Un Robot Dotato di Saggezza Pratica Etica

Antonio Chella, Arianna Pipitone, Valeria Seidita
RoboticsLab, Dipartimento di Ingegneria, Università degli Studi di Palermo
{nome.cognome}@unipa.it

Abstract

Questo articolo descrive le ricerche in atto presso il RoboticsLab dell'Università degli Studi di Palermo relativamente alla creazione di un robot dotato di saggezza pratica etica.

1 Introduzione

La competenza nel ragionamento morale nei robot è un problema difficile. In letteratura si è dibattuto lungamente sulla natura di questo problema, sulla vasta conoscenza pratica necessaria per il ragionamento etico competente e sulla possibilità per un robot di raggiungere una sorta di saggezza etica [Lin, 2012] [Lin, 2017] [Liao 2020]] [Gunkel, 2012].
Qui descriveremo il quadro di riferimento complessivo adottato al RoboticsLab del Dipartimento di Ingegneria dell'Università degli Studi di Palermo per lo sviluppo di un robot dotato di saggezza pratica etica in grado di lavorare in team eterogenei composti da persone e robot a cui viene presentato un problema con risvolti etici da risolvere in modo collaborativo. Il sistema di ragionamento del robot è basato su agenti ed è in grado di generare un monologo interiore relativo ai suoi processi di inferenza interni.
L'ipotesi è che un team che condivide il processo di ragionamento interno sia maggiormente consapevole delle problematiche etiche che il robot si pone, e quindi l'intero processo decisionale possa portare ad un aumento di fiducia verso i robot.

2 Il Quadro di Riferimento

Attualmente, diversi robot e sistemi autonomi sono stati progettati per assistere le persone nella vita quotidiana, ma il loro impatto etico non è sempre evidente in relazione alla loro autonomia [Wallach, 2009]. Mentre in alcune situazioni la valenza etica è chiara, come nel caso delle auto a guida autonoma o nel caso dei droni e robot militari, ci sono altre situazioni in cui l'impatto etico è meno intuitivo. È il caso, ad esempio dei sistemi di assistenza domestica che interagiscono con le persone e ne acquisiscono le abitudini personali. Questi sistemi di Intelligenza Artificiale sono diventati pervasivi e gli utenti non sono sempre pienamente consapevoli delle problematiche etiche annesse, ovvero di come i loro dati personali vengano usati, manipolati e gestiti. [Coeckelbergh, 2020].
Oggi i robot assumono fattezze diverse e si presentano sotto forma di automobili a guida autonoma, di elettrodomestici con funzionalità avanzate, di sistemi domotici e così via. L'importanza di implementare abilità etiche in questi sistemi è essenziale: i robot devono essere in grado di prendere decisioni basate su parametri che tengono conto di norme, leggi e regole di sensibilità garantendo al contempo la sicurezza delle persone.
Considerata l'estrema importanza di avere accettabili comportamenti etici per i robot, è indispensabile allineare i valori etici dei robot con i valori etici umani [Vallor 2016] [Christian 2020]. È quindi auspicabile porre l'attenzione sulla definizione di regole universalmente riconosciute dalla comunità [Havens, 2016] [Dignum, 2019]. Tali regole dovrebbero essere prese come riferimento nella costruzione di robot morali per assicurare la massima cura e attenzione degli esseri umani e dei loro diritti. Tuttavia, questo non è banale e non è sempre possibile. I valori morali sono soggettivi e variano in base alla cultura, al credo, alla sensibilità e all'attitudine personale. Inoltre, ogni contesto ha una sua valenza morale. In campo bellico, per esempio, le questioni morali tengono conto di aspetti totalmente diversi dal contesto sociale comune [Arkin, 2009] [Walsh, 2017].
I problemi etici interessanti sono difficili da affrontare per diverse ragioni [Kearns, 2020]. In generale, un problema etico interessante è un problema che non si è mai presentato prima. Quindi l'agente, umano o artificiale, si trova di fronte a una situazione nuova e generalmente inaspettata da valutare. Tale valutazione implica una scelta, detta virtuosa perché coinvolgente aspetti etici e morali.
La scelta virtuosa in una situazione etica difficile non è semplicemente il risultato di inferenze logiche, basato su regole del tipo "Se sei in situazione X allora esegui l'azione Y". Allo stesso tempo, essendo un problema etico interessante nuovo e inaspettato, è difficile ottenere una soluzione attraverso sistemi basati su tecniche di apprendimento, come le reti neurali, in quanto non è sempre possibile disporre di un insieme di esempi di addestramento [Dietrich, 2021] [Christian, 2020].
Il problema quindi rimane aperto, ed innovative soluzioni devono essere delineate e sperimentate.

3 Un Modello ad Agenti di Saggezza Pratica

Il RoboticsLab dell'Università di Palermo ha attivato una stretta collaborazione con il filosofo americano John Sullins che ha proposto un quadro teorico basato sulla "artificial phronesis", un termine usato da Aristotele per descrivere le capacità legate alla saggezza pratica [Sullins, 2016, 2019, 2021]. Secondo Sullins, il comportamento di un agente etico dovrebbe essere guidato dalla saggezza pratica per compiere azioni virtuose.

Sullins propone un modello ad agenti per la saggezza pratica basata sul concetto di gioco sociale: "L'etica e la moralità potrebbero essere viste come una sorta di gioco sociale complesso, ad alta posta in gioco, che gli agenti umani giocano tra di loro. Dato che i computer hanno dimostrato di essere abili nel giocare, il gioco dell'etica dovrebbe essere alla fine risolvibile" ([Sullins, 2016] p. 38, nostra traduzione).

Consideriamo un esempio del modello computazionale composto da agenti software interagenti all'interno di un sistema di controllo del robot: alcuni agenti software saranno utilizzati per la classificazione dei problemi e altri agenti software per l'analisi dei casi. Gli agenti per la classificazione dei problemi analizzeranno le caratteristiche della situazione in atto, come se fosse una posizione degli scacchi.

Gli agenti software per l'analisi dei casi lavoreranno con gli agenti per la classificazione dei problemi. Dopo un'analisi iniziale ricevuta dagli agenti per la classificazione del problema, questi agenti inizieranno una ricerca in diversi archivi per trovare casi etici simili e correlati. I possibili archivi sono i database di casi etici forniti dalla comunità di ricerca etica. Gli agenti potranno anche eseguire una ricerca generale attraverso siti web e social media per analizzare e cercare casi simili e casi correlati. Gli agenti potranno chiedere aiuto a un mentore umano.

Come menzionato sopra, le due classi di agenti software interagiranno continuamente in un ciclo ricorrente: quando un agente per l'analisi dei casi troverà effettivamente casi interessanti che mostrano somiglianze con il problema in questione, allora interrogherà uno o più agenti per la classificazione dei problemi per analizzare meglio le somiglianze dei casi trovati in termini di contesto, agente, utente, e così via.

L'agente per la classificazione dei problemi potrà quindi raffinare la sua analisi del caso tenendo conto del ruolo del team member o del contesto. Questa analisi raffinata potrà a sua volta aprire nuove possibilità per gli agenti di analisi dei casi, e così via.

Quando le due classi di agenti raggiungeranno un accordo su una lista di casi correlati o su una combinazione di essi, allora il sistema software deputato alla pianificazione genererà diversi piani, ossia diverse sequenze di possibili azioni del robot. Questi piani saranno classificati secondo l'urgenza e la fiducia. Pertanto, un piano potrà essere scelto al posto di un altro perché è più veloce o più affidabile. Il sistema potrà anche simulare un piano prima che venga eseguito per valutare meglio la sua affidabilità. Infine, il robot si muoverà secondo il piano prescelto.

Come in un gioco come gli scacchi, i risultati delle mosse dell'agente saranno rivalutati dal sistema per decidere se continuare o cambiare strategia. A seconda dei risultati generali delle operazioni dell'agente, il robot riceverà un riscontro adeguato e potrà adattarsi meglio. Pertanto, a lungo termine, il robot acquisirà una capacità operativa appropriata, cioè una sorta di saggezza pratica [Sullins, 2021].

4 Il Monologo Interiore nei Robot

Questo modello ad agenti è applicato all'interazione tra componenti del team composto da persone e robot, ed è integrato da tecniche per la generazione del discorso interno nei robot sviluppate presso il RoboticsLab dell'Università di Palermo [Chella, 2020] [Pipitone, 2021a, 2021b, 2021c]. Queste tecniche forniscono un accesso uditivo ai processi degli agenti, descritti precedentemente, che controllano il comportamento del robot e che gli consentono di prendere decisioni mediante l'interazione con l'utente con il quale sta collaborando per portare a termine un compito condiviso.

Il robot dotato di monologo interiore può presentare i fatti materiali, i valori etici e i costumi sociali di cui è consapevole, e che ritiene siano rilevanti per il caso in questione, e che sta usando nel suo processo di ragionamento.

Anche se questi potrebbero non essere sufficienti da soli a risolvere problemi etici complessi, questo porta comunque alla consapevolezza dell'utente delle idee o dei concetti a cui non avrebbe potuto pensare se avesse dovuto risolvere il problema da solo, e porterà quindi a una migliore soluzione complessiva del problema [Chella 2022].

L'obiettivo finale della ricerca è di stabilire se l'accesso uditivo alla descrizione dei processi interni influenzi la fiducia che l'utente umano ripone nel sistema di ragionamento della stessa macchina [Geraci 2021].

5 Esperimenti

Gli esperimenti empirici in atto presso il RoboticsLab riguardano attualmente un semplice compito collaborativo quale è quello di apparecchiare una tavola da pranzo.

In questo caso le problematiche etiche sono semplificate e riguardano le eventuali richieste da parte dell'utente di violare l'etichetta informale.

Successivamente, il problema sarà generalizzato relativamente a problematiche di interazione tra persone e robot in ambito medico [Lanza 2020].

Ringraziamenti

Gli autori ringraziano John Sullins per le discussioni e i suggerimenti sulle tematiche del presente articolo.

La ricerca descritta nel presente articolo è parzialmente supportata dal progetto AFOSR RESPECT FA9550-19-1-7025.

Riferimenti bibliografici

[Arkin, 2009] Ronald C. Arkin. *Governing Lethal Behavior in Autonomous Robots.* CRC Press, 2009.

[Chella, 2020] Antonio Chella, Arianna Pipitone, Alain Morin and Famira Racy. Developing Self-Awareness in Robots via Inner Speech. *Frontiers in Robotics and AI*, 7:16, 2020 doi: 10.3389/frobt.2020.00016

[Chella, 2022] Antonio Chella and John P. Sullins. Building Competent Ethical Reasoning in Robot Applications: Inner Dialog as a Step Towards Artificial Phronesis In P. Wu et al. (eds.): *Papers of the 2022 Ethical Computing: Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning Symposium* AAAI Spring Symposium Series (AAAI SSS-22).

[Christian, 2020] Brian Christian. *The Alignment Problem. Machine Learning and Human Values*. W.W. Norton & Company, 2020.

[Coeckelbergh, 2020] Mark Coeckelbergh. *AI Ethics*. MIT Press, 2020.

[Dietrich, 2021] Eric Dietrich, Chris Fields, John P. Sullins, Bram van Heuveln and Robin Zebrowski. *Great Philosophical Objections to Artificial Intelligence.* Bloomsbury Academics, 2021.

[Dignum, 2019] Virginia Dignum. *Responsible Artificial Intelligence- How to Develop and Use AI in a Responsible Way.* Springer, 2019.

[Geraci, 2021] Alessandro Geraci, Antonella D'Amico, Arianna Pipitone and Antonio Chella. Automation Inner Speech as an Anthropomorphic Feature Affecting Human Trust: Current Issues and Future Directions. *Frontiers in Robotics and AI,* 8:620026. doi: 10.3389/frobt.2021.620026

[Gunkel, 2012] David J. Gunkel. *The Machine Question*. MIT Press, 2012.

[Havens, 2016] John C. Havens. *Heartificial Intelligence. Embracing Our Humanity to Maximize Machines*. Pen guin Random House, 2016.

[Kearns, 2020] Michael Kearns and Aaron Roth. *The Ethical Algorithm*. Oxford University Press, 2020.

[Lanza, 2020] Francesco Lanza, Valeria Seidita and Antonio Chella. Agents and robots for collaborating and supporting physicians in healthcare scenarios, *Journal of Biomedical Informatics,* 108, August 2020, 103483

[Liao, 2020] S. Matthew Liao (ed.). *Ethics of Artificial Intelligence*. Oxford University Press, 2020.

[Lin, 2012] Patrick Lin, Keith Abney and George A. Bekey (eds.). *Robot Ethics*. MIT Press, 2012.

[Lin, 2017] Patrick Lin, Ryan Jenkins and Keith Abney (eds.). *Robot Ethics 2.0*. Oxford University Press, 2017.

[Pipitone 2021a] Arianna Pipitone, Alessandro Geraci, Antonella D'Amico, Valeria Seidita and Antonio Chella. Robot's Inner Speech Effects on Trust and Anthropomorphic Cues in Human-Robot Co-operation In *Proceedings of SCRITA 2021* (arXiv:2108.08092), a workshop at IEEE RO-MAN 2021

[Pipitone, 2021b] Arianna Pipitone and Antonio Chella. What robots want? Hearing the inner voice of a robot. *iScience* 24, April 23, 2021.

[Pipitone, 2021c] Arianna Pipitone and Antonio Chella. Robot passes the mirror test by inner speech. *Robotics and Autonomous Systems,* 144, 2021

[Sullins, 2016] John P Sullins. Artificial phronesis and the social robot. In *What Social Robots Can and Should Do: Pro-ceedings of Robophilosophy 2016/TRANSOR 2016* 290 (2016), pp. 37–39.

[Sullins, 2019] John P Sullins. The Role of Consciousness and Artificial Phronesis in AI Ethical Reasoning. In A. Chella et al. (eds.): *Papers of the 2019 Towards Conscious AI Systems* AAAI 2019 Spring Symposium Series (AAAI SSS-19)

[Sullins, 2021] John P. Sullins. Artificial Phronesis: What It Is and What It Is Not. In *Science, Technology, and Virtues: Contemporary Perspectives*, Ch. 7, Oxford, 2021, Oxford University Press.

[Vallor 2016] Shannon Vallor. *Technology and Virtues*. Oxford University Press, 2016.

[Wallach, 2009] Wendell Wallach and Colin Allen. *Moral Machines. Teaching Robots Rights from Wrong.* Oxford University Press, 2009.

[Walsh, 2017] Toby Walsh. *It's Alive. Artificial Intelligence from the Logic Piano to Killer Robots*. La Trobe University Press, 2017.