Lo stato della Machine Ethics applicata a Sistemi a Guida Autonoma: una review sistematica

Alberto Giunta[1]

Abstract— **Insieme alla diffusione sempre più capillare di veicoli autonomi e semi-autonomi sulle strade, cresce anche la necessità di capire più a fondo le capacità di interazione che questi veicoli hanno nella società. Sebbene statisticamente più sicuri di veicoli guidati da esseri umani, rimangono sistemi fallaci, che saranno causa o rimarranno vittime di incidenti a causa di comportamenti a loro dire corretti, ma che magari difficilmente si allineano con comportamenti morali ed etici raffinati nei millenni dalla specie umana. In questo studio si proverà a definire una panoramica su quelli che sono i principali problemi etici nello sviluppo di AV, quali sono i principali framework di pensiero con cui il problema viene affrontato attualmente, e quali sono i plausibili sviluppi e roadmap futuri.**

I. Introduzione

Tra il 2016 e il 2018 sono state registrate almeno quattro vittime di incidenti stradali in cui è stato verificato il coinvolgimento di veicoli con sistemi a guida autonoma [1], [2], [3], [4], [7], [8]. Questi eventi hanno ogni volta immancabilmente generato grandi discussioni di carattere etico, morale e legale e hanno rappresentato il seme per importanti spunti di riflessione riguardo alle responsabilità degli ingegneri che progettano questi veicoli, di coloro che ne sono alla guida e dei legislatori che devono regolarne gli errori e i malfunzionamenti.

Già nel 2016 si sono raggiunte le centinaia di milioni i kilometri di strada percorsi da automobili guidate in maniera semi-autonoma [5], ovvero con un guidatore in teoria sempre pronto a prendere il controllo dell'automobile in caso di malfunzionamenti. Inoltre, uno studio della NHTSA ha evidenziato tra le altre cose, una riduzione degli incidenti del 40% nelle auto Tesla, prima e dopo l'introduzione della funzionalità *Autopilot* [9].

Ogni incidente stradale avvenuto mentre la funzionalità di guida semi-autonoma era attiva ha rappresentato inoltre un importante crocevia in cui non solo le aziende potevano prendere atto dell'accadimento, degli eventuali errori da loro commessi e dell'uso più o meno corretto dei loro prodotti, ma è stato e sarà fondamentale per capire come questi sistemi, agenti già potenzialmente autonomi dal punto di vista tecnologico, si comportino nel mondo reale. Il loro comportamento in situazioni reali - spesso complesse più di qualsiasi possibile simulazione - è fondamentale anche per ridurre il Frame Problem, ovvero la tendenza umana di attribuire significati non effettivamente reali o corretti a determinati comportamenti di agenti o robot [6].

[1]Alberto Giunta, Department of Computer Science and Engineering

Un aspetto fondamentale nello sviluppo di questi sistemi risiede nell'introduzione della capacità di prendere decisioni, da cui l'autonomia che li caratterizza: ad esempio, usando in maniera estensiva la sensoristica di cui dispongono possono decidere e modificare la migliore velocità di crociera in ogni dato momento, riescono a rimanere nella carreggiata appropriata anche su strade urbane con una pavimentazione e linee di demarcazione non regolari, e possono superare altri veicoli o frenare per evitare ostacoli.

Fino a questo momento le maggiori compagnie produttrici (Tesla, Uber, Google per elencarne alcune) sono riuscite a mantenersi nella legalità riversando la responsabilità delle decisioni prese dai loro sistemi sul conducente, che in tutti i momenti deve essere vigile e pronto a prendere il controllo del veicolo, prevalendo sulle decisioni del sistema autonomo. Un esempio è Tesla, che prima che venga attivata la funzionalità di *Autopilot* richiede al conducente di prendere coscienza del fatto che ogni azione rimarrà sotto la sua responsabilità [7].

È apparentemente inevitabile però che lo sviluppo di queste tecnologie, che secondo alcuni ridurrà il numero di incidenti stradali del 90% (Gogoll et al. [14]), stia andando verso un'autonomia sempre più completa e una presenza sul mercato sempre più ampia, rendendo però sempre più probabile che questi veicoli saranno fatalmente causa di incidenti a loro volta. Non ci sono mai state tecnologie infallibili, specie nei loro primi anni di utilizzo, e gli AV non sono un'eccezione, specialmente perché se un umano non sarà in grado di prendere il controllo del veicolo in tempo, i sistemi saranno responsabili del cosiddetto comportamento di *precrash*. Tutto ciò richiederà necessariamente la definizione di policy etiche e morali che saranno alla base delle decisioni prese da questi veicoli una volta che si troveranno nel mondo reale e dovranno avere a che fare con gli esseri umani.

Lo scopo di questo studio è quindi quello di effettuare una raccolta dei principali argomenti di discussione nell'ambito di *Machine Ethics* applicata ai sistemi a guida autonoma. Oltre a questo, ci si è posti anche l'obiettivo di verificare quali possano essere delle possibili soluzioni o framework di pensiero per progettare agenti autonomi che dovranno forzatamente compiere delle scelte che, se compiute da umani, definiremmo *etiche e morali*. Nella sezione 2 verrà descritta la metodologia con cui è stato svolto questo studio, e nella sezione 3 ne verranno presentati i risultati e verrà data risposta alle principali domande di ricerca che erano state formulate. Nella sezione 4 verranno infine stilate le conclusioni di questo studio.

II. Metodo

È stato applicato il metodo di SLR base come descritto da Kitchenham [12], e qui di seguito verranno forniti ulteriori dettagli sul metodo di ricerca, review e selezione degli studi trovati.

A. Domande di ricerca

Le domande di ricerca fondamentali di questo studio sono:

RQ1. Quali sono i maggiori e principali problemi di Machine Ethics applicata ai sistemi a guida autonoma, anche detti AV (Autonomous Vehicles)?

RQ2. Quali sono attualmente i principali framework di design, architettura, pensiero nella progettazione e legislazione di sistemi a guida autonoma?

RQ3. In che direzione sta andando il mondo della ricerca e della legislazione riguardo a questo argomento? È già stata delineata una roadmap per studi futuri?

B. Processo di ricerca

Il processo di ricerca è consistito in una ricerca manuale sui principali aggregatori di letteratura scientifica:

- Google Scholar
- Scopus

La ricerca e raccolta degli studi sono state effettuate il giorno 15/06/2018. Per effettuare la ricerca è stata raffinata ed utilizzata la query seguente:

- Google Scholar: `"machine ethics" AND ("road vehicle automation" OR driverless OR crash OR vehicle? OR car? OR drive OR autonomous OR automated)`
- Scopus: `TITLE-ABS-KEY ("machine ethics" AND ("road vehicle automation" OR driverless OR crash OR vehicle? OR car? OR drive OR autonomous OR automated)) AND (LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014))`

Un altro criterio di ricerca e selezione è stato l'anno di pubblicazione dei paper. Si è scelto di non andare più indietro del 2014 per il semplice fatto che il topic di questa SLR è piuttosto specifico e relativamente recente. Gli AV sono diventati un argomento relativamente mainstream sono negli ultimi anni, con la diffusione di auto come quelle prodotte da Tesla o Uber, o le più prototipali auto di Google, Apple, Mercedes, BMW ecc. La maggiore presenza sulle strade di queste auto ha aiutato ad evidenziare e dare il meritato peso a problemi che prima non avevano la stessa rilevanza. Per questo motivo è stato quindi considerato un intervallo di tempo che va dal 2014 (compreso) a Giugno 2018.

Relativamente al tipo di studi che sono stati considerati, si è optato per includere Conference Proceeding e Journal Paper, tralasciando altre eventuali SLR. Le riviste e conferenze che sono state selezionate sono le seguenti:

- Springer
- Science
- American Journal of Public Health (AJPH)
- arXiv
- Transportation Research Board (TRB)
- Nature
- Association for the Advancement of Artificial Intelligence (AAAI)
- PhilPapers
- White Rose
- EmeraldInsight

C. Criteri di inclusione ed esclusione

Sono stati esclusi innanzitutto gli studi con un numero di citazioni non sufficiente. Seppure conscio del fatto che le citazioni siano una metrica ambigua nella definizione della qualità di un paper di ricerca, si è preferito evitare l'inclusione di studi con meno di 5 citazioni per far sì che fossero inclusi invece solo studi che erano stati validati, anche in minima misura, dalla comunità scientifica.

Sono inoltre stati esclusi tutti gli studi che non rispondevano o non erano inerenti alle domande di ricerca alla base di questa SLR.

D. Raccolta dei dati

I dati estratti da ciascun studio sono stati:

- Il motore di ricerca che li ha indicizzati.
- La fonte (rivista o conferenza) su cui sono stati pubblicati.
- L'autore o gli autori
- La data di pubblicazione (con giorno e mese se disponibili).
- I problemi evidenziati relativamente all'argomento di discussione (Machine Ethics e AV)
- Le possibili soluzioni trovate o pensate, dove disponibili.

III. Risultati e Discussione

L'obiettivo di questa review sistematica è quello di delineare lo stato dell'arte riguardo all'implementazione di Machine Ethics nei veicoli a guida autonoma. Ciò che è emerso è un'area di studio ancora in stato embrionale, che sta solamente in questi ultimi anni vedendo una crescita di interesse generale, data l'importanza del tema. I veicoli autonomi (semi-autonomi per ora) stanno diventando sempre più presenti sulle strade di tanti paesi, ed è inevitabile il loro sempre più alto coinvolgimento in situazioni che potrebbero portare a incidenti stradali, ed in cui si troveranno a dover reagire in maniera appropriata al codice della strada e a leggi morali ed etiche. Sia i produttori di AV che le Istituzioni dovranno affrontare situazioni e dilemmi spinosi, e mai realmente affrontati prima. La maggior parte degli studiosi afferma che non ci si debba fermare ad una visione

semplicistica delle questioni etiche, come quella offerta da esercizi di pensiero come il Trolley Problem. Bensì, la questione andrebbe affrontata da un punto di vista ben più approfondito e generale. Vengono proposte diverse filosofie morali come Utilitarismo e Consequenzialismo, e diversi framework di design per lo sviluppo di questi agenti. La visione meglio costruita e definita è probabilmente quella espressa da Goodall [17]. Per il resto, l'unica certezza, come spesso è accade, è che non vi sono certezze: volendo essere realisti non ci potranno essere reali passi avanti finché l'argomento non diventerà problematico per almeno una delle parti coinvolte: i produttori, i membri della società, e le Istituzioni.

Nell'appendice all'ultima pagina di questo studio sono riportati i risultati congiunti della ricerca sui due diversi motori di ricerca: sono stati in tutto esaminati 25 studi tra il 2014 e il 2018, pubblicati da 11 diverse fonti. Di questi, 12 corrispondevano positivamente ai criteri di qualità e alle domande di ricerca precedentemente elencate, e di seguito verranno aggregati i punti di maggiore valore espressi nei vari studi per ciascuna domanda di ricerca.

A. RQ1. Quali sono i maggiori e principali problemi di Machine Ethics applicata ai sistemi a guida autonoma, anche detti AV (Autonomous Vehicles)?

Bonnefon et al. [11] definisce come il problema più controverso creato dagli AV, la necessità di decidere *come* un veicolo debba scegliere se sacrificare o meno il proprio passeggero per salvare diverse altre vite sulla strada.

Secondo gli autori, quelli che vengono chiamati *algoritmi morali* dovranno raggiungere tre obiettivi principali: (1) essere consistenti, (2) non causare pubblico sdegno, (3) non scoraggiare gli acquirenti. Mentre le motivazioni alla base di (1) e (2) sono piuttosto intuitive, la motivazione dietro l'affermazione (3) consiste nel fatto che i vantaggi degli AV sono innumerevoli, e creare delle policy di sicurezza svantaggiose per gli acquirenti risulterebbe in una mancata diminuzione di incidenti stradali, inquinamento e quanto di più vantaggioso possono offrire gli AV.

Menon et al. [16] dà una definizione più formale di quelli che secondo gli autori sono i maggiori problemi degli AV:

- **Accettabilità del rischio**: non è chiaro se l'opinione pubblica sia pronta ad accettare lo stesso rischio quando questo è presentato da una macchina, rispetto a quando lo stesso rischio è presentato da un umano.
- **Trasferimento del rischio**: se il rischio viene trasferito da alcune persone ad altre, allora il trasferimento deve essere esplicitamente motivato, anche quando il rischio generale viene ridotto grazie al trasferimento.
- **Consenso**: le parti esposte al rischio devono tutte accettare e prendere coscienza della porzione di rischio che gli viene "allocata".
- **Visione degli AV come dei SoS (Systems of Systems)**: la potenziale intersezione con altri agenti deve essere presa in considerazione, così come l'interazione con l'infrastruttura, operatori umani o terze parti. Secondo gli autori, specialmente dove gli agenti fanno uso di algoritmi di machine learning è maggiore la probabilità di assistere a interazioni mai viste prima e comportamenti emergenti.

In Bonnefon et al. [10] e [11] vengono evidenziati tre tra i più classici dilemmi etici che gli studiosi si sono trovati ad affrontare nell'applicazione di Machine Ethics agli AV. Nell'esempio che viene riportato si è messi di fronte a tre possibili situazioni stradali che portano inequivocabilmente ad un incidente. Un'automobile deve decidere se:

- uccidere diversi pedoni per salvarne uno, o viceversa.
- uccidere un pedone per salvare il proprio passeggero, o viceversa.
- uccidere diversi pedoni per salvare il proprio passeggero, o viceversa.

Questo esempio viene non a caso utilizzato: è ispirato al Trolley Problem, un esercizio di pensiero a cui molto si fa riferimento in ambito di etica e morale. Più in generale, in questo studio viene approcciato il problema dal punto di vista della teoria etica dell'Utilitarismo, il cui enunciato è assimilabile al fatto che potendo scegliere, è sempre meglio scegliere l'opzione che massimizza una funzione di utilità definita a priori. Una visione Utilitaristica nel dominio di questo studio significa quindi implementare un meccanismo di scelta che apporta sempre il danno minore agli esseri umani da parte dei veicoli.

Bonnefon et al. [10] suggeriscono però che nella realtà, gli *"algoritmi morali"* dovranno risolvere problemi ben più complessi di semplici rivisitazioni del Trolley Problem, tra cui ad esempio situazioni in cui un AV deve scegliere tra diverse opzioni tenendo in considerazione anche le probabilità di sopravvivenza (è più vantaggioso schiantarsi per l'AV rispetto ad investire un motociclista?) oppure le età delle persone coinvolte (è più corretto salvare la vita di un adolescente o di tre cinquantenni?).

Holstein et al. [18], insieme a Nyholm et al. [19] e Etzioni et al. [21] sono concordi nell'affermare che il Trolley Problem sia un problema posto in maniera ingannevole, e non rappresenti un buon punto di partenza per la creazione di algoritmi di "*crash*" per AV. Etzioni et al. [21] lo ritengono un buon strumento di insegnamento e di introduzione all'etica e alla morale, ma nulla più. Secondo gli autori, ci si dovrebbe concentrare su soluzioni pratiche e sulle loro conseguenze *sociali* piuttosto che su problemi idealizzati e irrisolvibili come il tanto discusso Trolley Problem, in quanto problemi come quest'ultimo hanno come unica soluzione il fatto di non avere soluzioni valide in ogni circostanza, e offuscano quelli che sono i veri dilemmi etici che andrebbero davvero affrontati. Andrebbe inoltre discussa la linea di demarcazione che separa ciò che è tecnicamente possibile da ciò che è eticamente giustificabile, sotto l'aspetto economico, tecnico, di processo e organizzazione. Come si può leggere in Etzioni et al. [21], ciò che secondo gli autori è più sbagliato del Trolley Problem è che vengono lasciate al protagonista due sole scelte moralmente ripugnanti, senza un modo di aiutare realmente a risolvere la situazione, rendendo evidente che

questo quesito ha come premessa una visione altamente impoverita della natura umana.

Anche in Gogoll et al. [14] viene ritenuto inadeguato l'utilizzo del Trolley Problem per risolvere dilemmi etici come quello in oggetto, in quanto viene ritenuto una situazione troppo semplicistica e non capace di inquadrare tre aspetti strutturali fondamentali dei vari Traffic Dilemma (come il Tunnel Dilemma):

- **Interazione Strategica** ed **Iterazione**: nel Trolley Problem è unicamente la nostra azione a determinare l'esito della situazione, al contrario di ciò che succede nel mondo reale, dove ci sono azioni e reazioni, e dove una decisione può essere influenzata da ciò che sappiamo delle policy etiche e morali degli altri elementi della società.
- **Posizione variabile di ogni individuo**: ognuno di noi può essere toccato in maniera attiva o passiva dalle scelte degli AV, a seconda che il nostro veicolo debba prendere una decisione o subisca la decisione presa da altri veicoli. Al contrario, questo nel Trolley Problem non accade, in quanto l'unica possibilità di scelta è data al protagonista del problema, colui che ha in mano la leva del cambio.

Holstein et al. [18] hanno inoltre stilato una lista strutturata di ciò che ritengono essere le maggiori sfide sul lato tecnico e le raccomandazioni per una loro risoluzione ottimale, raggruppate per requisiti. In termini più generali, esprimono la necessità di tenere in considerazione gli aspetti etici in ogni fase dello sviluppo del software, dalla definizione dei requirement al testing, alla manutenzione ed evoluzione.

B. RQ2. Quali sono attualmente i principali framework di design, architettura, pensiero nella progettazione e legislazione di sistemi a guida autonoma?

In Malle [13] viene formulata un'idea iniziale di framework per sviluppare agenti autonomi morali, dove l'intenzione è quella di usare quella che viene chiamata *competenza morale* per risolvere non tutte, ma almeno una parte delle preoccupazioni di tipo etico della società relativamente ai robot.

Al posto di intendere un agente *morale* come un'entità che agisce in accordo a cosa è giusto e sbagliato, in Malle [13] viene tentato un diverso approccio, ovvero si cercano di identificare le capacità che sono alla base della competenza morale umana per poi usarle come mattoni fondamentali di un framework di sviluppo futuro.

I seguenti quindi sono gli elementi chiave di competenza morale che dovrebbero essere presenti in un eventuale agente autonomo e morale:

- Vocabolario morale
- Sistema di norme
- Cognizione e influenza morale
- Processo decisionale morale e azioni morali
- Comunicazione morale

Viene avanzata inoltre una discussione riguardo ai dilemmi morali, che secondo gli autori richiederebbero una scelta genuina tra opzioni imperfette, e molto spesso ciascuna delle opzioni può essere moralmente giustificata riferendosi a delle specifiche norme. Guardandoli da prospettive diverse quindi, le azioni svolte dai robot potrebbero essere contemporaneamente moralmente giuste o sbagliate. Resterebbe però il fatto che i robot, un'azione, la compirebbero.

In Munean et al. [15] si avanza l'idea di un modello per un Artificial Autonomous Moral Agent (AAMA). Esso sarebbe in grado di imparare e sviluppare una propria *moral competenecy* tramite l'utilizzo di una "popolazione" di reti neurali e di un dataset di *dati morali* che descrivano i tipici comportamenti umani. In questo modo riuscirebbe a classificare delle azioni come moralmente giuste, neutre o sbagliate in maniera consistente con il suo training set. La moralità sarebbe una caratteristica emergente di questo sistema, e non una sua funzionalità fondamentale. In altri termini, essa sarebbe una feature statistica di una popolazione di reti neurali, più che la proprietà di una sola rete.

Leben et al. [20] fornisce un'alternativa alla linea di pensiero Utilitaristica. L'autore suggerisce l'utilizzo della teoria Rawlsiana per far sì che un AV possa stimare la probabilità di sopravvivenza per ogni persona e azione, scegliendo poi l'azione in linea con le scelte che farebbe una persona che agisce nei propri interessi. L'algoritmo che viene proposto ha dei risultati secondo gli autori ben diversi (e migliori) da quelli ottenuti tramite l'utilizzo di teorie Utilitaristiche. Il più grande vantaggio di un algoritmo Rawlsiano sarebbe il suo trattamento eguale di ogni individuo, e la sua non predisposizione a sacrificare gli interessi di uno per gli interessi di altri. Gli autori specificano inoltre che in questo studio i veicoli agirebbero in maniera isolata rispetto agli altri veicoli, ma la ritiene una semplificazione, in quanto nel mondo reale i produttori di AV sarebbero in grado di interconnettere una flotta di veicoli facendoli comunicare e aprendo orizzonti a nuove modalità di decisione ed elaborazione delle informazioni.

In Millar et al. [22] viene affrontato inoltre il problema dell'inevitabile influenza e imprinting che i designer e sviluppatori del software hanno sulle loro creazioni, e quando le loro creazioni sono algoritmi che devono eseguire azioni ritenute *etiche*, la questione diventa rapidamente problematica. Una soluzione che viene discussa viene dalla definizione di cosiddetti *moral proxy*, ovvero tecnologie che agiscono su mandato di umani e che forniscono risposte materiali a domande morali. Quando un dispositivo reifica una decisione morale istanzia una relazione in quanto moral proxy. I moral proxy, insieme a pratiche di consenso informato provvedono un framework concettuale che può essere applicato al processo di design del software per evitare del paternalismo accidentale. Gli autori riconoscono però anche la presenza di un gap notevole tra la teoria proposta e le tecnologie disponibili per metterla in pratica.

Etzioni et al. [21] fornisce una comparazione tra i due principali punti di vista con cui si può affrontare il discorso di agenti autonomi morali:

- **Top-down**: i principi etici sono programmati nel sistema di guida dei veicoli. Questi principi potrebbero

essere le Tre Leggi della Robotica di Asimov, i Dieci Comandamenti o qualsiasi altra filosofia morale, come Utilitarismo, Consequenzialismo ecc. I veicoli sarebbero quindi in grado di prendere scelte etiche basandosi sui principi morali che sono stati impiantati al loro interno. Una critica a questa visione è sicuramente la difficoltà intrinseca nello sviluppo di principi morali che rimarrebbero validi e sarebbero robusti ad ogni possibile situazione. È infatti fondamentalmente impossibile che non si riesca a trovare una situazione che causi il fallimento e la contraddizione in termini di questi precetti.

- **Bottom-up**: in questo approccio gli agenti sono tenuti a imparare quelle che possono essere scelte etiche osservando il comportamento umano in situazioni reali, senza che gli venga impartita nessuna regola formale o filosofia morale. Per ora questo approccio apprendimento non supervisionato è stato utilizzato per gli altri aspetti *non etici* di funzionamento degli AV.

C. RQ3. In che direzione sta andando il mondo della ricerca e della legislazione riguardo a questo argomento? È già stata delineata una roadmap per studi futuri?

Lo studio effettuato da Bonnefon et al. [10] e [11] suggerisce che una regolazione degli AV possa essere necessaria ma anche controproducente. È stato verificato infatti come l'implementazione di comportamenti Utilitaristici nei modelli di scelta di veicoli a guida autonoma porti al crearsi di quello che è a tutti gli effetti un dilemma sociale: quando intervistata, la maggior parte delle persone ha infatti detto di volere e apprezzare questo tipo di etica all'interno dei veicoli, ma anche detto che non li comprerebbe mai per sé: "Nessuno vuole una macchina che si interessi ad un bene superiore. Le persone vogliono una macchina che si interessi a loro stesse." (Etzioni et al. [21]).

Bonnefon et al. [10] suggeriscono addirittura che una soluzione iniziale al problema sarebbe quella di evitare completamente una legislazione che si occupi di questi argomenti, in modo da non frenare lo sviluppo e quindi non impedire la messa in commercio di veicoli autonomi, che sono pur sempre più sicuri di veicoli guidati da umani.

Anche secondo la riflessione fatta in Holstein et al. [18] i sistemi legislativi attuali non sono minimamente pronti e adattabili agli AV, e secondo l'autore sarà compito degli enti di supervisione nazionali e internazionali accelerare la discussione e la raccolta di dati a supporto di future legislazioni.

Menon et al. [16] pensano, al contrario di altri prima citati, che la natura commerciale degli AV e una loro mancata regolazione, possano portare a più danni del previsto, in quanto le compagnie produttrici sarebbero per natura poco (se non per nulla) stimolate a condividere con la concorrenza e le Istituzioni ciò che per loro potrebbe essere un vantaggio competitivo. Lo stesso accadrebbe con i problemi che vengono riscontrati nei loro prodotti: senza obbligo alcuno da parte delle Istituzioni, non ci sarebbe alcun incentivo alla condivisione e al progresso comune, ai danni di ciascun membro della società.

Etzioni et al. [21] afferma però, che come si dice in campo giuridico: "i casi estremi creano cattive leggi", e lo stesso vale nel campo dell'etica. I casi limite ed eccezionali su cui si focalizza l'attenzione e la discussione dell'opinione pubblica hanno come risultato la creazione di leggi che risolvono il problema specifico nell'immediato, ma costituiscono regolamenti poco generali e scalabili nel tempo.

In Gogoll et al. [14] viene fatto notare come nelle società moderne e più liberali, gli scontri e i disaccordi morali vengano risolti dando la possibilità a ciascun individuo di scegliere secondo i propri i principi e in autonomia (entro certi limiti, ovviamente). Questa modalità equivarrebbe quindi a quelle che in questo studio vengono definite Personal Ethics Settings, o PES. Purtroppo però, dare la possibilità a ciascuno di scegliere le proprie PES nel dominio degli AV equivarrebbe a ricreare l'effetto sociale descritto dal Dilemma del Prigioniero: ogni individuo tenderebbe a scegliere in maniera egoistica ciò che ritiene essere meglio per sé, portando però ad una peggiore qualità della vita per tutti i membri della società, egli compreso. Gli autori deducono quindi che sia più pratico ed efficace che siano le Istituzioni ad istituire una legislazione coercitiva ed eguale per ogni membro della società, in modo da minimizzare i pericoli e i danni agli essere umani nell'insieme.

Bonnefon et al. [11] credono che presto ci sarà il bisogno impellente da parte di produttori e legislatori di trovare una risposta alle numerose domande di carattere etico e morale che questo dominio naturalmente presenta. In questo potranno essere aiutati dagli studi effettuati via via negli anni, tenendo in considerazione però che l'opinione pubblica rimane l'opinione di non esperti e con una visione d'insieme che potrebbe non essere la migliore, se non fuorviante. Gli sforzi del futuro, dicono, andranno incanalati verso il testing di scenari che introducano il concetto di rischio atteso, valore atteso, e di assegnazione di colpa e responsabilità.

Malle [13] denota la necessità di parlare in termini delle competenze che agenti autonomi, come gli AV, dovrebbero avere nella nostra società. Ad esempio: *I robot dovrebbero sempre obbedire ai comandi umani*? Oppure: *Dovremmo permettere ai robot di uccidere degli esseri umani?*

Goodall [17] propone un iter di sviluppo per veicoli *etici* diviso in tre fasi distinte:

- **Fase 1: Rational Ethics**: un sistema *razionale* di regole per AV, che premi i comportamenti che minimizzano il danno globale. Data la difficoltà di predire il comportamento e le regole corrette per ogni possibile situazione, dove ci sia conflitto o si sia scoperti da regole predefinite, l'autore suggerisce ad esempio che il veicolo si fermi o evada dalla situazione.
- **Fase 2: Hybrid Rational and Artificial Intelligente Approach**: richiede software ancora non disponibile, che sia in grado di usare tecniche di Machine Learning per capire la decisione etica corretta data una situazione di pericolo, aderendo anche a ciò che era stato definito nella Fase 1. Premettendo la necessità di impostare

dei limiti iniziali, un approccio incrementale di auto-training permetterebbe lo sviluppo di sistemi *morali* sempre più avanzati anche in assenza di un sistema morale perfettamente definito.

- **Fase 3: Feedback with Natural Language**: le reti neurali, lo strumento che andrebbe comprensibilmente utilizzato in questa teoria, hanno ancora il grave problema di essere difficili da "decifrare" per gli esseri umani. Rimaniamo all'oscuro dei pattern che hanno portato ad un output di un valore al posto di un altro. Per questo l'autore fa notare la necessità di fornirsi di sistemi per estrarre "spiegazioni" dalle reti neurali che risultino comprensibili anche all'umano. Come caso d'uso immediato, si pensi ai momenti successivi ad un incidente, in cui si cerca di capire quali azioni hanno portato a quell'avvenimento e perché, e questa spiegazione viene fornita in linguaggio naturale.

D. Limitazioni di questo studio

Questa SLR ha sofferto di alcune limitazioni, tra le quali:

- Scarsità di risorse mancata eterogeneità di opinioni: questa SLR è stata interamente realizzata da una sola persona, e questo potrebbe aver portato alla presenza di bias nella fase di inclusione ed esclusione dei paper. Se ci fossero state almeno altre due persone a lavorare a questo studio, si sarebbe potuto dividere il lavoro in modo da avere sempre un mindset libero dai pregiudizi o preferenze del singolo.
- Ricerca limitata a pochi aggregatori: la ricerca è stata effettuata in maniera manuale e su due soli aggregatori di studi scientifici: Scholar e Scopus. Al fine di ottenere uno spettro più ampio di studi andrebbe esteso lo scope di questo studio effettuando ricerche anche su altri motori di aggregazione o siti di riviste particolari, possibilmente raffinando ancor più anche la query ed eventualmente le domande di ricerca.
- Tema ancora in definizione ed espansione: il tema principale di questa SLR è piuttosto specifico, ed ancora in larga parte solamente teorico. Ciò che è possibile reperire per ora sono solamente studi teorici con rare applicazioni in campo pratico. Mancano quindi ogni sorta di studi che contengano misurazioni, comparazioni e analisi di dati e informazioni estratte da esperimenti sul campo o simulazioni.

IV. Conclusioni

In questo studio è stata delineata una panoramica di come il mondo della ricerca ha finora affrontato l'applicazione di Machine Ethics al mondo degli Autonomous Vehicles. Ciò che è emerso è che questa è un'area di studio ancora in grande espansione, che vedrà probabilmente nel futuro prossimo un largo aumento di partecipazione e interesse. Mano a mano che aziende come Tesla, Google, Apple, Mercedes, BMW ecc. si interessano sempre più allo sviluppo di veicoli a guida prima semi-autonoma, poi autonoma, emergeranno anche i principali problemi legati alla presenza di questi veicoli sulle strade di tanti paesi e città del mondo. Prima di tutto dovranno essere preferiti ed acquistati al posto di veicoli non autonomi; ciò significa che dovranno fornire un livello di sicurezza ed affidabilità superiore a qualsiasi altro veicolo (non-autonomo e di pari prezzo) al momento sul mercato. Il loro comportamento su strada dovrà inoltre rispettare non solo il codice della strada, ma anche leggi morali ed etiche che ciascun essere umano si aspetterebbe da chiunque altro. Le Istituzioni dovranno inoltre provvedere a fare il possibile per incentivare l'uso di questi veicoli, e allo stesso tempo tracciare una marcata linea di demarcazione tra ciò che è ritenibile un comportamento corretto, e ciò che invece costituisce un malfunzionamento. Specialmente in riferimento all'assegnazione di responsabilità, dovrà essere reso chiaro e il più possibile trasparente cosa è responsabilità di chi, ed in quali circostanze. Gli studiosi hanno per ora individuato alcune linee di pensiero su come sarebbe più appropriato sviluppare l'intelligenza dietro questi veicoli e la loro capacità di scelta. Ciò che è comune a quasi tutti gli studi esaminati è il rifiuto di esercizi di pensiero come il Trolley Problem per esprimere concetti estremamente più complessi e profondi, non riducibili ad una semplice scelta binaria.

References

[1] https://www.mercurynews.com/2018/03/30/tesla-autopilot-was-on-during-deadly-mountain-view-crash/
[2] https://qz.com/783009/the-scary-similarities-between-teslas-tsla-deadly-autopilot-crashes/
[3] https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk
[4] https://news.vice.com/article/kzxq3y/self-driving-uber-killed-a-pedestrian-as-human-safety-driver-watched
[5] https://www.theverge.com/2016/5/24/11761098/tesla-autopilot-self-driving-cars-100-million-miles
[6] https://plato.stanford.edu/entries/frame-problem/
[7] https://www.tesla.com/blog/tragic-loss
[8] https://www.tesla.com/blog/update-last-week's-accident
[9] https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF
[10] J. F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles, June 24, 2016
[11] J. F. Bonnefon, A. Shariff, I. Rahwan, Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?, October 13, 2015
[12] B. Kitchenham, Procedures for Peforming Systematic Reviews, July, 2004
[13] B. F. Malle, Integrating robot ethics and machine morality: the study and design of moral competence in robots, July, 2015
[14] J. Gogoll, J. F. Müller, Autonomous Cars: In Favor of a Mandatory Ethics Setting, July 14, 2016
[15] I. Munean, D. Howard, A minimalist model of the artificial autonomous moral agent (AAMA), 2016
[16] C. Menon, A. Menon, R. David, Ethics and the safety of autonomous systems, 2018
[17] N. Goodall, Ethical Decision Making During Automated Vehicle Crashes, 2014
[18] T. Holstein, G. Dodig-Crnkovic, P. Pelliccione, Ethical and Social Aspects of Self-Driving Cars, February 5, 2018
[19] S. Nyholm, The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?, November, 2016
[20] D. Leben, A Rawlsian algorithm for autonomous vehicles, March 4, 2017
[21] A. Etzioni, O. Etzioni, Incorporating Ethics into Artificial Intelligence, December 1, 2017
[22] J. Millar, Technology as moral proxy: Autonomy and paternalism by design, September 8, 2014

Titolo	Autori	Fonte	Rivista	Data	Motivo esclusione	Problemi evidenziati	Soluzioni trovate
Machine Ethics and Automated Vehicles	Goodall N.J. (2014) Machine Ethics and Automated Vehicles. In: Meyer G., Beiker S. (eds) Road Vehicle Automation. Lecture Notes in Mobility. Springer, Cham	Scholar	Springer	01/06/2014	Non affronta la questione in maniera completa ed esaustiva abbastanza.		
The social dilemma of autonomous vehicles	**Jean-Francois Bonnefon, Azim Shariff, Iyad Rahwan**	**Scholar**	**Science**	**24/06/2016**		**Vengono poste una serie di domande, viene analizzato il fenomeno degli AV dal punto di vista sociale (tramite sondaggi)**	**Non vengono fornite soluzioni vere e proprie poichè sono solo state analizzate le risposte ad alcune domande dal punto di vista sociale**
Public Health, Ethics, and Autonomous Vehicles	Janet Fleetwood	Scholar	AJPH	01/04/2017	Analizza il problema da un punto di vista di regolamentazioni e leggi da parte dei governi		
Are We Ready for Utilitarian Cars?"	**Jean-Francois Bonnefon, Azim Shariff, Iyad Rahwan**	**Scholar**	**arXiv**	**13/10/2015**		**Analizza i risultati di 3 sondaggi e li discute, fornendo ulteriori domande a cui si dovrà dare risposta**	
Investigation into the Role of Rational Ethics in Crashes of Automated Vehicles	Wesley Kumfer, Richard Burgess	Scholar	TRB		Non accessibile		
Integrating robot ethics and machine morality: the study and design of moral competence in robots	**Bertram F. Malle**	**Scholar**	**Springer**	**01/07/2015**		**Analizza il problema da un punto di vista più ampio rispetto solo agli AV, però solleva questioni di design "morale"**	**Fornisce alcuni spunti per inserire dei tratti "umani" all'interno della moralità dei robot**
Evolutionary Machine Ethics	Han, T. A., Pereira, L. M., (2018) 'Evolutionary Machine Ethics' in Bendel, O. Handbuch Maschinenethik. Springer	Scholar	Springer	01/01/2018	Non accessibile		
Logical Limitations to Machine Ethics with Consequences to Lethal Autonomous Weapons	Matthias Englert, Sandra Siebert, Martin Ziegler	Scholar	arXiv	11/11/2014	Affronta la questione dal punto di vista della armi autonome e non degli AV. Fornisce comunque spunti interessanti.		
Autonomous Cars: In Favor of a Mandatory Ethics Setting	**Jan Gogoll, Julian F. Mvºller**	**Scholar**	**Springer**	**14/07/2016**		**Presenta due modi per avere a che fare col trolley problem**	**Presenta delle possibili soluzioni nell'interpretazioni di questi dilemmi morali negli AV**
The Robot's Dilemma	Boer Deng	Scholar	Nature	01/01/2015	Non discute di questioni inerenti alla domanda di ricerca		
Against the moral Turing test: accountable design and the moral reasoning of autonomous systems	Thomas Arnold, Matthias Scheutz	Scholar	Springer	29/03/2016	Non discute di questioni inerenti alla domanda di ricerca		
Using The Machine Stops for Teaching Ethics in Artificial Intelligence and Computer Science	Emanuelle Burton, Judy Goldsmith, Nicholas Mattei	Scholar	AAAI	01/02/2016	Non discute di questioni inerenti alla domanda di ricerca		
A minimalist model of the artificial autonomous moral agent (AAMA)	**Ioan Munean, Don Howard**	**Scholar**	**PhilPapers**	**01/01/2016**		**Propone alcuni design e architetture per moral agents (senza AV)**	**Propone soluzioni alle scelte morali che devono essere fatte da un agente autonomo**
Programming Machine Ethics	Luis Moniz Pereira, Ari Saptawijiaya	Scholar	Springer		Non discute di questioni inerenti alla domanda di ricerca		
Ethics and the safety of autonomous systems	**Catherine Menon, Alexander Menon, Robert David**	**Scholar**	**White Rose**	**01/01/2018**		**Evidenzia rischi e ethical drivers di comportamenti degli agenti autonomi, la necessità di trasparenza.**	**Presenta una metodologia per capire se un behaviour è etico o meno**
Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm	Michael Anderson, Susan Leigh Anderson	Scholar	Emerald Insight	02/07/2015	Non accessibile		
Ethical Decision Making During Automated Vehicle Crashes	**Nolah Goodall**	**Scholar**	**TRB**			**Comportamento di un AV durante un incidente**	**Approccio a 3 fasi per sviluppare algoritmi di crashing etici**
Ethical and Social Aspects of Self-Driving Cars	**Tobias Holstein, Gordana Dodig-Crnkovic, Patrizio Pelliccione**	**Scholar**	**arXiv**	**05/02/2018**		**Presenta strumenti regolatori, standard, design e implementazioni di componenti sistemi e servizi**	**Presenta sfide pratiche sociali ed etiche che devono essere incontrate, così come nuove aspettative per il software engineering**
The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?	**Sven Nyholm**	**Scholar**	**Springer**	**01/11/2016**		**Trova una serie di disanalogie tra accident-algorithms e il trolley problem**	**Propone una serie di spunti su cui riflettere nel design di agenti che devono affrontare il trolley problem**
A Rawlsian algorithm for autonomous vehicles	**Derek Leben**	**Scholar**	**Springer**	**04/03/2017**		**Solleva la discussione intorno alla non adattabilità del trolley problem**	**Propone un'alternativa alla soluzione utilitaria usata finora**
A vindication of the rights of machines	Gunkel, D.J.	Scopus	**Springer**	01/03/2014	Non discute di questioni inerenti alla domanda di ricerca		
Autonomous vehicles: challenges, opportunities, and future implications for transportation policies	Bagloee, S.A., Tavana, M., Asadi, M., Oliver, T.	Scopus	**Springer**	29/08/2016	Non discute di questioni inerenti alla domanda di ricerca		
Incorporating Ethics into Artificial Intelligence	**Etzioni, A., Etzioni, O.**	**Scopus**	**Springer**	**01/12/2017**		**Solleva il problema che il termine "autonomo" è fuorviante nel contesto degli AV, e porta a conclusioni errate nello stabilire la capacità di prendere decisioni etiche da parte di questi agenti.**	**Sostiene di dover cambiare prospettiva perchè quello su cui si basano i ragionamenti fatti finora sono basi errate.**
Technology as moral proxy: Autonomy and paternalism by design	**Millar, J.**	**Scopus**	**IEEE**	**08/09/2014**		**Solleva la necessità di avere un moral proxy a prendersi carico delle decisioni degli agenti autonomi**	**Propone anche per gli AV dei moral proxy, che siano i designer o i proprietari**