Una Architettura Cognitiva per Interazione Affidabile tra Uomo e Robot

Filippo Cantucci, Rino Falcone, Cristiano Castelfranchi
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, (ISTC-CNR), Roma
filippo.cantucci@istc.cnr.it, rino.falcone@istc.cnr.it, cristiano.castelfranchi@istc.cnr.it

Abstract

La crescente complessità degli scenari di interazione tra uomo e robots autonomi, impone l'abilità da parte di tali sistemi di instaurare una forma di collaborazione avanzata, sia con gli umani stessi che con altri sistemi artificiali, più o meno sofisticati. La capacità dei robots di interagire efficacemente in domini distinti è un processo che sta evolvendo rapidamente. Tale prospettiva risulta tanto più utile, quando tali sistemi si dimostrano affidabili nelle loro finalità e comportamenti. Nei suddetti scenari, il concetto di affidabilità non è solo legato alla sicurezza o alla predicibilità del comportamento, ma è soprattutto da considerarsi in termini di capacità di adattare il proprio comportamento ai bisogni e scopi dell'umano. Tuttavia esiste una ulteriore dimensione che gioca un ruolo fondamentale, se propriamente sviluppata, ed è la capacità dei robots di fidarsi di altri sistemi artificiali intelligenti o degli umani stessi. Capacità di fidarsi e affidabilità, sono proprietà cruciali per definire una vera e profonda cooperazione e necessitano lo sviluppo di robots che siano in grado di avere una teoria della mente, sia in una prospettiva di lettura della mente dell'altro, sia in quella di strutturare la "propria mente" in modo tale da fidarsi di altri sistemi, artificiali o umani, sia nel saper rendendere il proprio comportamento spiegabile in termini mentali/cognitivi.

1 Introduzione

Una società guadagna risorse attraverso la collaborazione tra gli individui che la compongono [Kuipers, 2018]. Gli umani usualmente condividono conoscenza, delegano e adottano compiti, raggiungono e condividono obiettivi, al fine di ottenere risultati, in ogni dominio siano essi coinvolti. Per ognuna di queste attività, essi hanno una naturale predisposizione a collaborare in modo efficiente, comprensibile ed affidabile. Questo grazie a complesse capacità cognitive, apprese nel tempo. Consideriamo un ipotetico scenario nel quale un agente umano delega un compito ad un altro agente umano, che si impegna ad adottarlo (Figura 1). In questo caso, colui che adotta il compito è in grado di considerare diversi aspetti

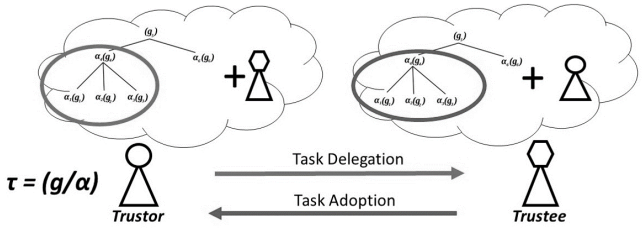

Figura 1: Concettualizzazione di delega e adozione di compiti. τ è il compito delegato, inteso come azione complessa o stato del mondo da realizzare

del contesto in cui si sviluppa il processo di delega e adozione ed è in grado di aggiustare la propria autonomia con l'obiettivo di fornire livelli di aiuto adatti a colui che ha delegato il compito. Grazie a queste capacità, l'adottante può svolgere il compito delegato nel modo più appropriato al contesto collaborativo, che include il delegante e i suoi stati mentali, ossia credenze, scopi, intenzioni, piani etc. La capacità di rappresentare gli stati mentali di altri agenti viene definita teoria della mente (ToM). Un'efficace collaborazione è assicurata anche ogni volta che l'adottante è consapevole delle proprie capacità di interpretare e rappresentare il mondo, i tratti e gli stati mentali del delegante.

Una grande sfida nell'ambito della interazione uomo-robot (HRI) è la progettazione di robots autonomi in grado di collaborarare efficacemente con gli esseri umani. I robots sono entrati a fare parte della vita quotidiana e sono presenti in molteplici ambienti (es. ospedali [Khan e Anwar, 2019], uffici [Veloso *et al.*, 2015], scenari turistici [Ivanov *et al.*, 2017] e così via). In questi contesti essi devono coesistere e interagire con un ampio spettro di utenti non necessariamente in grado (o disposti) ad adattare il proprio livello di interazione al tipo richiesto da una macchina: gli utenti devono confrontarsi con sistemi artificiali i cui comportamenti devono essere comprensibili e affidabili. In questi contesti, il concetto di affidabilità non è solo legato alla sicurezza o alla predicibilità del comportamento, ma va soprattutto considerato in termini di capacità di adattare l'autonomia all'utente interagente e di comportamento trasparente. In contesti di cooperazione tra uomo e robot è auspicabile che il ruolo dei robots non sia

quello di esecutori passivi, ma diventi quello di collaboratori attivi, dotati di iniziativa, anche intesa come elaborazione critica nella realizzazione di un compito. Consideriamo lo stesso scenario collaborativo introdotto sopra, ma sostituendo l'adottante con un sistema robotico. In questo contesto, l'agente umano $\mathcal{X}$ si affida al robot $\mathcal{R}$ per la realizzazione di una parte del piano che intende realizzare; da parte sua, $\mathcal{R}$ decide di aiutare $\mathcal{X}$, adottando parte del piano di $\mathcal{X}$ e/o raggiungendo alcuni sotto-obiettivi. Per fare qualcosa *per* (adozione di scopi di $\mathcal{X}$), $\mathcal{R}$ deve comprendere gli obiettivi e le intenzioni di $\mathcal{X}$, intese come aspettative di $\mathcal{X}$ sul comportamento di $\mathcal{R}$. La cooperazione, e di conseguenza la delega/adozione di compiti, implica qualcosa di più della semplice *obbedienza* agli ordini o della semplice *esecuzione* di un'azione prescritta. L'adozione di compiti presuppone intelligenza e *autonomia*. Tuttavia, anche nella semplice obbedienza ed esecuzione in realtà, il delegato $\mathcal{R}$ ha necessariamente un certo grado di autonomia e deve adeguare l'azione o lo scopo delegato al contesto e alle sue capacità attuali, intese anche come limiti nell'interpretare l'utente e le sue caratteristiche. Un vero robot collaborativo dovrebbe fornire all'utente diversi tipi di aiuto: $\mathcal{R}$ dovrebbe fare più di quanto esplicitamente richiesto da $\mathcal{X}$, al fine di raggiungere alcuni suoi obiettivi superiori (*over help*); inoltre, $\mathcal{R}$ dovrebbe essere in grado di correggere il piano di $\mathcal{X}$ e il compito assegnato, nel caso questo possa essere sbagliato, non ben concepito per gli stessi scopi di $\mathcal{X}$, o non perseguibile / realizzabile in quel contesto o da $\mathcal{R}$ stesso (*critical help*). $\mathcal{R}$ deve sfruttare la propria autonomia, competenza e capacità cognitive per trovare una soluzione migliore o possibile rispetto all'obiettivo di $\mathcal{X}$. Questo non deve necessariamente richiedere una trattativa, una discussione, un accordo; potrebbe essere un'iniziativa di $\mathcal{R}$. Questo è esattamente ciò che i robots dovrebbero mostrare e questi rappresentano il tipo di partner affidabile di cui gli esseri umani hanno bisogno. Come sarebbe possibile questa forma avanzata di cooperazione e collaborazione senza una reciproca ascrizione e lettura mentale? Non c'è vera cooperazione senza mente che rappresenta la mente del proprio interlocutore. $\mathcal{R}$ ha bisogno di capire fini, obiettivi più alti, aspettative di $\mathcal{X}$, per poterlo aiutare. Allo stesso modo, $\mathcal{X}$ deve capire cosa $\mathcal{R}$ sta facendo o ha fatto e perché, in termini mentali. Ed $\mathcal{R}$, se necessario, deve poter spiegare a $\mathcal{X}$ cosa ha fatto o ha intenzione di fare, in termini di proprie *ragioni*: dati, previsioni, obiettivi, soluzioni e così via. La figura 2 sintetizza i livelli di aiuto, concettualizzati in [Castelfranchi e Falcone, 1998] che un sistema robotico dovrebbe essere in grado di sfruttare per adattare la adozione di un compito alle esigenze dell'utente che glielo delega, alle condizioni di contesto in cui si svolge la collaborazione e alle proprie capacità di portare a termine un compito delegato (es: capacità di percepire le caratteristiche dell'utente e di attribuirle specifici stati mentali).

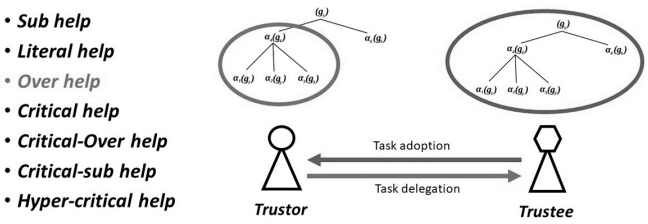

Figura 2: Livelli di aiuto nella adozione di un compito delegato

2 Architettura cognitiva

In questo lavoro presentiamo una architettura cognitiva [Cantucci e Falcone, 2020] che permette ad un robot autonomo di integrare diverse capacità cognitive, con lo scopo di fornire un aiuto trasparente, adattabile ed affidabile ogni volta che un umano gli delega un compito da raggiungere. L'architettura permette di considerare un robot come un *sistema intenzionale* [Dennett, 1971] in grado di:

- modellare e rappresentare gli stati mentali dell'umano interagente, in termini di credenze, scopi, piani [Scassellati, 2002];
- modulare il proprio livello di autonomia collaborativa, restringendo o espandendo un compito delegato e adottarlo a diversi livelli di aiuto[Castelfranchi e Falcone, 1998]. L'adozione è realizzata sulla base di diversi fattori contestuali come per esempio i bisogni e gli scopi non necessariamente dichiarati dall'utente umano;
- costruire il profilo dell'utente interagente e classificarlo, sulla base della capacità del robot di percepire e inferire specifiche caratteristiche dell'utente stesso, sia fisiche che sociali;
- decidere la modalità di adozione del compito delegato anche sulla base della *disponibilità* dell'utente, intesa come volontà da parte dell'utente di accettare un livello di aiuto (e di conseguenza di autonomia da parte del robot) differente da quello esplicitamente richiesto nella delega. Questa dimensione è molto importante e si basa sia su fattori intrinseci, legati alla personalità dell'utente [Matthews *et al.*, 2003], che su problemi "pratici" dell'utente stesso (es: mancanza di risorse in termini di tempo);
- autovalutare, sulla base del livello di adeguatezza del compito al profilo dell'utente (e alla sua disponibilità), l'affidabilità delle capacità su cui il robot si basa per profilare l'utente;
- generare una spiegazione delle ragioni che hanno portato il robot ad adottare il compito delegato con quello specifico grado di autonomia, con l'obiettivo di aumentare la trasparenza del comportamente e giustificare cambi rispetto alla richiesta letterale fatta dall'utente.

L'architettura proposta si basa sui principi fondamentali del Procedural Reasoning System (PRS) [Georgeff e Lansky, 1987], una delle architetture principali che esplicitamente integrano i concetti di credenze, desideri, intenzioni. La PRS definisce un sistema cognitivo basato su modellazione BDI ed esplicita ragionamenti basati su credenze, scopi, piani e intenzioni di un agente razionale.

3 Esperimenti

Per testare il modello computazionale descritto, abbiamo implementato una serie di esperimenti, sfruttando la piattaforma software *JaCaMo* [Boissier *et al.*, 2013], che simulano il processo di delega e adozione di compiti tra un robot e più utenti, raggruppati in classi di utenti, in uno specifico dominio applicativo. Abbiamo immaginato uno scenario collaborativo in cui il robot è un assistente turistico che aiuta le persone ad organizzare diverse attività turistiche offerte da una città (es. mangiare al ristorante, visitare un museo, visitare un monumento, bere qualcosa in un bar, godersi la città facendo più attività quotidiane). L'esperimento si basa sull'interazione tra due agenti: l'utente e il robot. Entrambi sono implementati come agenti Jason [Bordini e Hübner, 2005]. L'utente ha propri stati mentali rappresentati sotto forma di credenze, scopi, piani e interagisce con il robot delegandogli un compito. Da parte sua, il robot è in grado di rappresentare e attribuire stati mentali all'utente e a se stesso e, sulla base delle proprie capacità di profilare l'utente e costruirne un modello, di adottare il compito delegato a diversi livelli di aiuto.

Ad esempio, l'esperimento proposto in [Cantucci *et al.*, 2020] è stato progettato con l'obiettivo di mostrare l'importanza per un robot di autovalutare il livello di affidabilità associato alla sua capacità nella costruzione di un profilo dell'utente con cui interagisce. Questa capacità consente al robot di scegliere il compito migliore e adatto da adottare, rispetto alle caratteristiche dell'utente, anche quando le sue capacità si degradano progressivamente e possono essere considerate non affidabili. Il robot, infatti, è in grado di ordinare queste competenze sulla base del corrispondente livello di affidabilità, e fare leva su quelle più affidabili per decidere come adottare il compito delegato. Altri esperimenti sono stati condotti e esempi di spiegazione del compito da parte del robot sono stati descritti in lavori ancora sottoposti a processo di revisione.

4 Conclusioni e sviluppi futuri

In questo articolo abbiamo presentato un'architettura cognitiva di supporto al sistema decisionale di un robot sociale. Partendo dai principi di modellazione degli agenti BDI, abbiamo costruito un sistema intenzionale che integra i complessi stati mentali di delega e adozione di compiti e una teoria della mente, al fine di interagire efficacemente con gli esseri umani. L'architettura offre a un robot la capacità di costruire un profilo dell'utente e di sfruttare le sue capacità di profilazione in modo adattivo, considerando quelle abilità che massimizzano la valutazione delle prestazioni del compito dell'utente; consente al robot di ragionare su credenze, scopi, piani e intenzioni attribuiti all'utente e lo rende in grado di modulare la propria autonomia per il raggiungimento del compito delegato.

Se da un lato i robots, per poter collaborare in maniera efficace con gli umani, devono essere in grado di dimostrarsi affidabili nella collaborazione, facendo leva su alcune delle complesse abilità cognitive descritte sopra, dall'altro tali sistemi devono avere anche la capacità di fidarsi di altri sistemi intelligenti o degli umani stessi, con l'obiettivo di delegare compiti se non sono in grado di portarli a termine senza essere aiutati. La fiducia non è solo il risultato della frequenza con cui un agente produce un risultato o un comportamento atteso; la fiducia è un attitudine molto più complessa, che include attribuzione causale, stima, attribuzione di diversi fattori interni che giocano un ruolo causale nell'attivazione e nel controllo del comportamento; la fiducia è la controparte mentale della delega [Castelfranchi e Falcone, 2010].

L'architettura cognitiva proposta è pensata e progettata per supportare robots che sono integrati in ambienti complessi, in cui vi è la presenza sia di umani che di altri sistemi artificiali, più o meno complessi. Il principale lavoro futuro sarà quello di supportare un sistema robotico nello stimare il proprio grado di fiducia in altri agenti (umani e non) e, sulla base di tale stima, decidere se e a chi delegare parte del compito che a sua volta gli è stato delegato.

Inoltre continueremo a definire esperimenti, sia in simulazione che in reale, che ci consentano di indagare diversi aspetti del concetto di collaborazione affidabile tra robots e umani e tra robots ed altri sistemi artificiali, che considerano i robot come agenti cognitivi in grado di interagire con altri agenti, come fanno gli esseri umani quando interagiscono tra loro.

Riferimenti bibliografici

[Boissier *et al.*, 2013] Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, e Andrea Santi. Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6):747–761, 2013.

[Bordini e Hübner, 2005] Rafael H Bordini e Jomi F Hübner. Bdi agent programming in agentspeak using jason. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 143–164. Springer, 2005.

[Cantucci *et al.*, 2020] Filippo Cantucci, Rino Falcone, e Cristiano Castelfranchi. Investigating adjustable social autonomy in human robot interaction. *WOA 2021, 22nd Workshop "From Objects to Agents"*, 2020.

[Cantucci e Falcone, 2020] Filippo Cantucci e Rino Falcone. Towards trustworthiness and transparency in social human-robot interaction. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020.

[Castelfranchi e Falcone, 1998] Cristiano Castelfranchi e Rino Falcone. Towards a theory of delegation for agent-based systems. *Robotics and Autonomous Systems*, 24(3-4):141–157, 1998.

[Castelfranchi e Falcone, 2010] Cristiano Castelfranchi e Rino Falcone. *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.

[Dennett, 1971] Daniel C Dennett. Intentional systems. *The Journal of Philosophy*, pages 87–106, 1971.

[Georgeff e Lansky, 1987] Michael P Georgeff e Amy L Lansky. Reactive reasoning and planning. In *AAAI*, volume 87, pages 677–682, 1987.

[Ivanov *et al.*, 2017] Stanislav Hristov Ivanov, Craig Webster, e Katerina Berezina. Adoption of robots and service

automation by tourism and hospitality companies. *Revista Turismo & Desenvolvimento*, 27(28):1501–1517, 2017.

[Khan e Anwar, 2019] Arshia Khan e Yumna Anwar. Robots in healthcare: A survey. In *Science and Information Conference*, pages 280–292. Springer, 2019.

[Kuipers, 2018] Benjamin Kuipers. How can we trust a robot? *Communications of the ACM*, 61(3), 2018.

[Matthews *et al.*, 2003] Gerald Matthews, Ian J Deary, e Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[Scassellati, 2002] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.

[Veloso *et al.*, 2015] Manuela M Veloso, Joydeep Biswas, Brian Coltin, e Stephanie Rosenthal. Cobots: Robust symbiotic autonomous mobile service robots. In *IJCAI*, page 4423, 2015.