** sui dataset di addestramento.
  - Sviluppare **schemi etici di compensazione** per i titolari dei diritti.
  - Creare **meccanismi di audit robusti** per garantire equità e responsabilità.
  - Promuovere **quadri di governance multidisciplinari** che bilancino innovazione e diritti umani.

♦ In sintesi, **le sfide etiche riguardano equità, verità e diritti umani, mentre le sfide legali si concentrano su proprietà intellettuale, responsabilità e regolamentazione.** Insieme, richiedono una governance urgente e coordinata per garantire che l'IA generativa si sviluppi in modo responsabile.

7.4. Dalla Neuroetica nelle Neuroscienze Cognitive all'Etica nel Machine Learning (*Machine Ethics*)

- Uno dei contributi fondamentali alle **neuroscienze cognitive** per il superamento del primo approccio funzionalista è legato allo sviluppo della cosiddetta **Neuroetica**, originariamente ispirata dalla critica di Antonio Damasio contro l'interpretazione razionalista della mente autocosciente di Cartesio e Kant (Damasio, 1994).

7.4.1. La nascita della Neuroetica come critica al razionalismo etico di Cartesio e Kant

- La critica neurofisiologica di Damasio al dualismo cartesiano **di mente e corpo,** e quindi **di intelligenza ed emozione** è originata dalle prove sperimentali da lui raccolte negli anni '90 secondo cui, per un difetto **nell'area *emozionale* del loro cervello**, alcuni pazienti intelligenti, senza difetti di attenzione, di linguaggio e di ragionamento astratto, **manifestano un deficit sistematico** nel prendere decisioni che siano profittevoli per loro stessi non come individui isolati, ma come **persone o individui-in-relazione**, e quindi accettabili per la loro comunità, e **dunque *moralmente corretti*** (Damasio, 1994).

- Questa evidenza è di fatto una **falsificazione sperimentale del fondamento *razionalistico della morale*** di Cartesio e soprattutto di Kant. Quest'ultimo infatti criticava il tradizionale **fondamento *intenzionale* della norma morale** – perché basato su un **argomento ipotetico** (condizionale) "se vuoi (lo scopo) *X*, allora devi fare (l'azione) *Y*" –, per rivendicare un ***fondamento formalistico*** della norma morale.
- Basato, cioè, su **un argomento *apodittico* (incondizionato)**, ovvero **sull'universalità astratta del dovere morale** indipendente dai contesti.
- Si tratta della famosa nozione kantiana del cosiddetto **imperativo categorico** (del *sollen* come distinto dal *mussen,* in tedesco) del "dovere perfetto", **come incondizionato (tautologico) *dovere per dovere***, il "devi farlo, perché devi", che caratterizza per Kant **l'autonomia del giudizio morale e quindi il *libero arbitrio*** umano intesa come **volontà auto-legislatrice**, indipendente da qualsiasi **autorità eteronoma,** che costituisce il fondamento della dottrina kantiana della ***Ragion Pura Pratica*** in quanto distinta dalla ***Ragion Pura Teorica*** del sapere scientifico**.**
- Kant come Cartesio erano infatti affetti da un pregiudizio che tutt'ora affligge molti filosofi e cioè che **universalità sia sinonimo di necessità assoluta o incondizionata e quindi ultimamente di tautologicità.**

- Infatti, un fondamento intenzionale della norma morale si traduce per loro in un **imperativo ipotetico** che costringe all'azione in determinate circostanze, perché la lega al **perseguimento intenzionale di un certo scopo o *bene*** voluto o desiderato. Proprio come nel caso del soddisfacimento di un bisogno fisico: "se voglio/desidero dissetarmi allora devo necessariamente (sono obbligato a) bere".
- Di qui la formulazione dell'imperativo categorico che deve per Kant guidare la ragion pratica dell'uomo secondo le quattro definizioni datene nella *Fondazione di una metafisica dei costumi* (1785):

1. «*Agisci secondo quella massima che al tempo stesso puoi volere che divenga una legge universale*» (Kant, 1979, p. 79).
2. *«Agisci in modo tale che la massima della tua azione possa diventare una legge universale della natura»* (Kant, 1979, p. 79).
3. *«Agisci in modo da trattare l'umanità, sia nella tua persona sia in quella di ogni altro, sempre anche come fine e mai semplicemente come mezzo»* (Kant, 1979, p. 88).

4. *«La volontà non è semplicemente sottoposta alla legge, ma lo è in modo da dover essere considerata auto-legislatrice e solo a questo patto sottostà alla legge»* (Kant, 1979, p. 91).

- È chiaro come Kant, attraverso la distinzione fra **Ragion Pura Teorica** e **Ragion Pura Pratica** intendeva soddisfare la cosiddetta *Legge di Hume* che ha riproposto alla filosofia moderna la distinzione di per sé ben nota alla **logica modale** e ben chiara alla filosofia medievale, fra **verità aletiche** e **verità deontiche** andata perduta nella modernità per l'abbandono moderno dal XV secolo in poi della logica modale.
- Ma come l'analogia che Kant stesso pone nella seconda definizione dell'imperativo categorico fra **leggi matematiche della natura e leggi deontiche della morale,** egli intende perseguire il suo scopo fondativo delle une e delle altre seguendo lo stesso approccio.
- Giustificare **l'universalità** attraverso la **necessità incondizionata o *analiticità tautologica*** del fondamento logico delle une e delle altre – rispettivamente dell'**essere** e del **dover essere**.

- D’altra parte, sappiamo come il formalismo etico kantiano abbia prodotto nell’Ottocento la reazione dei cosiddetti **“tre maestri del sospetto”**, Freud, Marx e Nietzsche.
- Essi da tre prospettive diverse (psicanalitica, socioeconomica e filosofica) erano convergenti nel rivendicare **l’inseparabilità di componente emozionale e razionale** nel pensiero teorico e pratico dell’uomo.
- Tuttavia, perseverando nell’errore di confondere **universalità e vuota tautologicità** del pensiero frutto di ignoranza della logica, questa rivendicazione ha portato al **relativismo e al nihilismo** di gran parte del pensiero filosofico del ‘900 e quindi alla crisi della **rilevanza accademica, sociale e culturale** del sapere filosofico rispetto a quello scientifico.
- In questo senso le scoperte di Damasio hanno sostanziato dal punto di vista empirico delle **neuroscienze cognitive** questo recupero dell’inseparabilità fra componente emozionale e razionale, per l’irriducibile **carattere intenzionale** del pensiero umano, tipico della critica fenomenologica al formalismo cartesiano-kantiano nella modernità, e su cui E. Husserl ha scritto pagine memorabili, imputando al

razionalismo cartesiano-kantiano **la crisi della cultura scientifica e filosofica** dell'Europa del '900 (Husserl, 1970b).

- Non casualmente, Damasio stesso **ha esplicitamente aderito all'approccio intenzionale alle neuroscienze cognitive e all'etica** nei suoi libri più recenti (Damasio, 2010; 2018), testi che costituiscono lo sfondo dell'attuale programma di ricerca della neuroetica.
- Egli così rivendica che **l'*aboutness* intenzionale**, cioè la relazione intenzionale diretta-a-uno-scopo (*purposeful*) soggetto-oggetto, contro il solipsismo soggettivista dell'identificazione catesiana-kantiana del sé con l'autocoscienza, riguarda **la biologia prima che la psicologia** come, d'altra parte, lo stesso **Franz Brentano**, maestro di Husserl e ispiratore del movimento fenomenologico, aveva anticipato (Brentano, 1874).
- Infatti, tutti gli organismi viventi in quanto sistemi dissipativi **soddisfano una relazione omeostatica bidirezionale con i loro ambienti**.
  - Dove con **omeostasi** in sistemi dissipativi (o "termodinamicamente aperti") quali i viventi**,** ci ricorda Damasio, si intende termodinamicamente non un sistema **all'equilibrio con l'ambiente** – che nel caso specifico è un cadavere non un

vivente – né semplicemente un **sistema bilanciato fuori-dall'equilibrio** (= a **somma nulla dell'energia dal/al l'ambiente)** – ma un sistema bilanciato con l'ambiente mediante **complessi processi non-lineari di auto-regolazione** che sono ciò che distingue il vivente dal non-vivente.

- Nei cervelli, esiste così **una relazione bidirezionale tra cervello e corpo** e, attraverso la superficie corporea, **con l'ambiente fisico e umano esterno**, al fine di soddisfare gli obiettivi biologici e sociali dell'individuo umano (cfr. (Damasio, 2010) pp.91-108), e da cui dipende in ultima analisi **la stessa costruzione delle diverse culture**.
- In altri termini, **l'intenzionalità non è solo psicologica**, né riguarda solamente la sfera della coscienza ma è **primariamente biologica**. Su questa **base omeostatica**, Damasio raffinò la triangolazione delle neuroscienze cognitive (cfr. Prima Parte §3.3.1 e Figura 3), nella distinzione tra: 1) **la consapevolezza in prima persona** da lui definita il "sé-come-testimone" o la soggettiva "presenza-a-sé-stessi" o "io" di ogni persona umana; 2) **l'autocoscienza, come auto-oggettivazione** sempre necessariamente parziale a noi stessi del nostro sé-testimone; 3) **l'attribuzione del comportamento morale anche a qualcun altro** della nostra capacità di agire morale,

cioè l'attribuzione in terza persona di una coscienza morale (= o **valutazione consapevole della relazione scopo-mezzi**) anche ad altri esseri umani, sia individui che gruppi.

- Damasio, dunque, ha individuato **nel passaggio da 1) a 3) il proprio della ricerca neuroetica**, avendo nella persona, cioè "l'individuo-consapevole-in-relazione con il suo ambiente" l'agente morale, a causa dell'irrilevanza di 2), erroneamente identificato da Cartesio come l'**io in prima persona (l'io come *res cogitans, "cosa/oggetto pensante")***, soggetto dell'agire morale umano.
- Da questo punto di vista, è facile capire quale sia l'errore simmetrico rispetto a quello di Cartesio compiuto nella modernità. Cioè, l'attribuzione del ruolo di controllore del comportamento umano **al cervello isolato dal suo ambiente** corporeo e ambientale, come ha evidenziato uno dei più famosi neurofisiologi dei nostri tempi, Michael S. Gazzaniga (Gazzaniga, 2011).
- Questo errore, come è noto, si basa sull'evidenza neurofisiologica che ogni atto di volizione cosciente è preceduto dalla presenza di **potenziali elettrici d'azione** nei gruppi di neuroni causalmente coinvolti nella produzione di qualche azione senso-

motoria, come abbiamo visto quando abbiamo studiato il meccanismo di trasmissione sinaptica nelle RNN (cfr. Parte Prima §3.3.3 e Figura 9).

- I potenziali di azione **precedono di alcuni millisecondi** (Bernard Libet) (Libet & et al., 1983), o addirittura di alcuni **decimi di secondo** (John-Dylan Haynes) (Haynes, Roth, Stadler, & Heinze, 2007) **la manifestazione cosciente a noi stessi di compiere un atto volontario**, cosicché, usando il titolo dell'articolo appena citato di Haynes e dei suoi collaboratori, è possibile "leggere le intenzioni nascoste nel cervello umano".
- Da questa e da una quantità impressionante di altre evidenze neurofisiologiche al riguardo si è arrivati all'errore, che è simmetricamente opposto all'errore di Cartesio e di Kant quello per cui, **"la volizione cosciente, l'idea che tu stia volendo che un'azione accada, è un'illusione"** (Gazzaniga, 2011, p. 129).
- In ultima analisi, non l'io autocosciente di Cartesio, ma il suo cervello è **l'autore delle azioni dell'uomo,** visto che le evidenze neurofisiologiche dimostrano ampiamente non solo che la coscienza non può avere **nessun ruolo causale** in neurofisiologia – il che è banalmente ma sostanzialmente vero – ma che in ogni caso **arriverebbe sempre dopo l'azione causale dei neuroni.**

- In effetti soprattutto sui media statunitensi si è diffusa la falsa idea che quella precedente **dell'illusorietà della volizione cosciente** sia la tesi fondamentale della neuroetica. In particolare, un libro dello psicologo sociale Daniel M. Wegner, il cui titolo è tutto un programma *The Illusion of Conscious Will* pubblicato dallo MIT ha avuto un effetto devastante al riguardo (Wegner, 2002).
- Viceversa, al contrario l'obbiettivo scientifico della critica neuroetica e in genere delle neuroscienze è **l'identificazione cartesiana moderna dell'"io" come agente morale, con il "me" dell'autocoscienza,** e l'assurdità connessa **del ruolo causale della coscienza sul cervello** nell'agire umano.
- Viceversa, l'errore simmetrico di identificare materialisticamente nel **cervello il "vero" agente delle nostre scelte** facendo della libertà e della responsabilità morali un'illusione come fa Wegner deriva dalla falsa supposizione di considerare **il cervello come un agente isolato**. Come afferma Gazzaniga,

  «Il punto è che ora capiamo che dobbiamo guardare l'intero quadro, **un cervello al centro dell'interazione con altri cervelli**, non solo un cervello isolato (...). La natura umana rimane costante, ma fuori, nel mondo sociale, il comportamento può

cambiare e dei vincoli possono essere così posti sulle intenzioni inconsce» (Gazzaniga, 2011, p. 216).

- In altri termini, ciò che Gazzaniga sottolinea è che, alla luce delle neuroscienze, l' "essere responsabile" (*responsibility*) e il "dover rendere conto a" (*accountability*), come evidenzia lo stesso termine "responsabile" (= "rispondere a qualcun altro"), hanno principalmente **una natura relazionale sociale e quindi biologica** che non si oppone affatto alla natura razionale/personale di un atto morale.
- Basta recuperare l'equivalenza aristotelica della definizione dell'uomo come *biòs loghikòs* e *biòs politikòs,* "animale razionale" perché "animale sociale" e viceversa.
- A tale equivalenza la lettura cristiana di questa antropologia ha aggiunto il fondamento **di una relazionalità "verticale" di ogni uomo con l'Assoluto Trascendente** (= Dio Personale biblico) come **fondamento necessario** dell'essenziale uguaglianza **di tutte e di ciascuna persona umana** – aldilà dell'appartenenza a determinate culture/società e quindi aldilà di una **relazionalità puramente "orizzontale"**.

- Tale relazionalità è il fondamento trascendentale o metafisico, dello stesso **controllo responsabile** che ciascuna **persona come "individuo-in-relazione"** – non il suo solo cervello o la sua sola mente prese **isolatamente –** possono esercitare sulla componente inconscia ed emozionale del suo comportamento.
- È qui il fondamento della **civiltà occidentale** e della sua **etica personalistica** nelle sue storiche, imprescindibili **radici cristiane** che in questo senso ha una compiuta espressione nella terza definizione kantiana dell'imperativo categorico: *«Agisci in modo da trattare l'umanità, sia nella tua persona sia in quella di ogni altro, sempre anche come fine e mai semplicemente come mezzo».*
- In altri termini, i risultati della neuroetica e delle neuroscienze cognitive non si contrappongono affatto alla nozione di **responsabilità e libertà *personali.*** Si contrappongono solo **al monismo e/o al dualismo individualistici,** che sono le due ideologie opposte che hanno contaminato il pensiero occidentale per millenni, e contro le quali le neuroscienze possono dare un contributo essenziale e in qualche modo definitivo di decontaminazione.
- Per riassumere, **noi come "persone" siamo il nostro cervello e il nostro corpo in relazione con il nostro ambiente (metafisico-fisico-sociale):** il nostro

comportamento cosciente e la nostra consapevolezza “in prima persona”, il nostro “io”, sono solo **la cima dell'iceberg**, non l'intero iceberg del nostro **libero arbitrio morale**.

- Per dire la stessa cosa con una sapiente esemplificazione di Tommaso d’Aquino, affermare che solo la mente o solo il cervello e non **l’unità psicofisica di noi come *persone*** siano gli “autori” del nostro comportamento è altrettanto stupido quanto affermare che solo il martello o solo lo scalpello e non lo scultore siano gli autori della statua, semplicemente **perché lo scultore non può fare a meno di questi strumenti**.
- Nella terminologia scolastica, **la causa efficiente dell’azione cognitiva/deliberativa è la persona (individuo-in-relazione)**, la cui mente e il cui cervello (in relazione col loro ambiente fisico-sociale) sono solo **cause strumentali necessarie**.
- In questo **contesto relazionale: metafisico, fisico ed etico**, possiamo ridefinire la **libertà** come la facoltà di ogni **persona umana nel suo insieme**, comprese le sue imprescindibili relazioni col suo ambiente fisico e sociale, di **“autodeterminare il proprio comportamento in vista del raggiungimento effettivo di un certo scopo”**.

- Dove “scopo” è sinonimo di “fine consapevole”, così che la moralità di un atto libero – con buona pace di Kant e della sua falsa identificazione fra “universalità” e “non-condizionalità” della legge morale – dipende **dal carattere etico non individualistico** dello scopo perseguito (cfr. (Basti, 2003), cap. 5 per una sintesi).
- Per concludere con le parole di un altro rappresentante dell'approccio neuroetico, Neil Levy – Fondatore e Direttore del *Centre for Neuroethics,* ora *The Oxford Uehiro Centre for Practical Ethics* presso l’Università di Oxford che ha nella *Neuroethics* ma anche nell’*AI and Digital Ethics* due dei suoi principali campi di ricerca: https://www.practicalethics.ox.ac.uk/ – :

  «Pretendiamo giustamente che le nostre azioni e i nostri pensieri siano controllati da un agente, da noi stessi, e pretendiamo che a noi stessi siano attribuite le proprietà che valutiamo moralmente. **Ma l'unica cosa nel mente/cervello che risponde alla descrizione di un agente è l'intero insieme**, costituito da vari moduli e meccanismi sub-personali. Ed è l'intero agente che deve essere considerato il controllore dei processi controllati» (Levy, 2007), Kindle ed., pos. 395-396).

7.4.2. La distinzione neuroetica fra "responsabilità" personale e "responsività" cerebrale

- Per i nostri scopi fondativi di un'adeguata ME, è importante citare la posizione di Neil Levy sul dibattito neuroetico perché in un altro libro egli introduce una preziosa distinzione fondamentale per i nostri obiettivi.
- Cioè, la distinzione tra **"la responsabilità cosciente (*conscious responsibility*) *lenta*"** di noi come persone, *versus* "**la responsività inconscia (*unconscious responsiveness*) *veloce*"** dei nostri cervelli nel rispondere ai **vincoli ambientali fisici, sociali ma anche etici**.

  «Gli agenti coscienti possono esercitare un alto grado di controllo sui loro comportamenti **pur non essendo consapevoli dei fatti ai quali rispondono.** (…) [Un pilota come Ayrton] Senna reagisce velocemente alle auto attorno a lui e alle curve e ai tornanti della pista. Ma il tutto accade troppo velocemente perché i suoi processi coscienti possano tenere il passo. **È caratteristico dei processi coscienti essere molto più lenti di quelli inconsci**; la rapida responsività di agenti altamente addestrati come Senna deve certamente essere guidata da quest'ultima e non dai primi. **È**

dunque evidentemente 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** [non che sia consapevole dei processi cerebrali veloci di elaborazione dell'informazione da cui causalmente tale azioni dipendono]» (Levy, 2014, p. 114; 121).

- In tal modo viene introdotto nel dibatto neuroetico l'importante distinzione fra la **responsabilità morale di agenti di comunicazione coscienti** quali sono le nostre persone *versus* **la responsività etica di agenti di comunicazione inconsci** quali i nostri cervelli, ma anche i supporti decisionali dell'IA di cui facciamo uso, che possono entrambi essere inclusi nella categoria dei **moduli sub-personali** delle azioni di cui siamo moralmente responsabili.
- Questa distinzione può far luce però anche sulla questione dei sistemi **autonomi** di IA dotati di ML**,** intesi dunque non come oggetti ma come **soggetti inconsapevoli** di decisioni moralmente rilevanti**,** o se vogliamo come **agenti morali artificiali altamente addestrati,** mediante le tecniche di ML, ormai parte della nostra **società delle comunicazioni.**

- La domanda, dunque, è:
  - "In che senso possiamo applicare ad agenti inconsapevoli quali i sistemi di IA addestrati mediante ML una qualche forma di **responsabilità etica condivisa** con gli agenti morali consapevoli quali gli umani"?
- La preziosa notazione di Levy appena ricordata che per essere moralmente responsabile ciò che si richiede a un agente morale consapevole addestrato sia non la consapevolezza dei complessi processi di elaborazione dell'informazione che avvengono velocemente nei moduli neurali del suo cervello ma "una consapevolezza della **rilevanza morale** delle sue azioni", ci porta a riformulare la precedente questione in maniera più precisa per la ME:
  - "Come si può applicare tutto questo anche ad un sistema di IA *inconsapevole*, ma addestrato mediante ML"?
- **Tre passi** sono fondamentali per una risposta consistente a questa domanda:
  1. **Estensione anche ai moduli delle RNA della nozione di *responsività veloce (fast responsiveness)* dei moduli delle RNN cerebrali.** Il concetto di "responsività" o ***reattività veloce*** è infatti un concetto essenzialmente biologico che si applica in generale agli organismi e più specificamente alle **cellule** per significare

la capacità di una cellula sana di **adeguarsi velocemente mediante complessi processi di auto-regolazione alle modificazioni del suo ambiente fisico-chimico.** Se vogliamo, tale nozione è parte di quella più generale di **omeostasi** individuato da Damasio come fondamento biologico dell'**intenzionalità inconscia** (cfr. §7.3.1). È evidente che con processori nelle RNA che sono sette ordini di grandezza più veloci dei neuroni delle RNN il concetto di ***fast responsiveness*** versus the ***slow responsibility*** si applica a maggior ragione ai sistemi di IA. In questo senso occorre parlare di **responsività etica** e non di **responsabilità etica** dei sistemi di IA quando trattiamo il problema della **responsabilità condivisa uomo-macchina** in ME.

2. **Possibilità di implementare negli algoritmi di ML supervisionati, determinati criteri di *ottimizzazione etica* dell'apprendimento,** ovvero di **minimizzazione dell'errore rispetto a vincoli etici.** In altri termini, **implementare degli algoritmi di logica deontica (= "algoritmi *buoni* di ML"),** il che significa che il sistema diviene in grado di soddisfare **inconsapevolmente, ma effettivamente** nelle sue decisioni ultraveloci dei vincoli etici che rendano **socialmente**

e/o moralmente buoni le sue azioni, così da rendere **scientificamente consistente la nozione di *agente morale artificiale* in ME.**

3. **Possibilità di implementare nei sistemi di IA autonomi un *auditing etico automatico*** per il controllo che le decisioni prese dal sistema **soddisfino effettivamente i criteri etici posti nell'algoritmo di ML**. In effetti, è lo stesso che avviene nei giudizi pratici di noi umani. Ciascuno di noi **prima di emettere un giudio pratico,** su quale sia la decisione eticamente corretta da prendere in un determinato contesto pone **consapevolmente** dei vincoli etici alla sua operazione. Quindi, attraverso dei processi **del tutto inconsapevoli,** noi formuliamo/effettuiamo il nostro giudizio/decisione, per poi **controllare consapevolmente se effettivamente la decisione presa soddisfi ai criteri etici** che avevamo posto. Oppure – come si dice nelle Lettere di Paolo – "abbiamo fatto il male che non volevamo fare". In un certo senso, molto appropriato noi facciamo ***auditing etico* a noi stessi** nel ragionamento morale che segue al giudizio/decisione pratica che abbiamo effettuato, e che Kant avrebbe messo nell'ambito del **giudizio riflettente** della ragione nel suo uso pratico. Quella che nell'ambito della filosofia e della teologia morale si chiama **"esame di coscienza".** La

consapevolezza o, o se vogliamo **"la trasparenza-a-noi-stessi"** dell'eticità del nostro giudizio/decisione è solo dunque ***a parte ante*** **e** ***a parte post*** **del processo decisionale medesimo del tutto opaco e inconsapevole,** non solo agli altri, ma anche a noi stessi**.**

- Tutto questo ci fa capire (ma questa sarà la **Conclusione Generale** del nostro corso) che il vero problema del ME non è tanto quello della **irriducibile opacità del processo di ML** (che anzi rende questi sistemi del tutto simili alla mente umana), ma **nel non aver dotato questi sistemi -** ***a parte ante*** **nell'algoritmo di ML (= logica deontica del primo ordine), e** ***a parte post*** **negli algoritmi di IA simbolica** che devono effettuare un controllo "trasparente" (= logica deontica del second'ordine) dell'eticità delle decisioni prese dall'algoritmo di ML **-** di implementazioni algoritmiche adeguate di **calcoli di logica deontica,** rispettivamente al primo (***a parte ante***) e al secondo (***a parte post***) ordine.

7.5. Logica Deontica e la Nozione di "Algoritmo Buono" nel Machine Ethics

7.5.1. Alcune nozioni base di logica modale

- La **Logica Modale (LM)** è la logica della "necessità" e della "possibilità" – una distinzione di per sé priva di significato nella logica matematica dove tutte le affermazioni dimostrabili in una teoria **sono ugualmente necessarie**. In questo senso la LM – o logica filosofica – si distingue dall'ordinaria **logica matematica.**
- Il **Calcolo Modale (CM)** si caratterizza perciò rispetto al calcolo logico dell'ordinaria logica proposizionale e dei predicati per l'aggiunta di due nuovi simboli al loro alfabeto. **L'operatore di necessità** $\Box\alpha$ e **l'operatore di possibilità** $\Diamond\alpha = \neg\Box\neg\alpha$, dove α è il meta-simbolo per qualsiasi proposizione del calcolo logico.
- Il che significa, usando **la semantica *relazionale* modale "a molti mondi"** di Saul Kripke (Kripke, 1963; 1965), che **nella teoria modale dei modelli** abbiamo a che fare con verità o falsità di proposizioni che **non riguardano un solo stato di cose, o "mondo reale"**, come nella teoria standard dei modelli in logica matematica

(Tarski, 1935), ma anche con verità o falsità in altri possibili stati di cose o "mondi possibili" **che possiedono qualche relazione con quello reale**.

- Di conseguenza, in LM, una proposizione **sarà *necessaria* in un mondo**, **se è vera in *tutti* i mondi possibili relativi a quel mondo**, e ***possibile* in un mondo, se è vera *almeno in un altro mondo*, relativamente a quello precedente**.
- Ciò implica che in LM i connettivi logici (predicati proposizionali) **non sono vero-funzionali**, almeno nel senso di Frege legato all'uso delle **tavole di verità** per i connettivi/predicati proposizionali ("non", "e", "o", "se... allora", ...).
- Cioè, la verità delle proposizioni complesse che i connettivi logici formano non può essere semplicemente dedotta dalla verità dei loro argomenti (proposizioni elementari), usando le classiche tavole di verità {0,1} a due valori, ma richiedono **ulteriori assiomi che regolano l'uso dei due operatori modali, $\Box$, $\Diamond$** e che caratterizzano il CM rispetto all'ordinario calcolo proposizionale (Cresswell & Huges, 1996; Galvan, 1991).
- Per i nostri scopi, ricordiamo qui:
  - L'assioma **T** $:= \Box\alpha \rightarrow \alpha$ ("se α è vero in tutti i mondi possibili, allora è vero anche in quello attuale"), tipico delle logiche modali **aletiche** che hanno a che fare

cioè con la verità/falsità (**T** sta per *truth*) di proposizioni descrittive di stati di cose **necessariamente o possibilmente *vere*** in logica (necessità logica), oppure in fisica e metafisica (necessità causale, o **ontica**).

- L'assioma **D** $\coloneqq \Box\alpha \rightarrow \Diamond\alpha$ ("se α è necessario allora α è possibile"), tipico delle logiche modali **deontiche** (**D** sta per *deontic*) dove l'operatore $\Box$ sta per l'operatore deontico di **obbligo, O,** e l'operatore $\Diamond$ sta per l'operatore deontico di **permesso, P.** In questo senso l'assioma **D** è la formalizzazione della massima fondamentale di qualsiasi logica deontica *impossibilia nemo tenetur*: "nessuno può essere obbligato a qualcosa di impossibile".

♦ Per quanto riguarda le **logiche deontiche,** i "mondi possibili" riguardano qui i mondi "idealmente buoni" del **dover-essere** relati ai **valori etici** o **scopi** da essere perseguiti, in quanto distinti dal "mondo reale" dell'**essere**. O, più precisamente, sono quelli relati a **criteri di ottimalità** (per qualsiasi sistema di valori) e/o di **massimalità** (per un determinato sistema di valori) della "bontà" che determinate azioni devono soddisfare.

♦ In altri termini, nel caso **dell'obbligatorietà deontica** distinta dalla necessità logica, i "mondi possibili" interessati sono gli **stati buoni *s*** del mondo (personale,

fisico, sociale, economico, lavorativo, …) da raggiungere mediante azioni oggetto della norma *p*, per un dato **soggetto individuale/collettivo *x***. P.es, la norma per lo studente "devi studiare" dipende dalla bontà del valore "essere istruito" che altro non è che uno stato "buono" e quindi "desiderabile" fra quelli possibili del "mondo vitale" (*Lebenswelt*) dello studente, e che per essere effettivamente perseguito richiede necessariamente l'azione dello "studiare" che diventerà così moralmente buona, perché necessaria al conseguimento effettivo di un fine buono.

- Come si vede, ci muoviamo nell'ottica valoriale della **fondazione intenzionale dell'obbligo morale**, ovvero del suo carattere **condizionale** (con buona pace di Kant)**:** "Se vuoi essere istruito allora devi studiare".
- In tal modo, si introduce un operatore di **ottimalità deontica Ot** tipico della **logica assiologica (**"la logica dei valori") a due argomenti, cioè **Ot(*x, s*).**
- Pertanto, l'obbligatorietà etica espressa **dalla norma morale/giuridica *p***, cioè **O*p*,** regola le azioni che un soggetto morale individuale/collettivo *x* deve obbligatoriamente eseguire per raggiungere effettivamente **nel mondo reale** un dato stato di cose buono per lui/lei e/o per la società.

- Cioè l'operatore di obbligatorietà deontica **O***p* per le norme morali/legali *p,* che fa sì che la sua obbligatorietà renda la norma effettivamente seguita nel mondo reale delle azioni di un soggetto etico, ovvero diventi *p* per *x,* p_x, soddisfa il seguente assioma che include un criterio di ottimizzazione deontica. Ovvero, $\mathbf{O}p_x \coloneqq (\mathbf{Ot}(x,s) \wedge c_a \wedge c_{ni}) \rightarrow p_x$.
- Dove c_a e c_{ni} sono rispettivamente: la "condizione di accettazione" da parte di *x* della ottimalità dello stato *s* da perseguire mediante l'azione normata da *p*; e la "condizione di non-impedimento" per il soggetto *x* ad eseguire effettivamente l'azione normata da *p* per perseguire lo stato *s*.

7.5.2. Gli algoritmi "eticamente buoni" di ML

- Per passare dai principi generali di logica modale alla loro **implementazione algoritmica** bisogna innanzitutto ricordare la possibilità di sviluppare **logiche booleane modali** (Kupke, Kurz, & Venema, 2004; Venema, 2007) grazie alle quali è possibile implementare in esse e quindi in sistemi di IA la semantica relazionale di Kripke e quindi **l'etica relazionale**, compresa **un'etica dell'equità** (*fairness*) per evitare le "ingiustizie algoritmiche" negli algoritmi di ML di cui abbiamo accennato in

§7.2 (a tal proposito, cfr. (Gajane & Pechenizkiy, 2018) e (Basti, Capolupo, & Vitiello, 2020)).

- Infatti, ciò che è proprio della **semantica relazionale dei mondi possibili (modale) di Kripke** implementata in una logica algebrica booleana di tipo modale, e quindi **algoritmizzabile**, è che essa possa essere sviluppata, sia **al secondo ordine (*Higher Order Logic,* HOL) su un intero *universo* di mondi possibili**, sia **su *partizioni* (filtri o somme disgiunte o coprodotti) dell'universo di mondi possibili.**
- In questo caso, si tratterà di una **logica booleana del primo ordine (*First Order Logic,* FOL**) che alcuni fondamentali teoremi di logica dimostrati da Bjorn Jónsson e da Alfred Tarski negli anni '50 del secolo scorso garantiscono che si tratti di un **frammento protetto decidibile (*guarded decidable fragment*) della FOL** (Jónsson & Tarski, 1948; 1952a; 1952b)**,** che non ricadono cioè sotto le indecidibilità dei **Teoremi di Incompletezza di Gödel** (Gödel, 1931) e sono dunque **computabili da una macchina a stati finiti** quale un computer digitale o un sistema di IA (Goranko & Otto, 2007; Venema, 2007).
- Passando perciò all'**interpretazione deontica** della semantica relazionale di Kripke, ciò significa che possiamo implementare negli algoritmi "eticamente buoni" di

ML sia dei **vincoli etici in termini di <u>ottimalità</u> del valore da perseguire** (= valori validi per qualsiasi sistema etico, p.es., i diritti della *Dichiarazione Universale dei Diritti dell'Uomo* del 1948), sia dei **vincoli etici in termini di <u>massimalità</u> relativa del valore da perseguire** (= valori validi in un determinato contesto socio-economico o per un determinato sistema etico di valori di un particolare gruppo, etnia o religione).

- **In concreto,** parleremo di **stati possibili** del mondo secondo i diversi **contesti linguistici**. L'ottimizzazione deontica riguarda così **il perseguimento di stati valutati buon**i o desiderabili del mondo in base ad un determinato sistema valoriale, cosicché la norma *p* rende **obbligante** il perseguimento di quello stato ottimale del mondo e gli stati intermedi che lo precedono.
- P. es.: la scolarizzazione dei giovani è certamente un valore e quindi gli "stati buoni" desiderabili del mondo sono tutti quelli che favoriscono la scolarizzazione.
- In ogni caso, per tornare alla differenza fra **ottimalità e massimalità,** nel primo caso, si tratterà di un'implementazione in un algoritmo di ML della formula standard di obbligatorietà deontica di una norma *p* per un soggetto etico *x* per

ottimalità del valore da perseguire per tutti i contesti etici (effettivamente una formula di HOL) ricordata in § 7.4.1: $\mathbf{O}p_x \coloneqq (\mathbf{Ot}(x,s) \wedge c_a \wedge c_{ni}) \rightarrow p_x$.

- Nel secondo caso, si tratterà di un'implementazione in un algoritmo di ML di una formula **locale** di **massimalità** del valore da perseguire (stato "buono" del mondo, fisico, sociale, economico, …) **relativamente** a un contesto **etico specifico** per un dato soggetto (individuale o collettivo) $x \neq y$**.** Si tratterà dunque di una formula di FOL del tipo: $\mathbf{O}p_x \coloneqq (\mathbf{Max}(x, s_x) \wedge c_a \wedge c_{ni}) \rightarrow p_x \neq \mathbf{O}p_y \coloneqq \big(\mathbf{Ot}\big(y, s_y\big) \wedge c_a \wedge c_{ni}\big) \rightarrow p_y$.
- Inutile dire che un'etica dei valori basata su criteri di **massimalità relativa** è la logica deontica dei sistemi di ML e in particolare dei Transformer.
- In quest'ultimo caso tipico dell'**etica relazionale,** diventa fondamentale lo **"Assioma di Identità (Equivalenza) per Posizioni Differenti"** di soggetti (individuali e collettivi) *x, y,* sviluppato da A. Sen per formalizzare nella sua **Teoria delle Scelte Sociali** basata sui principi della *Giustizia Distributiva Comparata,* il fondamentale **principio di equità (*fairness*) socio-economica del *maximin***: "il massimo delle risorse a chi ha il minimo delle opportunità" visto come **principio di aggregazione delle variabili** (Sen, 2017, pp. 214-217)**.**

- Ovvero come **definizione corretta del numero dei gradi di libertà** di una funzione di distribuzione statistica "equa" delle risorse. Se vogliamo usare il linguaggio del ML, si tratta **di una definizione non-supervisionata "equa"** della funzione-costo il cui errore è da minimizzare **in un algoritmo supervisionato "eticamente equo" di ML** che eviti le "ingiustizie algoritmiche" cui si accennava in § 7.2).
- Per capirci aldilà delle formule, siccome nella "lotteria dell'esistenza" come la definiva il grande economista Adam Smith, non tutti nasciamo con le medesime opportunità, l'astratta o "incondizionata" **uguaglianza di tutti i cittadini di fronte alla legge** diventa foriera di ingiustizie (*summum ius summa iniuria*) se non è mitigata da **principi di equità** di cui il **maximin** è il fondamentale come il grande filosofo politico John Rawls ha ricordato ai liberisti e comunisti moderni, ambedue affetti ai poli opposti da un'inguaribile sindrome **egualitarista**.
- Per capirci, poniamo che x sia uno studente "povero" ma "dotato", è ovvio che il suo stato di "massima istruzione" nel suo contesto s_x non sarà mai quello dello studente y "ricco" s_y che può frequentare migliori università. È chiaro che per il

maximin occorre dare a x maggiori risorse (p.es., borse di studio) per far proprio lo stato s_y: ecco spiegato il cuore dell'Assioma di Identità fra contesti diversi di Sen.

- Tuttavia, in questo modo di fatto abbiamo più applicato l'astratto principio di maximin teorizzato da Rawls (che per un residuo esplicito di normativismo kantiano identificava lo stato di equità con **un ideale stato di eguali opportunità per tutti,** "ponendo un velo" sulle effettive e sempre mutevoli disparità), piuttosto che l'uso raffinato che ne fa Sen dotato di ben altra cultura matematico-statistica ma anche logica che fa del principio del maximin un raffinato strumento di **aggregazione delle variabili** per la definizione del numero dei gradi di libertà dello spazio di probabilità di una distribuzione statistica "equa" delle risorse, capace di **riadattarsi a contesti sempre mutevoli.**
- Come ricordato e sviluppato altrove (Basti, Capolupo, & Vitiello, 2020), il soddisfacimento **computazionalmente effettivo** del fondamentale **Assioma di Identità** fra contesti diversi e sempre mutevoli di Sen sopra ricordato, si può ottenere usando il principio del **raddoppio dei gradi di libertà** fra un sistema e il suo ambiente "esterno" come **algoritmo non-supervisionato di ML** (cfr. Parte II, §6.2) per **l'aggregazione delle variabili** (= definizione di un numero finito di gradi di libertà

dello spazio di probabilità) per ottenere delle due distribuzioni statistiche da far corrispondere **ogni volta,** un'unica distribuzione statistica che "equamente" le soddisfi.

- In ogni caso, a parte questi approfondimenti che qui non possiamo sviluppare ma che ho solo indicato per far comprendere in che senso può essere implementata un'**etica relazionale** in algoritmi di ML un punto deve esserci chiaro.
- In linea di principio, per creare algoritmi "eticamente" buoni di ML è sufficiente che l'algoritmo di ottimizzazione e/o di minimizzazione dell'errore in cui un algoritmo di ML ultimamente consiste, soddisfi anche dei **vincoli (*constraint*) di ottimalità/massimalità deontica** per determinati soggetti e stati di cose **in continua evoluzione**.
- Per esempio, negli algoritmi di ML largamente usati in IA nei sistemi automatizzati di *trading* nei mercati mondiali finanziari e azionari, sarà necessario inserire criteri di minimizzazione dell'errore che non consistano nella **massimizzazione incondizionata del profitto** ma una **massimizzazione del profitto condizionata** al soddisfacimento di valori etici quali la **massimizzazione del benessere relativo** per la società (escludendo, per esempio risorse che provengano da capitali mafiosi, del commercio della droga, delle armi, etc.) e/o la **massimizzazione del benessere**

relativo per soggetti individuali e collettivi in specifici contesti (p.es., i produttori di materie prime nei paesi in via di sviluppo, etc.).

- Già questo piccolo esempio ci fa vedere come l'enorme complessità dello sviluppo di una finanza e di un mercato ***ethically compliant*** richiedono necessariamente sistemi di IA avanzati e che questi, a loro volta, soddisfino i **principi del *Machine Ethics***.
- Una neonata disciplina che, come abbiamo visto, richiede immediatamente un forte sviluppo, in cui i filosofi del Terzo Millennio potranno e dovranno svolgere un ruolo essenziale **se adeguatamente preparati**.

8. Conclusioni: responsabilità condivisa uomo-macchina nel Machine Ethics

8.1. Responsabilità umana lenta *vs.* responsività veloce della macchina ai vincoli etici

- Come abbiamo visto, il concetto di **responsabilità condivisa uomo-macchina** che Floridi ha correttamente inserito nel dibattito sull'**etica nell'IA,** acquista una **consistenza teoretica** quanto mai necessaria mediante la distinzione e la complementarità emersa nel dibattito neuroetico, ma applicabile a maggior ragione ai sistemi di IA fra **responsabilità lenta** degli agenti comunicativi consci (gli umani) **e responsività veloce** degli agenti comunicativi inconsci (le macchine).
- La **questione della velocità di risposta,** infatti, non è solo questione tecnica, ma **sostanziale** per il **controllo etico** applicato ai sistemi automatici o **a controllo attivo** non solo di IA, ma ormai di qualsiasi sistema automatico dotato di **controllo**

"intelligente" o ***smart system*** che ormai stanno sostituendo tutta la tecnologia tradizionale elettronica ed elettromeccanica in qualsiasi settore.

- L'irriducibile e sempre maggiore divario fra la lentezza con cui noi esercitiamo il controllo come agenti etici coscienti – per non parlare dei ritardi "epici" della burocrazia, dei governi e delle istituzioni – sulle decisioni dei sistemi intelligenti inconsci, li rende di fatto **autonomi** nelle loro decisioni anche quando non fossero progettati per essere tali.
- L'unica soluzione è dunque inserire non solo nei sistemi di ML dell'IA, ma anche negli ***smart systems*** dei vincoli etici nei loro algoritmi adattivi alle veloci variazioni dell'ambiente in cui operano, e che li rende "intelligenti" e perciò così indispensabili a tutti noi.

8.2. L'auditing etico nei sistemi di IA con o senza ML

- Di per sé, un altro contributo che viene dalla riflessione neuroetica e neuroscientifica è quello che riguarda un'ulteriore fondamentale analogia fra intelligenza artificiale e intelligenza naturale, e che soddisfa **il test dell'imitazione** in una maniera inaspettata, ma significativa.

- Ovvero, **l'analogia dell'opacità irriducibile dei processi di elaborazione dell'informazione** in ambedue come giustamente Levy faceva notare nel suo libro già citato (Levy, 2014), sebbene limitandosi alla questione neuroetica nell'intelligenza naturale.
- Già abbiamo notato, come la consapevolezza è un qualcosa di **inoggettivabile** a chiunque, compreso il medesimo **soggetto cosciente** che esperimenta il proprio stato di coscienza, tant'è vero che filosofi e psicologi parlano a tale riguardo della **inoggettivabile presenza-di-noi-a-noi-stessi**.
- Per questo motivo i fenomenologi parlano giustamente del linguaggio della mente intenzionale di soggetti personali **consapevoli** come di un linguaggio "in-prima-persona" di un soggetto consapevole, distinguendo fra **consapevolezza e auto-coscienza**, essendo la seconda una sorta di necessariamente parziale **auto-oggettivazione** "in terza persona" del "sé" (*self*) all' "io"(*I*), un "io" che proprio per questo non è mai riducibile al "sé".
- Tutto questo però significa, che non solo gli algoritmi di ML basati su architetture multistrato ma anche le menti umane sono caratterizzati da **un'insuperabile "opacità"** nella produzione delle loro decisioni.

- Chi di noi può mai ricostruire il percorso con cui un'idea ci è venuta in mente? Ovvero, nessuno di noi è consapevole dei complessi e veloci processi di elaborazione dell'informazione nei nostri cervelli che **entrano necessariamente** nella produzione di un'idea (concetto) e/o di una decisione (giudizio).
- Ciò che è effettivamente consapevole – trasparente e non-opaco, a noi stessi innanzitutto e non solo agli altri – in un processo decisionale è ciò che è ***a parte ante*** **e** ***a parte post*** il processo decisionale stesso.
- Ovvero,
    1. ***A parte ante,*** la consapevolezza che la decisione che stiamo prendendo dovrà soddisfare dei vincoli etici**;** e
    2. ***A parte post,*** la consapevolezza oggettivata in un ragionamento, o inferenza logica linguisticamente espressa e dunque accessibile anche al controllo altrui (*accountability*), **che giustifichi formalmente l'eticità della nostra decisione**.
- Si tratta in questo caso di un processo di ***auditing*** **(rendicontazione/controllo) etico** che noi facciamo a noi stessi, ma che possiamo condividere con altri:

1. **Sull'eticità delle decisioni (giudizi pratici)** che abbiamo effettivamente preso e quindi **sulle scelte comportamentali**, che abbiamo effettivamente eseguito, e

2. **Su se e quanto** queste decisioni (giudizi) e scelte (azioni) **corrispondano a quei vincoli etici che ci eravamo dati** *a parte ante* e che avrebbero dovuto guidare le nostre decisioni e quindi le nostre scelte. O, viceversa, su se e quanto, queste decisioni/giudizi **non corrispondano a quei vincoli etici** che ci eravamo dati, per cui **di fatto abbiamo compiuto quel male che *non volevamo* fare.** Fra l'altro, in questo consiste la **colpa morale** perché compiuto con "piena avvertenza" e "deliberato consenso" un'azione cattiva.

- Proprio perché si tratta di ragionamenti etici ***a parte post*** esplicitati in inferenze espresse in un linguaggio, essi sono **assolutamente trasparenti al controllo intersoggettivo** e sono quindi alla base di quella **rendicontazione etica (*ethical accountability*)** con cui giustifichiamo la moralità dei nostri giudizi (decisioni) e delle nostre scelte (azioni) non solo a noi stessi, **ma anche agli altri**.
- Quando dunque discutiamo della **rendicontazione etica trasparente al controllo sociale** in ME dobbiamo distinguere anche per gli **agenti etici artificiali** i due momenti che precedono e seguono **l'opacità insormontabile** del processo di apprendimento automatico o ML:

1. **Se, quali e come siano stati implementati vincoli di ottimizzazione/massimizzazione etica** nell'algoritmo di ML di un determinato sistema di IA, al fine di evitarne il più possibile le "ingiustizie algoritmiche".
2. **Se le decisioni e le scelte di un sistema di IA (con o senza ML) effettivamente compiute** soddisfino determinati criteri etici, eventualmente implementati nell'albero inferenziale di un sistema di IA senza ML, oppure nell'algoritmo di ML di un sistema di IA che ne è provvisto.

- Paradossalmente, ma realmente questo secondo tipo di **analisi metalogica** generalmente si pensa che suppongano logiche di ordine superiore al primo o HOL, quindi il supporto di **dimostratori automatici** specializzati in compiti di **logica deontica** di ordine superiore al primo.
- Questo non è del tutto vero. Infatti se usiamo la metalogica della TC e non della TI si può evitare il ricorso alle HOL nelle logiche deontiche basate su criteri di **massimalità locale** e non di **ottimalità assoluta** dello stato desiderabile da perseguire.
- Quindi tutto ciò che da qui in poi si dice in riferimento a ragionatori meta-etici per garantire la **spiegabilità dei sistemi di IA** e quindi la loro ***ethical accoumtability*** supponendo un HOL deontica – che computazionalmente è altamente *time/resource*

consuming quindi **poco efficiente –** può e deve essere riletto nei termini di una FOL deontica. Ed è in questo tipo di logica che già questi ragionatori etici vengono usati in ME.

- Sebbene sconosciuti al grande pubblico perché usati quasi esclusivamente nell'ambito della **ricerca matematica pura ed applicata,** esistono ormai da diversi anni sistemi di questo tipo per **analisi metalogiche usando logiche di ordine superiore al primo (HOL)** di tipo **logico-matematico** (logico aletico).
- Essi sono applicati per **analisi di consistenza** di teoremi e teorie matematiche complesse e/o di algoritmi per **sistemi di controllo automatici** particolarmente complessi e che richiedono **un altissimo grado di robustezza all'errore** perché applicati in campi molto delicati, dove un errore o un'inconsistenza del programma avrebbe conseguenze catastrofiche, **non riparabili**.
- Si tratta dunque di meta-controlli per **l'analisi della sicurezza (*safety*) e affidabilità** (***reliabiliy***) di **sistemi di controllo automatico** particolarmente delicati, prima della loro commercializzazione/applicazione sul campo.

- Un'analisi che in campo informatico va sotto il nome del cosiddetto ***functional programming, programmazione funzionale*** o, più esattamente, della "analisi funzionale dei programmi".
  - Si pensi, per esempio, ai sistemi per **il controllo automatico dell'atterraggio degli aerei nei grandi aeroporti** che avvengono sempre col pilota automatico (il pilota cede i comandi dell'aeromobile alla torre di controllo dell'aeroporto e, con il supporto umano dei controllori di volo, il pilota resta solo per interventi di emergenza, in cui, sia detto fra parentesi, spesso potrebbe fare ben poco…).
  - Si pensi poi **ai sistemi di controllo degli scambi ferroviari** delle grandi stazioni ferroviarie, **ai sistemi di controllo dei viaggi spaziali**, **delle centrali nucleari**, **dei grandi complessi industriali**, **delle grandi reti di distribuzione dell'energia (elettricità/gas) e di telecomunicazione** dove le risorse vanno allocate e riallocate continuamente in base al grado di consumo e/o occupazione della rete nei vari orari e situazioni, etc.
- Solo, recentemente, causa il prorompente sviluppo delle problematiche inerenti alla ME si è aperto il nuovo settore di ricerca per **la progettazione e**

l'ingegnerizzazione di meta-sistemi di controllo nel campo **dei "ragionatori" etici (*ethical reasoners*)**, **delle teorie normative e delle logiche deontiche.**

- Come il mio amico Christoph Benzmüller della *Freie Universität Berlin* – uno dei maggiori esperti mondiali nel campo dei dimostratori automatici di teoremi usando la HOL – ha sintetizzato all'inizio di un suo bellissimo articolo di review sul nuovo campo di ricerca legato alla ME ed in cui presenta il suo rivoluzionario approccio alla problematica, «la motivazione principale è lo sviluppo di strumenti adeguati **per il controllo e la gestione (*governance*) di sistemi autonomi intelligenti**» (Benzmüller, Parenta, & van der Torre, 2020, p. 1).
- Senza entrare in ulteriori approfondimenti di un campo molto tecnico, il suo sistema denominato LOGIKEY (***Logi**c and **K**nowledge **E**ngineering Framework and Methodology*) «si basa sull'incorporazione semantica di logiche deontiche, di calcoli logici e di teorie che appartengono a uno (specifico) dominio etico-legale, entro una logica classica di ordine superiore (HOL) (…), come **la teoria dei tipi di Alonso Church»** (Benzmüller, Parenta, & van der Torre, 2020, pp. 1-2)**.**
- A titolo puramente esemplificativo è significativo che Benzmüller definisca il **meta-sistema di controllo sulla congruenza etica delle decisioni prese ma non**

ancora attuate all'esterno da sistemi autonomi di IA un «***ragionatore etico esplicito*** per sistemi autonomi intelligenti».

- Questo per evidenziare che, trattandosi di un sistema inferenziale di metalogica deontica tipo logico-simbolico esso è totalmente **trasparente** rispetto all'eventuale **opacità** delle inferenze logiche sub-simboliche del sistema autonomo di IA controllato, qualora esso sia dotato di ML. Un sistema di IA, dunque, il cui output è l'input del meta-controllore deontico.
- Dico "eventuale opacità", perché il **meta-sistema di controllo etico** può lavorare sia con sistemi di IA simbolica, che con sistemi di IA sub-simbolica dotati di ML. Sistemi che, comunque, si suppongono sempre dotati di una loro "competenza etica" (nel caso sub-simbolico, di vincoli etici sull'algoritmo di ML).
- Il ruolo del "ragionatore etico" (meta-sistema etico di controllo), infatti, è **la valutazione (*assessment*) e il giudizio (*judgement*) etici automatici** sull'output del sistema di IA controllato, cioè **sulle decisioni prese** ma non ancora attuate all'esterno da sistemi autonomi di IA senza o con ML, quindi senza o con **opacità** del processo inferenziale di decisione che essi autonomamente sviluppano.

♦ La seguente Figura 1 e la spiegazione che ne danno Benzmuller e gli altri Autori dell'articolo sintetizzano quanto detto finora:

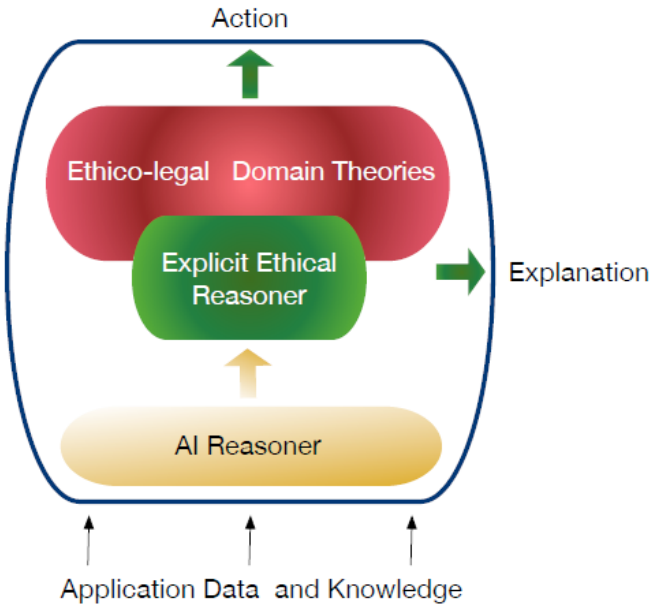

Figura 1. Ragionatore etico esplicito per sistemi intelligenti autonomi di IA (Benzmüller, Parenta, & van der Torre, 2020, p. 3).

«L'architettura visualizzata in Figura per un sistema autonomo intelligente con esplicita competenza etica distingue il **ragionatore etico esplicito** con il proprio dominio di

teorie etico-legale specifiche (cui applica le sue inferenze), dal **ragionatore/pianificatore di IA,** e da altri componenti, fra i quali ci sono anche i dati applicativi e la base di conoscenze disponibili per ambedue i ragionatori. Il ragionatore etico prende come input le azioni suggerite (decisioni) dal ragionatore/pianificatore di IA e le pone in riferimento: 1) da una parte alla base di dati e di conoscenze e 2) dall'altra, alle teorie di un determinato dominio etico/legale, e, alla fine, **produce valutazioni (*assessments*) e giudizi (*judgements*) riguardo quali azioni sono eticamente/legalmente accettabili o meno** e provvede anche **le corrispondenti spiegazioni**. In altri termini, le azioni suggerite dalle decisioni del ragionatore di IA in figura, **non sono eseguite immediatamente ma ulteriormente valutate dal ragionatore etico** che controlla se soddisfano o meno determinate teorie etico/legali. Questa valutazione è pensata per fornire **uno strato esplicito (trasparente) di spiegazione e controllo** all'apice del ragionatore di IA, che idealmente è già provvisto con una solida sua competenza morale. (…) E non è rilevante per l'architettura del sistema se il ragionatore di IA è basato su tecniche simboliche o sub-simboliche (ML), o da una combinazione di esse. Questo perché le azioni suggerite da esso **non sono eseguite subito**, (ma solo dopo la valutazione del

ragionatore etico), almeno quelle considerate le più critiche» (Benzmüller, Parenta, & van der Torre, 2020, pp. 3-4). Parentesi mie.

- Come si vede il sistema pensato da Benzmüller e i suoi colleghi soddisfa **a quel secondo momento della valutazione etica *a parte post* rispetto al ML** di un sistema di IA dotato di criteri etici nell'algoritmo di apprendimento automatico.
- Questo genere di **valutatore etico al secondo ordine** rispetto al quel soppesamento etico delle variabili che un sistema di ML, che incorpora nel suo algoritmo di ottimizzazione *constraints* etici **chiude il cerchio** che consente alla ME di parlare dei sistemi di IA autonomi sia di tipo simbolico o sub-simbolico (con ML) come veri e propri **agenti (soggetti) etici inconsapevoli.**
- Tutto questo completa il quadro della ME, anche se siamo davvero agli inizi di questo **nuovo campo di ricerca scientifica e filosofica** dagli impatti sociali, economici e culturali assolutamente rilevanti, anche per la **carriera professionale** dei futuri filosofi.

Bibliografia

Anderson, M., & Leigh Anderson, S. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine, 28*(4), 15–26.

Basti, G. (2003). *Filosofia dell'uomo. Second edition.* Bologna: ESD.

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

Basti, G., & Vitiello, G. (2022). A QFT Approach to Data Streaming in Natural and Artificial Neural Networks. *Proceedings, 81*, 106. doi:10.3390/proceedings2022081106

Basti, G., Capolupo, A., & Vitiello, G. (2020). The Computational Challenge of Amartya Sen's Social Choice Theory in Formal Philosophy. In R. Giovagnoli, & R. Lowe (A cura di), *The Logic of Social Practices. Studies in Applied Philosophy, Epistemology and Rational Ethics 52* (p. 87-119). Berlin-New York: Springer Nature. doi:10.1007/978-3-030-37305-4_7

Benzmüller, C., Parenta, X., & van der Torre, L. (2020). Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial Intelligence, 287*(103348), 1-50.

Birhane, A. (2021, February 12). Algorithmic Injustice: a relational ethics approach. *Patterns, 2*(2), 100205: 1-9. doi:10.1016/j.patter.2021.100205

Birhane, A., & Cummins, F. (2019, December 16). *Algorithmic injustice: toward a relational ethics.* Retrieved April 9, 2020, from https://arxiv.org/pdf/1912.07376v1.pdf

Brentano, F. (1874). *Psychologie vom empirischen Standpunkt.* Leipzig: Duncker & Humblot.

Cresswell, M. J., & Huges, G. E. (1996). *A new introduction to modal Logic.* London: Routledge.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: Putnam Publishing.

Damasio, A. (2010). *Self comes to mind: constructing the conscious brain* (1 ed.). London: Heinemann.

Damasio, A. (2018). *The strange order of things. Life, feeling and the making of cultures.* New York: Pantheon Books.

Dignum, V. (2018). Ethics in Artificial Intelligence: Introduction to the Special Issue. *Ethics and Inform. Techn., 20*(1), 1-3.

European Commission, Directorate-General for Research and Innovation, Unit RTD.01. (2018, March 9). *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems.* Retrieved from European Group on Ethics in Science and New Technologies: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/ethics-artificial-intelligence-statement-ege-released-2018-03-09_en

Gajane, P., & Pechenizkiy, M. (2018, May 28). *On Formalizing Fairness in Prediction with Machine Learning.* Retrieved from arXiv:1710.03184v3: https://arxiv.org/pdf/1710.03184v3.pdf

Galvan, S. (1991). *Logiche intensionali. Sistemi proposizionali di logica modale, deontica, epistemica.* Milano: Franco Angeli.

Gazzaniga, M. (2011). *Who is in charge? Free will and the science of the brain.* New York: Harper Collins Publ.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik, 38*, 173–98.

Goranko, V., & Otto, M. (2007). Model theory of modal logic. In P. Blackburn, F. J. van Benthem, & F. Wolter (Eds.), *Handbook of Modal Logic* (pp. 252-331). Amsterdam: Elsevier.

Haynes, J., Roth, G., Stadler, M., & Heinze, H. (2007). Reading intentions in the human brain. *Current Biology, 17*(4), 323-328.

Husserl, E. (1970b). *The Crisis of European Sciences and Trascendental Phenomenology. An Introduction to Phenomenological Philosophy.* (D. Carr, Trans.) Evanston, Illinois: Northwerstern UP.

Jónsson, B., & Tarski, A. (1948). Representation Problems for Relation Algebras. Abstract 89. *Bulletin of the AMS, 54*, 80 and 1192.

Jónsson, B., & Tarski, A. (1952a). Boolean algebras with operators, Part I. *American Journal of Mathematics, 73*, 891-939.

Jónsson, B., & Tarski, A. (1952b). Boolean algebras with operators, Part II. *American Journal of Mathematics, 74*, 127-152.

Kant, I. (1979). Fondazione della metafisica dei costumi. In I. Kant, *Scritti morali* (P. Chiodi, Trans.). Torino: UTET.

Kripke, S. A. (1963). Semantical analysis of modal logic I. Normal modal propositional logic calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik, 9*, 67-96.

Kripke, S. A. (1965). Semantical analysis of modal logic II. Non-normal modal propositional calculi. In J. W. Addison, L. Henkin, & A. Tarski (Eds.), *The Theory of Models* (pp. 206-220). Amsterdam: North Holland.

Kupke, C., Kurz, A., & Venema, Y. (2004). Stone coalgebras. *Theoretical computer science, 327*, 109-134.

Levy, N. (2007). *Neuroethics: Challenges for the 21st Century* . Cambridge UK: Cambridge UP.

Levy, N. (2014). *Consciousness and Moral Responsibility. Kindle edition.* Oxford UK: Oxford UP.

Libet, B., & et al. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act. *Brain, 106*(3), 623-642.

Müller, V. C. (2021, June 1). *Ethics of Artificial Intelligence and Robotics.* Retrieved November 6, 2022, from Stanford Encyclopedia of Philosophy, Edward N. Zalta (Ed.): https://https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Rutten, J. J. (2000). Universal coalgebra: a theory of systems. *Theor. Comp Sc., 249*(1), 3-80.

Sen, A. K. (2017). *Collective Choice and Social Welfare. Expanded Edition.* London: Penguin Ltd. Kindle Edition.

Tarski, A. (1935). The Concept of Truth in Formalized Languages. In J. Corcoran (A cura di), *Logic, Semantics, Metamathematics* (J. H. Woodger, Trad., 2 ed., p. 152–278). Indianapolis: Hackett, 1983.

Venema, Y. (2007). Algebras and co-algebras. In P. Blackburn, F. J. van Benthem, & F. Wolter (A cura di), *Handbook of modal logic* (p. 331-426). Amsterdam, North Holland: Elsevier.

Wegner, D. M. (2002). *The Illusion of Conscious Will.* Cambridge MA, USA: MIT Press.