UNIVERSITÀ DEGLI STUDI DI NAPOLI

FEDERICO II

Intelligenza Artificiale ed Etica:
La ricerca in IA alla prova delle sfide etiche

6 dicembre 2019

Aula Seminari del DIETI (ex aula Softel)

1° piano, Ed. 3, Via Claudio 21, Napoli

Prof.ssa Anna Corazza

De Lucia Gianluca (N97000311)

Sommario

1. Introduzione

In questa relazione si discuterà intorno alle questioni etiche e giuridiche sollevate dall'Intelligenza Artificiale (IA) e alle prospettive che tali questioni aprono per la ricerca scientifica e tecnologica in IA.
Il workshop dal titolo "*Intelligenza Artificiale Ed Etica. La ricerca in IA alla prova delle sfide etiche*" tenutosi il 6 dicembre 2019 nell'aula seminari del DIETI, Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, in via Claudio 21.

L'analisi di questioni etiche è stato il punto centrale degli interventi del workshop: la manipolazione dell'opinione pubblica e la creazione di "echo chambers" mediante l'IA, le decisioni delle armi autonome e il diritto internazionale, le responsabilità dell'uomo alla luce delle sue limitate capacità di interpretare, spiegare e prevedere i sistemi dell'IA.
Il workshop ha raccolto ricercatori che operano in vari settori - scientifici, tecnologici e umanistici.
Al centro di altri contributi ci sono stati i problemi di comprensione e controllo dei sistemi di IA. Si tratta di problemi per la ricerca in IA che hanno forti motivazioni di tipo etico-giuridico: trasparenza, interpretazione simbolica e spiegazione dei processi e delle decisioni dei sistemi dell'IA, conformità dei sistemi dell'IA alle preferenze o agli obblighi morali di gestori e utenti, la fiducia che l'uomo può in essi riporre.

Nel programma del workshop, i contributi dei due tipi (di analisi etico-giuridica dell'IA e di ricerca in IA con motivazioni etico-giuridiche) sono intercalati tra loro. Questa scelta riflette l'intenzione di favorire il dialogo e l'interazione produttiva tra ricercatori, diversi per formazione – umanistica, scientifica, tecnologica – ma accomunati dall'obiettivo di capire e affrontare le nuove sfide etiche per l'IA.

1.1. Relatori

Al workshop sono intervenute varie personalità di spicco del settore.

In ordine di apparizione i relatori sono stati:

- Daniele Amoroso, Dipartimento di Giurisprudenza, Università di Cagliari
- Roberto Prevete, DIETI, Università di Napoli Federico II
- Viola Schiaffonati, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
- Riccardo Guidotti, KDD-Lab – ISTI-CNR, Pisa
- Luciano Serafini, ICT, Fondazione Bruno Kessler, Trento
- José M. Galvan, Dipartimento di Teologia Morale, Pontificia Università della Santa Croce, Roma
- Paola Inverardi, DISIM, Università degli studi dell'Aquila
- Piero A. Bonatti, DIETI, Università di Napoli Federico II

2. Workshop

2.1. Mattina

2.1.1 Daniele Amoroso: *Sistemi d'arma autonomi e la (ir)responsabilità giuridica e morale*

Nella prima fase del seminario ci si è accostati al problema di interpretazione Black Box dei sistemi AI e la spiegabilità in termini etico-giuridico.

Ad aprire ed introdurre il workshop è stato il professore Guglielmo Tamburrini, il quale ha dato per prima la parola al professore Daniele Amoroso.

Il suo intervento è stato basato sui sistemi d'arma autonomi e la responsabilità giuridica e morale ad essi legati.

Ci troviamo in una società che comprende stati che vogliono utilizzare Armi autonomo con AI le quali selezionano e ingaggiano gli obbiettivi senza l'intervento umano, in maniera del tutto indipendente. Ma se l'ingaggio prevede obiettivi critici come ad esempio bunker militare ma viene colpita una scuola, chi ha la responsabilità?

Ci sono vari problemi, dal punto di vista tecnico una macchina svolge funzioni che normalmente esegue un uomo: è lecito lasciare a una macchina decidere sul diritto alla vita di un uomo?

Già nel 2004 Mattias e poi nel 2007 Sparrow parlarono ed iniziarono a discutere dell'imprevedibilità dei sistemi autonomi. Il professore Guglielmo Tamburrini nel 2016 disse che questa imprevedibilità era legata all'ambiente operativo. Dal punto di vista giuridico esiste il concetto di *meno rea*, ossia non si può dare la colpa a nessuno per l'imprevedibilità poiché ci sono molte mani ad operare su um progetto come quello delle armi autonome, motivo per cui è lecito allargare la responsabilità dall'individuo alla collettività.

Una proposta di riforma che si vuole portare avanti è l'abbassamento di responsabilità per l'individuo e aumentare quella degli stati che decidono di usare tali armi.

I professori Amoroso e Tamburrini in questo anno hanno quindi elaborato questa proposta, con l'aggiunta di un controllo umano significativo sulle decisioni di queste armi, distinguendo in base ai contesti, migliorando la partnership tra uomo e macchina tramite un training diverso.

2.1.2 Roberto Prevete: *Controllo etico dei sistemi di machine learning*

Il professore ha introdotto come metodo di controllo di un sistema autonomo, la spiegabilità delle decisioni di un Machine Learning System e la loro importanza.

Si è incentrato soprattutto nello spiegare come lavora un classificatore e come deve essere chiara e comprensibile la sua decisione.

La spiegazione deve essere comprensibile dall'uomo, attraverso pochi elementi che devono essere contrastivi e ben chiari.

Diventa quindi necessario l'utilizzo di un dizionario che rappresenti il dominio di decisione, ossia una piccola regione significativa che contraddistingue un'immagine. Tale dizionario deve essere ottenuto dal dataset di partenza, i suoi elementi sono detti atomi.

Da ogni elemento del dizionario potrò quindi ricostruire, con un minimo margine di errore, ogni elemento del dataset iniziale.

Essenzialmente si ha un oracolo che riceve un input ed elaborandolo restituisce un output. Per ottenere un controllo basta aggiungere un mediatore che risponde a un interrogatore per capire il motivo di una decisione.

Il mediatore risponderà con una percentuale, inoltre attraverso gli atomi abbiamo la validazione del risultato attraverso la corrispondenza con l'immagine in input.

2.1.3 Viola Schiaffonati: *AI e metodo scientifico: dall'epistemologia*

La professoressa Schiaffonati ha messo in discussione il metodo di analisi delle AI poiché spesso, da come emerge da più di 400 articoli scientifici, non è possibile avere la riproducibilità dei lavori e delle applicazioni. Un esperimento per essere tale deve avere due caratteristiche portanti: riproducibilità e ripetibilità. E' necessario replicare e ripetere un esperimento in maniera esatta per attenersi al metodo scientifico tradizionale ma questo non avviene ancora nel campo del Machine Learning in maniera corretta.

L'informatica per essere considerata una scienza deve usare i metodi delle scienze tradizionali, e quindi attenersi al metodo scientifico.

Secondo la docente bisogna applicare esperimenti esplorativi per AI e spiegare che sono tecnologie sperimentali.

Un esperimento che si può fare è quello di usare un framework etico: l'AI rispetta le 16 condizioni della bioetica?

Tra i più importanti si ricordano:

- **Principio di Autonomia**: l'utente ha diritto di rifiutare il trattamento e di prendere parte al processo decisionale;
- **Principio di Beneficenza**: l'AI deve agire tutelando l'interesse dell'utente;
- **Principio di Non Maleficenza**: l'AI non deve causare danno all'utente;
- **Principio di Giustizia**: in caso di risorse limitate, i trattamenti devono essere distribuiti tra gli utenti in modo equo e giusto.

Dato che non tutti i problemi possono essere risolti si torna al mondo degli esperimenti esplorati che possono impattare con persone. Bisogna fare attenzione e cercare quoindi l'approvazione da parte delle autorità competenti perchè a volte si possono infrangere delle leggi durante le sperimentazioni.

Molti dei problemi però possono essere risolte *by design,* agendo quindi sulla strutturazione. In conclusione si può dedurre quindi che l'AI e la robotica autonoma sono per il momento tecnologie sperimentali.

Alcuni dicono che informatica non ha bisogno di sperimenti perché vicina alla matematica, ma in realtà che c'è bisogno si esperimenti per le AI in blackbox, perché l'esperimento è necessario quando c'è interazione con l'uomo.

2.1.4 Riccardo Guiddotti: *Explaining Explanation Methods*

Il ricercatore Guidotti successivamente ci ha introdotto all'explanation di una macchina AI. Le sue ricerche si incentrano sull'inserire un interprete che ci renda le decisioni di un AI chiare. Spiegare le decisioni di un sistema autonomo però non è sempre facile, perché spesso si incorre in bias che forviano il senso di una scelta. Ad esempio un classificatore di cani husky e di lupi, come opera? Si è scoperto che il classificatore riconosceva il lupo dalla neve sullo sfondo, e non per le sue caratteristiche fisiche. Altro esempio lampante è il COMPAS CORE, un sistema utilizzato per decidere se concedere o meno la libertà condizionata; le persone di colore sono considerati sempre criminali, e questo è un problema.

In Europa per fortuna esiste la legge GDPR che prevede il diritto alla spiegazione,

delle decisioni di un sistema AI.

Per rendere possibile questo è necessaria una conversazione tra l'uomo e la macchina.

Ma risultano esserci dei problemi: di interpretabilità: quanto un modello interpretabile è paragonabile a un black box? Qual è il livello di fidelity?

Modelli interpretabili come i decision tree, linear model, oppure regole booleane non hanno problemi ma per rendere più interpretabile un sistema blackbox una possibile soluzione è di rendere l'ultimo livello della scatola nera trasparente e leggibile nelle decisioni.

Guidotti ha terminato il suo discorso spiegando il perché è fondamentale studiare le blackbox. Per essere sicuri con il loro utilizzo, per migliorare standard industriali, aiutare persone a prendere una decisione corretta.
Gli sviluppi futuri saranno presto pesi parte dagli studi per dare una spiegazione
tra la macchina e l'uomo in modo tale che possano conversare, definendo bene una spiegazione.

2.1.5 Luciano Serafini: *Integration of Learning and Reasoning in Logic Tensor Networks*

Introdotto dal professor Vanni, Luciano Serafini ha mostrato come definire la distanza tra misure in un sistema di Machine Learning, cosa che manca alla logica.

Nella logica classica un predicato unario è un classificatore, ma il nostro classificatore
con il Machine Learning ci restituisce un valore in percentuale, non rendendo
mutualmente esclusivo un risultato.
Ad esempio supponiamo di essere ad una partita di tennis che si svolge tra maschi contro femmine. Abbiamo un classificatore che associa ai maschi valori di feautures vicini allo 0 e alle femmine valori vicini all'1. Ho cosi creato il mio dataset di partenza.
So che, per la loro semantica, i valori delle feautures si trovano tra 0 e 1.
Se utilizzo solamente tecniche di Machine Learning delle nuove persone che parteciperanno al torneo saranno classificate in percentuale, e quindi saranno poste, in un piano cartesiano, nel mezzo. Ma questo non è corretto perché una persona deve essere o maschio o femmina, in maniera mutualmente esclusiva.
Bisogna quindi aggiungere delle restrizioni su valori della semantica per limitare il risultato,
avendo una nuova feautures vicino ai maschi o vicino alle femmine.
Quindi per la fase di training va aggiunto un predicato universale, che indichi che per ogni nuova variabile introdotta, un maschio deve giocare necessariamente contro una femmina.
In conclusione possiamo dire che con logica ho una conoscenza strutturale del problema, ma con il Machine Learning ho conoscenza numerica.

L'obiettivo è quindi quello di unire in maniera coesa la Logica e il Machine Learning.

2.2. Pomeriggio

2.2.1 José M.Galvan: *Dilemma etico della libertà di espressione nei networked*

Ad aprire la fase pomeridiana è stato José M.Galvan che ha iniziato il suo discorso legato all'etica e agli aspetti *dark side* della tecnologia.

Secondo lui il mondo ICT sarà umano solo quando rispetterà completamente la persona umana. Questo suo pensiero si trova contrapposto a quello dello studioso Kurzweil che afferma che nel 2045 si arriverà alla *singolarità tecnologica*: la macchina supererà l'uomo con la robotica e la Strong Artificial Inteligence.

Se si prende in considerazione solo la capacità calcolo di una macchina, allora la singolarità tecnologica è già arrivata con l'avvento della calcolatrice. Infatti l'AI molto meno deludente della persona, se pensiamo solo alla capacità di calcolo.

La vera rivoluzione tecnologica arriverà quando saremo in grado di virtualizzare la coscienza umana in un'AI.

Dal punto di vista etico la macchina è tale quando riesce a dare un giudizio su cose concrete, un uomo è uomo quando ha delle relazioni con altri uomini.

Se riduco la ragione al semplice calcolo senza un linguaggio umano allora perdiamo di senso, basti pensare alla torre di Babele e alla relativa perdita del linguaggio come la punizione dovuta al ridurre la ragione al semplice calcolo nello spazio e nel tempo. L'uomo è molto di più. L'uomo può cambiare il suo linguaggio poichè la sua parola è data dalla sua libertà.

Tenendo conto di tutto questo, AI migliorerà l'uomo?

2.2.2 Paola Inverardi: *La Soft-Ethics ed il progetto EXOSOUL*

Sempre inerente al tema della libertà la professoressa Paola Inverardi ha spiegato che l'uomo è ogni giorno in relazione con enti sociali (anche non umani) e spesso non è in pari per poter esercitare la sua libertà di scelta.
Al giorno d'oggi l'autonomia decisionale dei sistemi AI è sempre più in crescita, ad esempio il cellulare ci governa senza che ne siamo consapevoli (pochi sanno che saltano le preferenze di privacy ad ogni aggiornamento di sistema!).
Ancora una volta l'uomo è al centro, deve essere un *Empower User*, farlo diventare attore attivo del sistema e non solo usato per prendere dei dati e farne delle statistiche.
Il team di ricerca che la docente segue pone l'accento sul poter mettere le proprie preferenze su qualsia sistema automatico attraverso un software personale.

Ad esempio se ci sono A e B macchine che competono per il posto e se B è il più vicino ma in A ci trova una donna incinta, B deve poter sapere che lei è incinta ed esercitare la sua libertà di lasciargli il posto per il mio libero arbitrio, anche se l'algoritmo originale avrebbe assegnato il posto a B.
Chi inserisce la decisione etica nel sistema AI? Chi stabilisce qual è il bene universale da seguire?
Un utente deve poter entrare nella mia macchina vista come un esoscheletro e poter cambiare il set di impostazioni di default (Hard Ethics) con le mie impostazioni (Soft Ethics).

Quindi i prossimi lavori e ricerche su questo ambito saranno mirati a sviluppare un software per ogni individuo dove ci sono proprie preferenze potenzialmente integrabile con qualsiasi sistema.

2.2.3 Piero Bonatti: *L'Etica è sostenibile?*

Al termine della giornata è intervenuto il professore Piero Bonatti, il quale ha posto in discussione la sostenibilità dell'etica in ambito della privacy e della sicurezza legate alle AI.
Privacy non è confidenzialità, che diventa inutile quando i dati di un utente devono essere condivisi, spiega infatti il professore che esiste un AI che grazie alla quale con i like su Facebook si possono delineare i profili psicologici degli utenti.
Ma non vogliamo che i nostri dati siano abusati, quindi bisogna mantenere confidenziali i dati, ma i costi della privacy sono molto alti, comprendendo i costi di usabilità e i costi computazionali.
L'anonimato è difficile da mantenere ma quali sono i parametri da usare per considerare un dato anonimo? Non è ben chiaro ancora come difendersi da un attacco informatico, bisognerebbe avere la *metaconoscenza* dell'attaccante ma questo è un problema di NP hard complessità.
È in corso il progetto SPECIAL per migliorare l'usabilità dei sistemi di privacy policy ad esempio con un consenso dinamico della gestione dei dati. Verrà fatta una piccola richiesta ogni volta che serve una nuova informazione, ma questo è un problema perché genererei un flusso troppo alto di richieste.

A conclusione possiamo dire che la privacy ha vari livelli di sicurezza e la confidenzialità dei dati ha costi elevati ed è difficile da gestire, ciò su cui possiamo lavorare è fare sistemi AI di *analitics* che preservano anonimato migliorando anche efficienza degli elaboratori.

3. Conclusioni

Durante questo seminario si è quindi avuto una larga panoramica e visione di insieme sui moderni e sempre più usati sistemi di Intelligenza Artificiale.

C'è molto lavoro da svolgere ancora per rendere questi sistemi parte integrante della nostra società, senza che essi creino problemi e soccombano l'uomo assoggettandolo, ma in questo nuovo *Umanesimo* che la nostra società sta vivendo, è necessario mantenere al centro l'uomo senza che esso perda il controllo.

Grazie ai contributi di analisi etico-giuridica e di ricerca in IA ricevuti vediamo come è fondamentale favorire il dialogo e l'interazione produttiva tra ricercatori di ogni area accomunati dall'obiettivo di capire e affrontare le nuove sfide etiche per l'IA.

L'Intelligenza Artificiale va ottimizzata ancora per poter essere applicata in ambiti di comunicabilità e sicurezza, agendo molto sulla trasparenza, autonomia e grado di infallibilità. Sono stati esposti anche problemi di tipo etico-giuridici, come privacy o la responsabilità di alcune scelte, non sempre conformi al GDPR, spesso citato durante i vari interventi.

Parafrasando Turing, possiamo vedere solo una piccola distanza davanti a noi, ma c'è molto ancora su cui lavorare.