UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di laurea in

Scienze cognitive psicologiche e psicobiologiche

Elaborato Finale

Ragionamento etico, Intelligenze Artificiali e neuroscienze: stato dell'arte e future prospettive di indagine

Ethical reasoning, Artificial Intelligence and Neuroscience: state of the art and further analysis perspectives

Relatore:

Prof. Andrea Spoto

Correlatore:

Dott. Giovanni Bruno

Laureando:

Giacomo Ursella

Matricola:

1221870

Anno accademico:2021/2022

INDICE

CAPITOLO 1: ETICA E METODI DI STUDIO

1.1 Evoluzione del concetto di etica

Aristotele definisce l'etica come una virtù, che in quanto tale non può essere innata, bensì è sviluppabile soltanto attraverso l'esercizio e lo studio quotidiano. Tale esperienza si sviluppa dinanzi a una situazione in cui è necessaria una risposta complessa che supera il bene individuale,che viene quindi elaborata sulla base di una serie di compromessi, condividendo contemporaneamente il male e il bene della relativa scelta. (Aristotele, IV secolo a.C.). Aristotele si è ispirato all'etica socratica che si basava sulla consapevolezza dell'uomo come *animale pratico, che si rifà alla concettualizzazione di etica pratica*: se un uomo predica il bene allora è un uomo buono, al contrario esso è malvagio se predica il male. (Aristotele, IV secolo a.C.). Il concetto di etica si è poi sviluppato nei secoli ed in differenti campi di applicazione ed approcci di studio, portando alla luce diversi campi di ricerca ed essa relate: psicologia etica, antropologia etica, filosofia, diritto etico.

Il concetto di pensiero critico sulla correttezza delle intenzioni è attribuibile già alle popolazioni mesopotamiche; per esempio, all'interno delle gesta di Gilgamesh ne "l'Epopea di Gilgamesh"(III millennio a.C) è possibile individuare valori etici e morali che, per gli antichi, un capo villaggio deve avere per la gestione del popolo. Nella storia si sono poi definiti diversi studi 'antropo-etici' che hanno permesso lo sviluppo delle varie materie di studio focalizzate sulla discussione del concetto di etica. Fra questi vi è ad esempio lo studio dell'etica religiosa, spesso concettualizzata negli scritti sacri ("ama il tuo prossimo tuo come te stesso", Nuovo Testamento, Matteo 22, 37-40). Dalla

visione religiosa è altresì possibile osservare il modo attraverso cui le varie correnti filosofiche hanno poi evoluto il concetto di etica, sviluppando nuove domande e discussioni in grado di astrarre nuovi punti di riferimento e declinazioni del concetto. (Tredimensioni 2, 2005)

A questo proposito nel 1800, periodo peraltro di grandi rivoluzioni e cambiamenti sociali e geopolitici, viene introdotta una nuova e determinante concettualizzazione filosofica su cui verrà basata la politica e la filosofia successiva, definita 'Utilitarismo'. Tale interpretazione morale descrive la ricerca dell'utilità nella forma di una ricerca del bene personale e collettivo. Definisce, infatti, le azioni come atto strettamente personale, etiche quanto una azione è *utile* in termini di felicità soggettiva. Padre di questa scuola di pensiero è stato Bentham (1776), il quale ha definito l'utilità di ogni azione solamente tramite le conseguenze che porta in termini personali, quindi come una ricerca del personale benessere, non solo a breve termine ma anche sul lungo periodo. (Canova, 2013). L'Utilitarismo di Bentham prende velocemente piede nel XIX secolo, dal momento che, in piena rivoluzione industriale, la politica e la vita in generale si sta evolvendo verso un prospettiva positivistica e materialistica; la definizione Benthamiana, perciò, si ritrova nella frase: *"la maggior felicità per il maggior numero possibile di uomini"*(Engelmann, S., 2010). La sua filosofia muove dalla principale critica all'ipotesi contrattualistica dei giusnaturalisti. Infatti, per i giusnaturalisti, come Hobbes, Locke e J.J. Rousseau, lo Stato avrebbe origine da un accordo volontario, da un contratto (idealmente) sottoscritto da tutti i cittadini, i quali precedentemente si trovavano al di fuori di ogni organizzazione politica allo "stato di natura". Bentham sostiene invece che alla base dello Stato non vi è alcun contratto sociale bensì una necessità utilitaria di promuovere collettivamente la felicità, il piacere di tutti. Da qui la riformulazione benthamiana di utilità comune come un rapporto

tra le utilità dei singoli cittadini: *"la massima felicità per il maggior numero di persone è la misura del giusto e dello sbagliato"*. (Engelmann, S., 2010).

A questo paradigma filosofico si contrappone la filosofia di Immanuel Kant che, in Critica della Ragion Pura (1788) definisce il concetto attraverso l'imperativo categorico: *"Agisci in modo che la massima della tua volontà possa sempre valere in ogni tempo come principio di legislazione universale"*. L'etica kantiana conserva in sé la condizione "deontologica", lo stesso autore infatti afferma che le azioni hanno delle limitatezze (i.e., Imperativi Categorici) che devono essere considerate nel momento in cui le stesse azioni vengono svolte nei confronti degli altri. Tra questi vi è la maniera in cui le altre persone risponderanno all'azione da noi svolta: infatti, secondo Kant è insensato osservare le conseguenze delle nostre azioni se non abbiamo oltrepassato i limiti dell'altro. Il concetto determinante di questa visione è il Rispetto verso l'altro. In tal senso, Kant osserva come, non ponendo attenzione alle azioni altrui in risposta alle proprie, allora è possibile che si agisca in maniera del tutto disinteressata di chi si ha di fronte. (Larmore, 2008). Nel 2017 ci fu un Ted Talk a cui partecipò come ospite protagonista Iyad Rahwan; l'argomento della conferenza, la concezione dell'etica, venne spiegato attraverso un esempio molto attuale che può ricevere due tipologie di risposta, queste ci permettono di paragonare in maniera pratica le teorie etiche di Bentham rispetto alle teorie di Kant. la discussione si basava sull'Intelligenza Artificiale (AI) e la guida autonoma. Dopo aver definito e spiegato in che modo funzionano i veicoli a guida autonoma, dimostrò che ci possono essere le risposte di stampo Benthamiano o Kantiano. Secondo Iyad, infatti, seguendo una ipotesi prettamente Benthamiana, di fronte a un ostacolo il veicolo autonomo seguirebbe una linea di risposta utilitaristica, volta, quindi, a ridurre al minimo il danno generale riportato ai vari soggetti coinvolti nell'incidente (sia gli interni che gli esterni al veicolo). Pertanto,

se la quantità di persone esterne alla macchina è maggiore in caso di incidente, la loro protezione prevarrà sulle persone interne al veicolo. La risposta Kantiana, invece, stabilisce che non devono esserci azioni volontarie che arrechino danno consapevole agli altri individui, quindi, la macchina dovrebbe proseguire la sua strada non curandosi di eventuali individui che si trovano sulla sua traiettoria di passaggio, dal momento che non può essere responsabile delle loro decisioni (Tabarelli, N., 2017).

1.2 I dilemmi morali

L'Uomo, dopo essersi concentrato sul concetto di etica e di morale, ha riflettuto sulla necessità di investigare questi concetti etici e, in questo senso anche il comportamento umano, spesso di fronte a situazioni di natura dicotomica e critica. I dilemmi morali divengono così strumenti applicativi per lo studio delle scelte e delle risposte etiche. Il dilemma morale è una situazione di particolare criticità etica o fisica, nella quale si è costretti a scegliere tra due alternative che comportano, ciascuna, situazioni indesiderate e negative (Grion, 2009). In letteratura si riconoscono vari esempi di dilemma morale, spesso suddivisi secondo diverse categorizzazioni. Due sono i principali modelli: la classificazione di Greene (dilemmi personali e impersonali; Greene, J., et al., 2002), e quella di Lariguett (da perdite di incommensurabilità, per sorteggi, sacrificali; 2008)

Tra queste, il presente elaborato si focalizzerà sulla classificazione di Greene, raffigurata classicamente attraverso gli scenari morali di Foot (1967) e Thomson (1976). I due scenari vengono così descritti:

-*trolley dilemma*: un treno si sta spingendo a forte velocità verso cinque persone legate al binario e l'unico modo per salvarle è spostare la leva delle rotaie cosicché il treno cambierà binario su cui però vi è intrappolata una persona che morirà.

-*footbridge dilemma*: costituisce una variante del Trolley in cui le cinque persone legate possono essere salvate solamente se il partecipante decide attivamente di spingere una persona grassa sui binari in modo tale da fermare il treno.

1.3 Moralità e neuroscienze

Negli anni, svariati studi si sono focalizzati sulla osservazione e valutazione delle attivazioni neurali ed emozionali dell'uomo di fronte a situazioni eticamente problematiche, utilizzando come strumento di studio proprio i dilemmi morali. Nei vari studi che spaziano tra l' indagare le attivazioni neurali, al comprendere perché in alcune popolazioni l'attivazione è diversa dalle altre, è stato osservato come l'etica non sia un concetto condiviso allo stesso modo da tutti, già dallo sviluppo etico nel singolo individuo. Kohlberg è stato uno dei primi sperimentatori a individuare queste differenze, di cui pone in cima alla lista l'età; bambini e adolescenti, avendo un sistema mentale molto più schematico, dividono in due momenti l'etica e risponderanno a un dilemma morale diversamente rispetto a un adulto o a loro stessi da adulti. Kohlberg, sulla base degli studi di Piaget sullo sviluppo (1920-1930), introduce una visione particolarmente interessante riguardo lo sviluppo della morale nei bambini, suddividendola in diversi stadi differenziati nelle varie età (realismo morale, morale dell'autonomia, morale eteronoma) (Kohlberg, 1985). Greene et al. (2001,2008) hanno poi potuto individuare che nell'attuazione del ragionamento etico il cervello adulto umano può rispondere seguendo due vie, una più "automatica" che vede solamente il tentare di fare il bene maggiore, e una più "ragionata", il sistema cognitivo deliberato; hanno messo a paragone due tipologie di dilemmi morali che al loro interno raffiguravano due scene di differente natura , osservando le relative attivazioni cerebrali in fase di immedesimazione e ragionamento.

Il dilemma definito personal, ovvero il "footbridge", implica un'azione diretta e volontaria di omicidio (spingere materialmente una persona giù da un ponte) atta al salvataggio di un maggior numero di persone, mentre il dilemma impersonal (o "trolley") rende l'atto utilitaristico indiretto, permettendo all'agente morale di salvaguardare la vita dei tanti a discapito di uno tramite l'azionamento di una leva.

. Seppur concettualmente ed 'economicamente' questi due dilemmi siano identici, studi neuropsicologici hanno dimostrato che vi sono tempi di risposta e attivazioni cerebrali differenti fra dilemmi personali ed impersonali. Nel caso personal (Footbridge) le risposte dei partecipanti sono molto più controllate ed emotive; infatti, si nota un'attivazione maggiore della corteccia prefrontale mediale -sede delle funzioni esecutive (FE)-, la corteccia cingolata posteriore, l'amigdala-sede delle emozioni- e una porzione neurale all'interno solco temporale e la giunzione tempero-parietale (Greene et al, 2002). Le attivazioni legate ai dilemmi di tipo personal vedono quindi anche l'attivazione di una porzione cerebrale molto profonda. È interessante osservare in tal senso l'attivazione dell'amigdala; l'amigdala è sede centrale del nostro sistema emozionale e, congruentemente all'attivazione della corteccia prefrontale e del lobo temporale (porzione cerebrale che determina il controllo emotivo e la risposta cosciente) fa in modo che la risposta sia decisamente più soppesata (motivo per cui i tempi di risposta sono più elevati in questo esperimento).

Al contrario, il dilemma del "Trolley", che definisce appunto un dilemma non direttamente compromettente, porta la maggiore attivazione nelle porzioni cerebrali legate alla Working Memory (WM), corteccia prefrontale dorsolaterale e lobo parietale. Le attivazioni in risposta a questo dilemma sono molto meno profonde e specifiche, questa tipologia di risposta, infatti, ha un status di attivazione molto più e immediato, mira infatti all'istinto del "salvare il più possibile".

1.4 Metodi di indagine

Dal momento che, come è stato detto in precedenza, alcune attivazioni cerebrali si situano in aree piuttosto profonde, è stato da subito necessario l'utilizzo di specifiche tecniche e strumentazioni scientifiche. Tra questi, la più funzionale alla localizzazione della attivazione neurale è sicuramente la Risonanza Magnetica funzionale (fMRI). La fMRI consiste in una tecnica di brain imaging funzionale che permette di conoscere l'attività neurale sulla base delle modificazioni delle concentrazioni di ossigeno nel sangue. Questi cambiamenti producono, di conseguenza, una distorsione nel campo magnetico, nota come "segnale BOLD" (blue oxygen level dependent). L'fMRI viene ampiamente utilizzata nell'ambito medico e di ricerca per investigare i processi cognitivi, in quanto va ad indagare in maniera approfondita e completa le varie attivazioni neuro cerebrali. (David, A., et al., 1994). L'uso della fMRI ha permesso di poter trarre conclusioni molto importanti sulla condizione emozionale del decision-making. Analizzando anche la moralità da un punto di vista filosofico, avvalendosi anche di studi fMRI, è possibile affermare che Kant e l'Utilitarismo erano ben lontani dal concetto di etica scollegato dal decision-making emozionale (Jana Schaich Borg et al. 2006).

Un'ulteriore modalità di ricerca, sempre a livello neuro cerebrale, ma meno utile all'osservazione delle attivazioni cerebrali, è l'EEG. L'elettroencefalogramma consiste in una tecnica d'indagine non invasiva in grado di misurare in tempo reale le modificazioni e le differenze di attività elettrica di gruppi di elettrodi posti sullo scalpo. (permette di indagare e monitorare l'attività elettrica cerebrale attraverso la ricezione di

onde elettriche posizionando un preciso numero di elettrodi sul cuoio capelluto). Questa tecnica, ampiamente utilizzata in psicofisica, permette di ottenere un tracciato dell'attività elettrica osservata formato dalla loro diversa frequenza e ampiezza. L'EEG risulta particolarmente utile in quanto, a partire dai tracciati registrati, è possibile osservare i potenziali evento-relati (ERP), utili a osservare le attivazioni cerebrali associate a specifici eventi di tipo sensoriale, percettivo, motorio o cognitivo. Esso costituisce, perciò, un valido strumento di indagine, soprattutto se unito ad altre analisi neurologiche, come per esempio l'fMRI (Biasiucci et al., 2019).

CAPITOLO 2: UN FOCUS SULLE ATTIVAZIONI CEREBRALI

2.1 Etica: basi neurali e attività cerebrali

Funk e Gazzaniga (2009) nella loro overview, svolta prettamente su studi neuropsicologici e neuroscientifici, hanno potuto affermare che vi sono diversi modelli e differenti attivazioni cerebrali che si spostano attraverso un continuum post-posto alla definizione di: cultura, sesso, età, e altre variabili individuali che sono determinanti per le teorie emozionali. In particolare, si sono soffermati sullo studio pioneristico di Greene, il quale riporta le differenti zone di attivazione cerebrali in relazione alla tipologia di azione morale da svolgere (dilemma morale: trolley o footbridge). Lo studio con fMRI di Greene (2001) osserva come le regioni interessate in una situazione di tipo soggettiva, ossia il footbridge dilemma, attivasse zone emozionali: giro medio-frontale, giro posteriore-laterale, giro angolare; mentre situazioni morali più oggettive, o impersonali, riportavano attivazioni di zone legate alla working memory, giro medio-frontale, corteccia parietale posteriore. Studi più attuali, sempre basati sulla teoria emozionale di Greene, hanno dimostrato che la teoria emozionale si basa su uno spettro che oscilla tra la iniquità – rappresentata dall'attivazione dell'insula – e l'ammirazione e la compassione – osservate dall'attivazione di networks nella corteccia

cingolata anteriore, sempre una zona anteriore dell'insula, ipotalamo e regioni corticali posteromediali.

Sono state indagate le risposte ai dilemmi morali da parte di pazienti split-brain, che hanno permesso la comprensione delle differenze dei due emisferi; l'emisfero sx, deputato alla comprensione linguistica dei dilemmi, non riceve informazioni dai network implicati nelle risposte emotigene, per esempio dalla giuntura tempero-parietale. Successive evidenze hanno riportato che la parte destra della corteccia frontale è specializzata nella comprensione delle intenzioni altrui attraverso i neuroni specchio (sede dell'empatia). Infatti, sconnettendo questo network si elimina la possibilità di integrare le varie informazioni che derivano dall'analisi del dilemma morale proposto (Gazzaniga, M.S., 2009).

A rafforzare gli studi su pazienti split-brain e sul loro ridotto vissuto emozionale, vi è lo studio condotto da Conor M. Steckler (2017) nel quale vengono investigate in maniera approfondita le differenze di morale in seguito a disconnessione emisferica. In esso, vengono definite le due zone emisferiche in maniera categoriale, l'emisfero sinistro, dedicato alla comprensione linguistica, e l'emisfero destro, dedicato in parte alla comprensione del dilemma morale, ma noto come principale sede dell'intenzione del dilemma morale. Sempre dello stesso autore, in letteratura è possibile ritrovare un test finalizzato allo studio dell'emisfero sinistro che ha permesso di identificare le risposte dei pazienti; i risultati affermano che queste risposte sono completamente basate sul risultato ottenuto piuttosto che sull'intenzione, non erano quindi in grado di prevedere cosa sarebbe successo in maniera autonoma. È stato quindi dimostrato che l'emisfero destro è necessario e sufficiente al giudizio morale basato sull'intenzione; mentre l'emisfero sinistro, nel momento in cui è disconnesso dal destro probabilmente non è sufficiente alla produzione di una risposta istintiva al presentarsi di un dilemma morale.

Può rispondere solamente ex posto facto, ovvero solamente nel momento in cui sia già presentato un risultato che può comprendere (Steckler et al, 2017).

Altri importanti scoperte e inferenze svolte su queste argomentazioni portano la firma di Young e Saxe (2009). Attraverso studi effettuati con neuroimmagini funzionali gli autori hanno potuto osservare che il modello alla base dei giudizi morali e delle risposte sia la Teoria della Mente, ovvero la capacità umana di rappresentarsi soggettivamente lo stato delle cose e delle persone esterne. Il loro studio si è articolato sulla lettura di un brano in cui il protagonista svolge delle azioni cosiddette "morali" e una in cui erano considerate "non morali". I risultati fMRI hanno permesso di affrontare la considerazione che la giunzione temporo parietale destra, il precuneo e la corteccia prefrontale mediale erano implicate direttamente ed erano le principali aree attivate in risposta a situazioni morali. Importante osservare che queste azioni svolte erano prive di una conseguenza sui protagonisti, cosa che invece viene modificata nella seconda parte dell'esperimento, nel quale i protagonisti esplicitamente affermano che la loro azione potrebbe come non potrebbe arrecare danno a terzi. Nel caso di non-moralità, le risposte della giunzione temporo-parietale dx, corteccia parietale e giunzione temporo-parietale sx, portano le attivazioni maggiori rispetto alle altre zone attivate. (Young, Saxe, 2009). Inoltre, ulteriori studi si sono focalizzati sull'utilizzo del primo assunto di Greene, relativo all'esistenza di una teoria emotiva nelle risposte etiche, introducendo un'ulteriore variabile, ossia la ToM, nel disegno sperimentale. I risultati osservati in questi studi hanno evidenziato come le attivazioni neurali in risposta al tipo di dilemma morale proposto rimanevano differenti indipendentemente dall'introduzione della variabile ToM (Heekeren H. R., et al., 2003).

L'fMRI, infatti, ha permesso di individuare che le aree implicate nella risposta sono così suddivise:

-Nel lobo temporale: solco temporale superiore del lobo temporale, il giro temporale medio sinistro, e il lobo temporale in maniera bilaterale

-Nel lobo frontale: corteccia prefrontale laterale sinistra, corteccia prefrontale medio ventrale (soprattutto per le risposte morali piuttosto che semantiche), esattamente come l'attivazione del precuneo.

In generale, vi è un aumento del segnale BOLD (*Blue Oxygen Level Dependent*) - ossia il processo di trasformazione dell'ossigeno durante l'attività neurale - nelle regioni prefrontali laterali e temporali. Le risposte semantiche producono un aumento dei segnali BOLD nella corteccia prefrontale dorsolaterale, il giro precentrale destro e il corpo caudato sinistro (Hekeeren et al, 2003).

2.2 Indagini etiche: studi neuropsicologici di attivazione cerebrale

Come già espresso in precedenza, Greene svolse un'interessante distinzione di attivazioni cerebrali a seconda del tipo di risposta etica che si desiderava ricevere, ma cosa comportano e perché nel cervello si attivano differenti aree a seconda della tipologia di dilemma morale proposto?

Innanzitutto, è necessario osservare in maniera più macroscopica le due differenti vie proposte da Greene: il dilemma di tipo personale attiva le aree legate e limitrofe al giro medio-frontale, giro posteriore-laterale, giro angolare. L'attivazione del lobo frontale in generale è molto interessante da un punto di vista neuropsicologico, infatti, tra i vari network che si attivano in relazione a risposte soggettive troviamo per esempio il CON, network circolo-opercolare, che ci permette di mantenere il task-set ma anche di gestire i conflitti di natura etica. Di questo network fanno parte anche l'insula, che abbiamo visto avere diverse attivazioni in situazioni di necessità morale, oppure il FPN, il network fronto-parietale, fondamentale per la flessibilità cognitiva e il

checking delle regole. L'unione di questi due network, che si sviluppano per tutta la corteccia prefrontale e una porzione di corteccia orbito-frontale, permetterebbe di comprendere la motivazione per cui le risposte agli esperimenti di tipo Footbridge sono più lunghe; infatti, il partecipante deve valutare attentamente il fatto di stare per compiere un omicidio, valutando pro e contro dell'azione e la sua complicità. Il giro posteriore laterale e il giro anteriore, indicano l'attivazione delle aree motorie, in particolare della corteccia motoria primaria, ma rappresentano anche l'attivazione di zone emotive, infatti è presente una risposta da parte dell'amigdala attraverso la via corticale, la via più lenta tra i due network di cui è composta l'amigdala; la via talamica, infatti, è più rapida e probabilmente è maggiormente coinvolta in risposta a dilemmi di tipo Trolley (Sadaghiani, S., et al., 2015).

Un' altra visione sulla funzione cerebrale, invece, ce la procura il dilemma di tipo Trolley. Infatti, studi successivi a Greene hanno dimostrato che le attivazioni cerebrali sono molto diversificate rispetto al Footbridge, nonostante il concetto del dilemma sia sostanzialmente simile; il dilemma impersonale è completamente privo della componente emozionale, dunque le risposte target sono molto più rapide. È interessante osservare l'attivazione della zona dove risiede la working memory (WM), la quale rappresenta il concetto di dover svolgere un compito tenendo conto delle sue regole. Sostanzialmente, l'attivazione di questa area definisce che il partecipante non sia emotivamente legato alla risposta ma sia semplicemente attento alle richieste dello sperimentatore, senza badare alle conseguenze che un suo gesto possa provocare.

2.3 La teoria della mente

Non è mancata un'interessante discussione sul concetto di TOM, Theory of Mind, all'interno delle decisioni etiche e morali richieste dai dilemmi. Alla base delle risposte etiche, è possibile infatti dedurre e trarre alcune conclusioni avendo alla mano

studi di fMRI e EEG. Non è insolito individuare attivazioni di zone ippocampali e para-ippocampali, al cui interno, infatti, risiedono le nostre memorie; tuttavia, non è nemmeno strano osservare attivazioni in zona PFC e OFC, al cui interno è stata individuata la "neocorteccia", zona ancora non del tutto compresa in cui vanno a consolidarsi i ricordi più antichi e solidi (Purves, D., et al., 2018).

È necessario fare luce su quale sia l'effettivo significato della Teoria della Mente nelle decisioni etiche. Considerando il concetto di "gruppo" negli esseri umani, possiamo osservare l'etica e la morale come una norma "condivisa" ma anche differente in base alla cultura e società di appartenenza e alle differenze interindividuali. Dalsant et al. scrissero all'interno del Giornale di Filosofia e Psicologia (2015) che la ToM è da considerarsi nel migliore dei casi come empatia, che si basa su assunti multilivello che si generano da stadi precoci e avanzano verso elaborazioni più complesse e che rispondono alla necessità di (r)esistere all'interno di interazioni ambientali, sociali e culturali. La ToM consiste nella capacità di attribuire stati mentali propri e altrui e nel riconoscere che anche negli altri vi è la presenza di credenze, desideri e pensieri (Dalsant, A., 2015). Attili, il cui lavoro viene citato anche da Dalsant, osserva come la ToM sia una risposta evoluzionistica alla necessità dell'uomo di essere in grado di vivere, cooperare e competere in gruppo (Attili, G., 2015).

Si può affermare, quindi, che la ToM sia caratterizzata da una componente innata, successivamente plasmata in fase di sviluppo, influenzata dapprima dal rapporto con i genitori e successivamente indirizzata in base alle credenze personali e dal gruppo di cui fa parte nel corso della sua vita. È quindi composta sia da piani cognitivi e pragmatici ma anche da situazioni interattive e relazionali (Dalsant et al, 2015). La capacità di elaborare l'esistenza di credenze altrui ha una corrispondenza di basi neurali specifiche; Rizzolatti e colleghi (Rizzolatti, G., et al., 2004) hanno studiato le attivazioni cerebrali

nelle scimmie in seguito all'azione o alla visione di essa; la principale attivazione è stata osservata nella corteccia motoria, considerata la base dei neuroni specchio, ossia neuroni che si attivano nel momento in cui viene effettuata un'azione o in cui essa viene vista svolgere. Studi successivi hanno potuto identificare anche nell'uomo la presenza di questi specifici neuroni, soprattutto in zone come il lobo parietale inferiore, giro precentrale e giro frontale inferiore e, come nelle scimmie, nella corteccia motoria e premotoria (Rizzolatti, G., et al. 2004).

Ponendo in confronto le basi neurali della Teoria della mente e le attivazioni cerebrali osservate dagli studi neuropsicologici su Trolley e Footbridge, è possibile delineare una sorta di sovrapposizione neurale delle attivazioni. È quindi possibile affermare che le risposte etiche ai dilemmi morali, sia emozionali e soggettivi che più impersonali e oggettivi, sono dettate non solo dalla considerazione e dalle credenze del partecipante, ma anche dalla consapevolezza che l'ambiente, la società e la cultura gli hanno donato nella sua evoluzione personale, dalla nascita fino a quel preciso momento.

CAPITOLO 3: MORALITÀ E INTELLIGENZE ARTIFICIALI (AI)

3.1 AI: lo stato dell'arte

Fondamentale nell'ambito delle ricerche e degli studi in psicologia è stato l'avvento dell'AI, la cui nascita viene fatta risalire al periodo della Seconda Guerra Mondiale quando il matematico Alan Turing (1912-1954) creò la macchina per decifrare il codice segreto Enigma, utilizzato dai nazisti per comunicare tra loro. Prima di morire, nel 1950, Turing scrisse un saggio dove propose un test per stabilire se una macchina possa pensare: secondo lui sì, se risponde a domande in modo indistinguibile dagli umani (c.d. Imitation game). Dopo pochi anni l'uscita del saggio, nel 1956, si tenne al Dartmouth College di Hanover (New Hampshire) uno storico convegno fra matematici, scienziati e tecnologi nel quale si proponeva un progetto di studio basato sulla "congettura per cui, in linea di principio, ogni aspetto dell'apprendimento e qualsiasi altra caratteristica

dell'intelligenza possono essere descritti in modo così preciso da permettere di costruire una macchina che li simuli" ((Barberis, M., 2021).

Questo è l'inizio della storia delle AI; una storia che ha conosciuto molti alti e bassi. Inizialmente si riteneva che sarebbe stato facile "costruire", o meglio, "programmare", un cervello umano in poco tempo, ma ciò si è rilevato essere un obiettivo molto poco realistico. L'iniziale positività ha lasciato pian piano il posto allo scettiscismo.

Nel primo periodo di evoluzione i vari scienziati erano particolarmente fiduciosi di questa novità nel campo delle scienze: Marvin Minsky, considerato uno dei padri fondatori dell'intelligenza artificiale, ha dichiarato in un'intervista, dopo circa 10 anni dalla nascita dell'AI, che in meno di dieci anni anni avremo una macchina con un livello di intelligenza comparabile con un quella di un uomo medio (Crevier, Daniel 1993).

.Elon Musk, fondatore e CEO di Tesla, circa 50 anni dopo a Minsky, ha affermato che i Veicoli Tesla completamente autonomi saranno stati inseriti nel mercato entro la fine del 2017. È ovvio che entrambe le dichiarazioni sono state smentite: stando alle ultime dichiarazioni di Toby Walsh, Professore ed esperto alla UNSW di Sidney, si ritiene che l'AI raggiungerà quella umana tra poco meno di 50 anni, ponendo anche, con non poca audacia stando allo storia della AI, una data, il 2062 (Walsh, T., 2018). Per quanto riguarda il CEO di Tesla, i lavori in corso sul rendere completamente autonome le Tesla sono ancora in atto e probabilmente lontani dal loro obiettivo finale. Il motivo di cotanta attesa è dato dal fatto che riuscire a integrare nel database della macchina il concetto di etica e una sorta di pensiero critico umano è una sfida di estrema complessità.

3.2 Etica e AI

Il tentativo di rendere i comportamenti AI il più possibile simili a quello umano "umane" si trova quindi a dover affrontare il grande ostacolo del pensiero razionale, si evolve progressivamente nel corso della nostra vita. Attualmente le AI sono

enormemente integrate nel quotidiano, supportando un gran numero di attività umane. Sempre più spesso vengono utilizzati navigatori per viaggiare in automobile, spesso già pre-inseriti con tutte le informazioni relative al traffico e allo status dell'abitacolo nel veicolo, le piattaforme di ricerca consigliano l'utilizzo di "cookie" per aumentare la visione ai siti più targettizzati, ma anche nello sport il monitoraggio delle risposte corporee sotto sforzo sono fondamentali per gli atleti professionisti e non (Rezzani, A., 2019).

Un articolo pubblicato su Slate da Satya Nadella (2016), amministratore delegato di Microsoft, riporta una riflessione di Cynthia Breazeal, professoressa all' MIT, nella quale afferma che l'uomo è la specie più integrata a livello sociale, e completa delle più forti emozioni attraverso le quali è in grado di relazionarsi ("dopo tutto, come la nostra esperienza si basa sulla comunicazione e collaborazione, se noi vogliamo che le macchine lavorino con noi allora non possiamo ignorare l'approccio umano"). Da queste parole Nadella afferma che vi sono cinque regole che le AI dovrebbero rispettare per una coesistenza pacifica con l'umano: (i) devono essere create per assistere gli umani, ad oggi infatti possiamo contare su programmi e siti vari ma anche su "robot" presenti nelle nostre case che ci aiutano nel quotidiano; (ii) devono massimizzare l'efficienza senza distruggere la dignità delle persone; (iii) devono essere create per la sicurezza dei dati e della privacy; (iv) devono avere un particolare algoritmo in modo che gli umani possano annullare l'eventuale danno non intenzionale; (v) devono essere protette dai bias ed evitarli.

Queste leggi sono giustamente paragonabili alle famose tre leggi della robotica di Asimov (1942):

-Un robot non può recar danno a un essere umano né può permettere che, a causa del proprio mancato intervento, un essere umano riceva danno.

-Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non contravvengano alla Prima Legge.

*- Un robot deve proteggere la propria esistenza, purché questa autodifesa non contrasti con la Prima e con la Seconda Legge (*Asimov, I., 1950)

Di riflesso anche l'uomo, per Nadella, ha alcuni "obblighi" in relazione alla macchina, ossia: (i) l'empatia (l'uomo deve essere pronto a creare relazioni e collaborazioni con il mondo); (ii) l'educazione e l'istruzione (arriverà infatti un momento nel quale sarà necessario lo studio di nuovi ambiti accademici, bisogna essere pronti a snocciolare la novità); (iii) la creatività (attraverso la quale sarà possibile l'innovazione e il cambiamento); (iv) giudizio e responsabilità (osservando le leggi di Nadella e le leggi della robotica, l'uomo deve essere sempre consapevole di fare del giusto e non usufruire di questa Intelligenza per scopi oltraggiosi); (Nadella, 2016).

Oggi, l'etica delle AI riguarda, tra gli argomenti, anche il progresso dei meccanismi di guida autonoma.

Alvin Dubicki del MIT afferma che lo sviluppo etico delle AI non è al passo con lo sviluppo tecnologico che c'è stato negli ultimi 20 anni. E' utile osservare le risposte all'esperimento pubblicato da Awad nel 2018 su Nature, che permette di comprendere al meglio tutti i vari dilemmi etici a cui si va incontro. L'esperimento, attualmente ancora attivo, è il "moral machine", da cui gli studiosi che hanno preso parte al progetto hanno riscontrato, tra le milioni di persone che hanno partecipato da più di 10 paesi differenti, che le scelte su base etica sono molto differenti tra di loro a seconda della cultura e del paese di provenienza. Non bastasse questo grande limite, la difficoltà che attende i

programmatori aumenta nel momento in cui è necessario considerare le varie caratteristiche personali di ciascun partecipante, come le differenze di età, sesso ed educazione, che giocano un ruolo molto importante nella scelta della risposta etica (Awad *et al.* ,2018).

3.3 AI e guida autonoma

Per la guida autonoma le varie idee che si stanno avvalorando con il tempo per riuscire a ovviare il problema dell'etica e della decisione morale su strada si stanno ampliando nelle varie case di produzione, oltre a Tesla troviamo anche il Volkswagen Audi Group (VAG), Google, Nissan etc. I fondi stanziati per la ricerca sono di enormi dimensioni, secondo Felix Demaeght, sono più di 16 miliardi di dollari spesi per la ricerca. Attualmente, comunque, ci si può affidare a funzioni relativamente autonome, meglio definiti "di livello 3", come per esempio il sistema Advanced Driver-Assistance System (ADAS) che permette al conducente di avere un'assistenza in caso di emergenza o di stanchezza e lasciare al veicolo la completa autonomia di trasporto ma solo per breve tempo (Felic Demaeght, il giornale, quotidiano online)**.** Lo stato dell'arte attuale delle AI alla guida di automobili ha superato con un ottimo successo le aspettative dei conducenti di tutto il mondo, in trend molto più che positivo sono le vendite tanto che ne sentiamo parlare sempre di più ogni giorno che passa. Ormai l'autonomia dell'auto è una realtà, e stando agli attuali progetti in campo è necessario da parte nostra doverci abituare all'idea che cambio e frizione diverranno obsoleti nel giro di pochi decenni ma soprattutto deve partire da noi la capacità di accettare un cambiamento, probabilmente anche culturale, dal punto di vista dello spostamento; a Waymo nello stato di New York sono già stati attivati nel 2016 taxi di 5 generazione, ovvero completamente autonomi,

che in un lasso di tempo non troppo definito ma relativamente precoce è possibile che diventi una realtà anche in Europa e nel mondo.

3.4. Prospettive future

Per rispondere a questa domanda è necessario prendere ad esempio le critiche e le posizioni dei diversi studi affrontati in letteratura; gli autori dell'esperimento pubblicato su Nature, ad esempio, auspicano che in futuro le intelligenze artificiali saranno in grado di unire tutte le differenze personali e culturali attraverso le risposte al "the moral machine experiment" integrando una sorta di decision making che tutt'ora è una delle funzioni mancanti delle AI (Awad, E., et al., 2018). Un'intervista svolta durante il Festival dell'Economia di Trento alla fine di maggio di quest'anno riporta le parole di Francesco Profumo, uno dei leader della Fondazione Bruno Kessler (FBK), un'associazione in prima linea nella ricerca delle AI (come riporta la prima pagina del loro sito internet). Egli afferma che è necessario un "cambiamento paradigmatico", attraverso cui le macchine e l'umanità coesistono in un unico ecosistema nel quale sia possibile giungere a una intelligenza artificiale generativa di nuovi scenari e nuovi dati che il cervello umano non ha la capacità di generare o risolvere (il Trentino, quotidiano online della provincia autonoma di Trento, 2022).

Le prospettive future oggi si basano su un approccio molto sicuro e propositivo. Vi è una grande consapevolezza della necessità di cambiamento radicale sia da parte della tecnologia che degli atteggiamenti nei confronti di essa; siamo sulla strada che ci porterà a un mondo completamento interconnesso e probabilmente molto più accessibile, che però stando allo stato dell'arte attuale necessita di un'implementazione dal gusto più filosofico che prettamente scientifico.

CONCLUSIONI

Osservati i vari scenari attuali di evoluzione tecnologica della AI è immediata, e forse necessaria, la riflessione sulla figura dell'uomo e come egli si osserva di fronte alla costruzione di una nuova Intelligenza; l'uomo per natura non è perfetto e soprattutto ogni umano è unico e rappresentato dal proprio pensiero, le risposte ai test etici dimostrano solo poche delle differenze che contraddistingue ciascuno: è quindi possibile che si riesca a costruire un'intelligenza priva di errori e che funzioni in maniera autonoma? Probabilmente ci si riuscirà, per ora la condizione di autonomia intesa nel senso di autosostenibilità è forse ancora lontano; le ricerche però non devono fermarsi e anzi è necessaria un'abituazione al nuovo e, come ha affermato Satya Nadella, una nuova educazione alla collaborazione potrebbe aumentare le probabilità di successo della ricerca.

Riferimenti bibliografici

Aristotele, Etica Nicomachea, IV secolo a.C.

Asimov, I., 1950, Io, Robot

Attili, G. (2015). L'evoluzione della Teoria della Mente. *Rivista internazionale di Filosofia e Psicologia, 6*(2), 222-237. DOI: https://doi.org/10.4453/rifp.2015.0020

Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018). https://doi.org/10.1038/s41586-018-0637-6

Barberis, M. (2021). Ecologia della Rete - *come usare internet e vivere felici*. Mimesis Edizioni.

Biasiucci, A., Franceschiello, B., Murray, M. M. (2019). Electroencephalography. Current Biology: CB, 29(3), R80-R85 DOI: https://doi.org/10.1016/j.cub.2018.11.052

Bibbia, Nuovo Testamento, Matteo, 22

Borg, J.S., Hynes, C., Van Horn, J., Scott Grafton, Sinnott-Armstrong, W.; Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *J Cogn Neurosci* 2006; 18 (5): 803–817. doi: https://doi.org/10.1162/jocn.2006.18.5.803

Crevier, D., (1993). *AI: The Tumultuous Search for Artificial Intelligence*. New York, NY: BasicBooks. ISBN 0-465-02997-3.

Canova, L., 2013, “l’Utilitarismo da Jeremy Bentham a John Stuart Mill”, Greenreport.it quotidiano di ecologia online, https://greenreport.it/news/comunicazione/lutilitarismo-da-jeremy-bentham-a-john-stuart-mill-videolezione/

Dalsant, A., Truzzi, A., Setoh, P., & Esposito, G. (2015). Empatia e Teoria della Mente: un unico meccanismo cognitivo. Rivista internazionale di Filosofia e Psicologia, 6(2), 245-248. doi:https://doi.org/10.4453/rifp.2015.0022

David, A., Blamire, A., & Breiter, H. (1994). Functional Magnetic Resonance Imaging: A new technique with implications for psychology and psychiatry. *British Journal of Psychiatry,164*(1), 2-7. doi:10.1192/bjp.164.1.2

Engelmann, S. (2010). Philip Schofield, Utility and Democracy: The Political Thought of Jeremy Bentham (Oxford: Oxford University Press, 2006), pp. xii 370. *Utilitas, 22*(1), 98-101. doi:10.1017/S0953820809990446

“Epopea del Gilgamesh”. (III millennio a.C.)

Foot, P. (1967). *The problem of abortion and the doctrine of double effect*. Oxford: Blackwell. (Reprinted in Foot (1978), *Virtues and Vices*.)

Funk, C.M., Gazzaniga, M.S., The functional brain architecture of human morality, Current Opinion in Neurobiology, Volume 19, Issue 6, 2009, Pages 678-681, ISSN 0959-4388, https://doi.org/10.1016/j.conb.2009.09.011.

Gibbs J, C: Kohlberg’s Moral Stage Theory. Human Development 1979;22:89-112. DOI: 10.1159/000272431

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. The Neural Bases of Cognitive Conflict and Control in Moral Judgment, Neuron, Volume 44, Issue 2, 2004, Pages 389-400, ISSN 0896-6273, https://doi.org/10.1016/j.neuron.2004.09.027.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, *6*(12), 517-523.

Heekeren, H. R., Wartenburger, I., Schmidt, H., Schwintowski, H. P., & Villringer, A. (2003). An fMRI study of simple ethical decision-making. *Neuroreport*, *14*(9), 1215–1219. https://doi.org/10.1097/00001756-200307010-00005

Il Trentino, quotidiano online della provincia autonoma di Trento. (2022). Intervista a Francesco Profumo

Larmore, C. (2008). Dare ragioni: Il soggetto, l'etica, la politica. Torino: Rosenberg & Sellier. doi:10.4000/books.res.289

Kant, I. (1788), Critica della Ragion Pura

Nadella, S. (2016). The partnership of the Future, Slate

Purves, D., et al. (2018) Neuroscience. 6th Edition, Sinauer Associates.

Sadaghiani, S., & D'Esposito, M. (2015). Functional Characterization of the Cingulo-Opercular Network in the Maintenance of Tonic Alertness. *Cerebral cortex (New York, N.Y. : 1991)*, *25*(9), 2763–2773. https://doi.org/10.1093/cercor/bhu072

Steckler, C.M., Hamlin, J.K., MIller, M.B., King, D., Kingstone, A. (2017). Moral judgement by the disconnected left and right cerebral hemispheres: a split-brain investigation *R. Soc. open sci.***4**170172170172 http://doi.org/10.1098/rsos.170172

Tabarelli, N. (2017). Il Dilemma etico della guida autonoma, 2017, WU Magazine

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, 94, 1395–1415. doi:10.2307/796133

Tragici dilemmi e conflitti, Una investigación conceptual, Bogotá: Palestra-Temis, 2008, p.90

Tredimensioni 2. (2005) 1, 4-11

Rezzani A. (2019). L'intelligenza artificiale nella vita quotidiana, Data Skills understanding the world

Rizzolatti, G., Craighero, L. (2004), The Mirror-neuron System. In: «Annual Review of Neuroscience», vol. XXVII, pp. 169-192

Walsh, T., (2018), 2062: The World that AI Made.

Young L.,Saxe R. (2009). An fMRI Investigation of Spontaneous Mental State Inference for Moral Judgment,Journal of Cognitive Neuroscience, DOI 10.1162/jocn.2009.21137