Silvia Rossi, Alessandra Rossi, Raffaella Esposito[**]

Robot Sociali per l'Assistenza: Inganni Pro-sociali ed Etica

Abstract
Social Assistive Robotics aims to improve people's lives by providing assistance through interaction. There are settings where these robots need to encourage people to comply with instructions and advice, and also show empathetic reactions, which is a crucial aspect for achieving effective behaviour change. Such behaviors can be configured as deceptive. Deception is a complex behavior in human interactions and can compromise people's trust. In this paper, we introduce different types of potentially deceptive behaviors that can be used in socially assistive robotics and discuss the ethical implications in human-robot interaction.

Keywords
Robot Deception, Socially Assistive Robotics, Acceptance of Technology, Trust, Leaking deception

Introduzione

I progressi dell'intelligenza artificiale e dell'automazione stanno portando a una rapida integrazione di nuove tecnologie in vari ambiti della vita quotidiana, come l'assistenza, la riabilitazione, l'educazione, ma anche in attività apparentemente più semplici come il supporto e la compagnia in ambienti domestici[1]. Per poter essere efficacemente integrati

[*] Università degli Studi di Napoli, "Federico II"
[**] Questo articolo è stato realizzato con il supporto di Air Force Office of Scientific Research under award number FA8655-23-1-7060. Qualsiasi opinione, risultato, conclusione o raccomandazione espressa in questo materiale è da attribuirsi esclusivamente agli autori e non riflette necessariamente le opinioni della United States Air Force.
[1] G. Wilson, *et al. Robot-Enabled Support of Daily Activities in Smart Home Environments*', "Cognitive systems research", 54, 2019, pp. 258-272; G. Mois, J.M Beer, *Robotics to support aging in place*, in "Living with Robots", Elsevier, 2020, pp. 49-74; C. Di Napoli, G. Ercolano, S. Rossi, *Personalized home-care support for the elderly: a field experience with a social robot at home*, "User Model User-Adap Inter 33", 2023, pp. 405-440. https://doi.org/10.1007/s11257-022-09333-y

in tali contesti, l'attenzione si sposta dalle semplici abilità di dialogo, rese possibili dall'Intelligenza Artificiale (AI) generativa, alla capacità di questi sistemi di esibire comportamenti socialmente appropriati. Le capacità di interazione di questi robot si sono evolute da semplici azioni comunicative a interazioni socio-emotive più complesse. Questo progresso consente ai robot sociali di soddisfare i bisogni terapeutici degli utenti, poiché il contesto emotivo può influenzare significativamente l'efficacia dell'assistenza fornita.

Uno dei compiti in cui i robot sociali possono avere un impatto positivo nella società moderna è il supporto al cambiamento comportamentale, dove è necessaria l'adesione alle istruzioni del robot da parte degli assistiti. Studi precedenti hanno dimostrato che i robot sociali per l'assistenza possono incrementare l'aderenza e la conformità alle prescrizioni delle persone rispetto ad altri strumenti tecnologici come tablet o smartphone[2]. Tuttavia, per supportare pienamente le persone, i robot devono essere in grado di imitare le capacità cognitive umane attraverso interazioni verbali e fisiche naturali e coinvolgenti, e di comprendere e adattarsi autonomamente alle loro esigenze e differenze[3]. La robotica sociale deve quindi evolversi verso un livello di interazione che vada oltre la semplice reazione allo stato emotivo immediato di una persona, considerando in un contesto più ampio il suo benessere psicologico generale. Ad esempio, la capacità dei robot sociali di adattarsi alle esigenze terapeutiche degli utenti creando un contesto emotivo positivo permette loro di influenzare in modo significativo l'efficacia dell'assistenza fornita. Per questo motivo, un robot sociale potrebbe strategicamente omettere informazioni o persino arrivare a mentire al fine di migliorare la stabilità psicologica degli assistiti. Tali strategie fanno un uso intenzionale di comportamenti potenzialmente ingannevoli.

L'inganno è un comportamento complesso nelle dinamiche relazionali tra esseri umani ed è un tema controverso, poiché può avere sia benefici (si pensi alla necessità di tacere a volte sulle reali condizioni del paziente, quando si tratta di un malato terminale) che effetti negativi sulle persone e sulle loro relazioni. Nell'ambito dell'interazione persona-macchina

[2] M.C. Carter, *et al.*, *Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial*, "Journal of medical Internet research", 15(4), 2013, p. e32.

[3] A. Tapus, M. Mataric, B Scassellati, *Socially assistive robotics [Grand Challenges of Robotics]*, "IEEE robotics & automation magazine", 14(1), 2007, pp. 35-42; N. Lee, *et al. The influence of politeness behavior on user compliance with social robots in a healthcare service setting*, "International journal of social robotics", 9(5), 2017, pp. 727-743; G. Maggi, *et al. "don't get distracted!": The role of social robots' interaction style on users' cognitive performance, acceptance, and non-compliant behavior*, "International journal of social robotics", 13(8), 2021, pp. 2057-2069.

non vi è un accordo unanime tra gli studiosi sulla definizione stessa di inganno. Secondo alcuni ricercatori, qualsiasi comportamento sociale di un robot, inclusa la capacità di mostrare emozioni, simulare empatia e di avere caratteristiche antropomorfe, può essere considerato ingannevole, in quanto induce le persone a credere che tali comportamenti siano reali[4]. Tuttavia, la possibilità di sfruttare comportamenti ingannevoli può assumere un valore positivo[5] se mirato ad aiutare le persone, come, ad esempio, per prevenire possibili conflitti, per ridurre il disagio emotivo, ma anche per migliorare le relazioni stesse tra esseri umani e robot[6]. Bugie bianche ed errori intenzionali possono talvolta rivelarsi utili in contesti educativi, sanitari e assistenziali[7]. Ad esempio, tali comportamenti possono migliorare le capacità di apprendimento degli studenti, favorire la riabilitazione dei pazienti e rassicurare le persone in situazioni di emergenza[8]. In generale, anche l'utilizzo strategicamente errato di un comportamento può avere lo scopo di indurre un effetto positivo desiderato nell'interazione con le persone e rappresenta esso stesso una forma di inganno[9]. I robot che dimostrano di non essere perfetti attraverso errori ben temporizzati e deliberati hanno il potenziale di apparire più simili agli esseri umani, meno alieni e intimidatori, e di suscitare simpatia, migliorando così l'efficacia dell'interazione stessa. Nei contesti educativi, gli errori indotti dai robot facilitano l'apprendimento degli studenti, ren-

[4] H.S. Sætra, *Social robot deception and the culture of trust*, "Paladyn: journal of behavioral robotics", 12(1), 2021, pp. 276-286; A. Sharkey, N. Sharkey, *We need to talk about deception in social robotics!*, "Ethics and information technology", 23(3), 2021, pp. 309-316.
[5] E. Adar, D.S. Tan, J. Teevan, *Benevolent deception in human computer interaction*, in "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", *CHI '13: CHI Conference on Human Factors in Computing Systems*, 2013 New York, NY, USA: ACM. Available at: https://doi.org/10.1145/2470654.2466246; J. Shim, R.C. Arkin, R.C. *A taxonomy of robot deception and its benefits in HRI*, in "2013 IEEE International Conference on Systems, Man, and Cybernetics", 2013 IEEE International Conference on Systems, Man and Cybernetics (SMC 2013), IEEE. Available at: https://doi.org/10.1109/smc.2013.398.
[6] H.S. Sætra, *Social robot deception and the culture of trust*, cit,
[7] E. Adar, D.S. Tan, J. Teevan, *Benevolent deception in human computer interaction*, cit.; J. Shim, R.C. Arkin, *A taxonomy of robot deception and its benefits in HRI*, cit.
[8] R.C. Arkin, P. Ulam, A.R. Wagner, *Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception*, "Proceedings of the IEEE. Institute of Electrical and Electronics Engineers", 100(3), pp. 571-589; J. Shim, R.C. Arkin, *The benefits of robot deception in search and rescue: Computational approach for deceptive action selection via case-based reasoning*', in "2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)", 2015 IEEE. Available at: https://doi.org/10.1109/ssrr.2015.7443002.
[9] H.S. Sætra, *Machiavelli for robots: Strategic robot failure, deception, and trust*, in "2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)". 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE. Available at: https://doi.org/10.1109/ro-man57019.2023.10309455.

dendo gli utenti più consapevoli delle proprie competenze[10]. In ambito terapeutico, i fallimenti programmati strategicamente rappresentano un modo alternativo di comportarsi in situazioni sociali, in contrapposizione ad aspettative rigide di comportamento[11].

L'uso dell'inganno prosociale presenta principalmente il vantaggio di accrescere la fiducia delle persone nei robot[12] aspetto essenziale per il successo delle interazioni tra esseri umani e robot nel lungo periodo[13]. Tuttavia, l'inganno può anche minare tale fiducia con conseguenze negative sull'intera interazione, producendo una percezione negativa della loro affidabilità e dei loro obiettivi[14]. Una calibrazione errata della fiducia può anche portare a un effetto opposto, ovvero a una fiducia eccessiva[15]. Ad esempio, l'inganno robotico nell'assistenza agli anziani è considerato rischioso perché potrebbe indurli a sviluppare una dipendenza eccessiva[16] e compromettere il loro benessere emotivo se sviluppassero attaccamenti basati su interazioni ingannevoli[17].

In questo lavoro, intendiamo evidenziare la letteratura esistente sull'uso dell'inganno nell'interazione tra esseri umani e robot, delineando le definizioni e le tecniche maggiormente diffuse. Tale discussione mira ad accrescere la consapevolezza dei potenziali rischi etici legati all'uso di robot per l'assistenza, al fine di garantire che le innovazioni nel campo della

[10] R.C. Arkin, P. Ulam, A.R. Wagner, *Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception*, "Proceedings of the IEEE. Institute of Electrical and Electronics Engineers", 100(3), 2012, pp. 571-589.
[11] H.S. Sætra, *Machiavelli for robots: Strategic robot failure, deception, and trust*, cit.
[12] H.S. Sætra, *Social robot deception and the culture of trust*, cit.,
[13] Rossi, A. *et al.*, *How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario*, in "*Social Robotics*", Cham: Springer International Publishing (Lecture notes in computer science), 2017, pp. 42-52.
[14] C. Castelfranchi, *Ethics and information technology*, 2(2), 2000, pp. 113-119; A. Sharkey, N. Sharkey, *'We need to talk about deception in social robotics!*, *"Ethics and information technology*", 23(3), 2021, pp. 309-316.
[15] A.M Aroyo, *et al. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation*, "Paladyn: journal of behavioral robotics", 12(1), 2021, pp. 423-436.
[16] R. Carli, A. Najjar, *Reconsidering deception in social robotics: The role of human vulnerability (student abstract)*, "Proceedings of the … AAAI Conference on Artificial Intelligence", AAAI Conference on Artificial Intelligence, 37(13), 2023, pp. 16174-16175.
[17] M. Coeckelbergh, *Artificial companions: Empathy and vulnerability mirroring in human-robot relations*, "Studies in ethics, law, and technology", 4(3), 2011. Available at: https://doi.org/10.2202/1941-6008.1126; A. Sharkey, N. Sharkey, *Granny and the robots: ethical issues in robot care for the elderly*, "Ethics and information technology", 14(1), 2012, pp. 27-40. A. Matthias, *Robot Lies in Health Care: When Is Deception Morally Permissible?*, "Kennedy Institute of Ethics journal", 25(2), 2015, pp. 169-192; J.B. Bell, *Toward a theory of deception*, "International journal of intelligence and counterintelligence", 16(2), 2003, pp. 244-279.

robotica siano allineate ai principi etici e contribuiscano positivamente all'esperienza umana.

Che cosa è un comportamento ingannevole nell'interazione persona-robot

L'inganno può essere categorizzato in due tipi principali: nascondere la verità e mostrare il falso[18]. Queste categorie rappresentano strategie distinte per fuorviare gli altri riguardo alla realtà e possono essere attuate sia attraverso comportamenti verbali (cioè ciò che si dice) sia, nel caso di un sistema robotico, anche non verbali (cioè quello che si fa e come lo si fa). Nelle interazioni verbali, il *mostrare il falso* attraverso il linguaggio è definito falsificazione, e consiste nel fornire informazioni completamente false o bugie[19]. Si consideri, ad esempio, un robot che ricopre il ruolo di motivatore durante una terapia e che inciti il paziente con incoraggiamenti non basati sui fatti ("stai andando bene!", "avanti così!")[20]. Nascondere la verità nella comunicazione verbale si traduce in omissioni, occultamenti ed equivoci in cui alcune informazioni non vengono rivelate, lasciando l'ascoltatore con una comprensione incompleta o distorta della realtà[21]. Ad esempio, un robot potrebbe fornire dei consigli alla persona omettendo le "ragioni" di tali consigli o la spiegazione del meccanismo di ragionamento che ha generato il consiglio stesso. Nelle interazioni che prevedono comportamenti non verbali e ingannevoli, *mostrare il falso* può essere ottenuto attraverso l'espressione facciale, i gesti del corpo, la vicinanza fisica o le traiettorie di movimento, al fine di indurre interpretazioni errate della situazione[22]. D'altro canto, nascondere la verità si manifesta quando determinati segnali non verbali vengono intenzionalmente omessi. Ad esempio, si potrebbe evitare di mostrare determinati gesti ed espressioni facciali che normalmente trasmettono informazioni significative sullo stato interno del robot, o si potrebbero modificare pose e traiettorie di movimento per non rivelare elementi chiave necessari a una corretta interpretazione della situazione[23] e delle azioni del robot.

[18] J.B. Bell, *Toward a theory of deception*, cit.
[19] D.B. Buller, *et al.*, *Testing interpersonal deception theory: The language of interpersonal deception*, "Communication theory: CT: a journal of the International Communication Association", 6(3), 1996, pp. 268-289.
[20] D. Laparidou, *et al.*, *Patient, carer, and staff perceptions of robotics in motor rehabilitation: a systematic review and qualitative meta-synthesis*, "Journal of neuroengineering and rehabilitation", 18(1), 2021, p. 181.
[21] D.B. Buller, *et al.*, *Testing interpersonal deception theory: The language of interpersonal deception*, "Communication theory: CT: a journal of the International Communication Association", 6(3), 1996, pp. 268-289.
[22] J. Shim, R.C. Arkin, R.C. *A taxonomy of robot deception and its benefits in HRI*, cit.
[23] *Ibidem*

Chisholm e Freehan[24] distinguono tre dimensioni chiave dell'inganno, adottando una prospettiva logica e filosofica:

– *Commissione-omissione*, in cui l'ingannatore o causa attivamente un cambiamento di ciò che pensa il soggetto dell'inganno o, al contrario, permette tale cambiamento in maniera passiva;
– *Positivo-negativo*, in cui l'ingannatore induce il soggetto dell'inganno a credere che una proposizione falsa sia vera o che una proposizione vera sia falsa;
– *Intenzionale-non intenzionale*, che si riferisce al fatto che l'ingannatore alteri deliberatamente la credenza del soggetto dell'inganno o semplicemente la mantenga invariata.

In questo lavoro, sosteniamo che un robot non possa tecnicamente ingannare, in quanto macchina non è dotata della possibilità di un comportamento intenzionale. Essa può essere però uno strumento di inganno e a tal fine programmata per un obiettivo di commissione (si veda ad esempio il cambio di comportamento e l'aderenza); di conseguenza, gli esseri umani coinvolti nello sviluppo dei robot sociali sono gli unici responsabili delle conseguenze che l'inganno robotico ha sulle relazioni tra le persone e i robot[25]. Inoltre, sebbene l'inganno nei robot possa verificarsi spesso in modo non intenzionale[26], il nostro principale interesse è concentrato sull'inganno volontario.

De Paulo[27] classifica l'inganno in quattro aree principali: *contenuto, tipo, referente e motivazioni.* Il *contenuto* è l'oggetto dell'inganno e può riguardare sentimenti, risultati, azioni, spiegazioni o fatti. Le *motivazioni* dell'inganno possono essere *orientate verso se stessi* (a vantaggio dell'ingannatore) o *altruiste* (a beneficio della persona ingannata). I tipi di inganno possono essere diretti, esagerati o sottili, mentre il referente può essere l'ingannatore stesso, il bersaglio, un'altra persona o un oggetto/evento. Un sistema robotico è potenzialmente in grado di ingannare considerando tutte le possibilità di contenuto; noi ci aspettiamo che un utilizzo etico della robotica assistiva, così come in generale dell'AI, sia sempre focalizzato su un utilizzo dell'inganno "altruistico" e mai "utilitaristico". Se prendiamo in considerazione la classificazione di Erat e Gneezy[28] che suddividono l'inganno in base ai

[24] R.M. Chisholm, T.D. Feehan, *The intent to deceive*, "The journal of philosophy", 74(3), 1977, p. 143.
[25] H.S. Sætra, *Social robot deception and the culture of trust*, cit.
[26] *Ibidem*
[27] B.M. DePaulo, *et al.*, *Lying in everyday life*, "Journal of personality and social psychology", 70(5), 1996, pp. 979-995.
[28] S. Erat, U. Gneezy, *White lies*, "Management science", 58(4), 2012, pp. 723-733.

suoi effetti, identificando quattro tipologie: 1) le *bugie nere egoistiche*, che avvantaggiano l'ingannatore a scapito del bersaglio; 2) le *bugie nere maliziose*, che danneggiano sia l'ingannatore che il bersaglio; 3) le *bugie bianche di Pareto*, che apportano benefici a entrambe le parti; e 4) le *bugie bianche altruiste*, in cui il soggetto dell'inganno trae beneficio dall'inganno, mentre l'ingannatore subisce una perdita; le bugie bianche dovrebbero essere l'unica forma di inganno ammissibile, sebbene eticamente discutibile.

Tali categorizzazioni dell'inganno trovano riscontro in altrettante tassonomie presentate nella letteratura sull'interazione persona-robot. In particolare, Shim e Arkin[29] classificano l'inganno dei robot secondo tre dimensioni riconducibili alle teorie di De Paulo e che considerano: 1) l'*oggetto dell'interazione,* ovvero l'agente che viene ingannato (può essere umano o non umano); 2) lo *scopo dell'inganno,* che distingue tra inganno *rivolto a se stessi*[30] e inganno *orientato verso gli altri*[31]; 3) il *metodo di inganno*, che comprende l'inganno fisico/di apparenza, dove l'inganno viene manipolato attraverso l'embodiment del robot, e l'inganno mentale/comportamentale, dove l'inganno è nei comportamenti del robot, e può includere la falsa rappresentazione del proprio stato attuale, la modifica del suo comportamento in modo fuorviante, oppure l'uso di segnali verbali o non verbali per creare un'impressione falsa. L'oggetto dell'interazione può essere così anche un oggetto non umano. Le combinazioni tra questi livelli generano otto diversi tipi di inganno robotico. Danaher, invece, propone un approccio più incentrato sulla tecnologia, identificando tre categorie di inganno robotico che si concentrano sull'oggetto dell'inganno e, in particolare, sul fatto che esso riguardi uno *stato esterno, superficiale o interno* al robot[32]. Per stato esterno s'intende la presentazione di informazioni false o l'omissione di informazioni riguardanti il mondo che circonda il robot. L'inganno su uno *stato superficiale* riguarda, ad esempio, la falsa rappresentazione delle caratteristiche esteriori del robot, come un'espressione facciale o un'apparenza umanoide. L'inganno su uno *stato nascosto* implica la dissimulazione o la distorsione degli stati

[29] J. Shim, J. and Arkin, R.C. (2013) 'A taxonomy of robot deception and its benefits in HRI', cit.; H.S. Sætra, *Social robot deception and the culture of trust*, cit.

[30] I. Deweese-Boyd, *Self-Deception,* "The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab", Stanford University, 2023, https://plato.stanford.edu/archives/fall2023/entries/self-deception/.

[31] J. Shim, R.C. Arkin, *Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark*, in "2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)". 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2014, IEEE. Available at: https://doi.org/10.1109/robio.2014.7090385.

[32] J. Danaher, *Robot Betrayal: a guide to the ethics of robotic deception*, *"Ethics and information technology*", 22(2), 2020, pp. 117-128.

interni del robot, come i suoi meccanismi di decisione o i suoi obiettivi. Secondo Sætra[33] però questi due ultimi livelli di inganno sono intrinsecamente collegati poiché le persone tendono a inferire gli stati interni dai segnali superficiali. Questo si riflette anche nella letteratura corrente sulla trasparenza durante l'interazione persona-robot, dove il processo di comunicazione dei meccanismi interni di un robot non è solo legato a un processo di spiegazione verbale (eXplainable Artificial Intelligence – XAI), ma anche a meccanismi di comunicazione non verbale che rendano chiare le azioni e gli obiettivi del robot[34].

Collegate alla tassonomia di Danaher, Sætra introduce due forme di inganno nei robot sociali[35]: *inganno totale* e *inganno parziale*. L'*Inganno totale* si verifica quando una persona crede completamente che un robot sociale sia qualcosa di diverso da una macchina (ad esempio un essere umano o un animale). In questo caso, il soggetto è ingannato sia consapevolmente che inconsciamente, arrivando ad accettare pienamente il robot come un'entità diversa, in modo simile a quanto avviene nel superamento del test di Turing. L'*Inganno parziale* si verifica quando una persona reagisce inconsciamente al robot come se fosse un'entità reale, pur sapendo razionalmente che si tratta di una macchina. Ad esempio, i robot sociali possono suscitare risposte emotive simili a quelle provocate da esseri viventi. In tal caso i meccanismi di inganno in robotica sociale dovrebbero essere sempre confinati con possibili effetti di inganno parziale e mai totale.

Quali sono gli effetti di comportamenti ingannevoli in interazione persona-robot

Al fine di esplorare gli effetti di meccanismi di inganno durante l'interazione persona-robot abbiamo condotto una revisione sistematica dello stato dell'arte *[anomysed]*[36]. Tale revisione ci ha permesso di categorizza-

[33] H.S. Sætra, *Social robot deception and the culture of trust*, cit.
[34] G. Angelopoulos, *et al.*, *Using theory of mind in explanations for fostering transparency in human-robot interaction*, in "Lecture Notes in Computer Science", Springer Nature Singapore (Lecture notes in computer science), Singapore, 2024, pp. 394-405; N. Chandran Nair, A. Rossi, S. Rossi, *Impact of explanations on transparency in HRI: A study using the HRIVST metric*', in *Lecture Notes in Computer Science*, Springer Nature Singapore, Singapore 2024 (Lecture notes in computer science), pp. 171-180; A. Rossi, S. Rossi, *On the way to a transparent HRI*, in "Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization", UMAP '24: 32nd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA 2024: ACM. Available at: https://doi.org/10.1145/3631700.3664890.
[35] H.S. Sætra, *Social robot deception and the culture of trust*, cit.
[36] R. Esposito, A. Rossi, S. Rossi, *Deception in HRI and Its Implications: A Systematic Review*,

re i comportamenti ingannevoli dei robot nelle interazioni persona-robot mettendo in luce i vari effetti derivanti e evidenziando le risposte psicologiche umane sia positive che negative che qui discutiamo brevemente.

Un primo parametro di valutazione del comportamento ingannevole persona-robot riguarda la percezione che l'utente ha di tale comportamento e, di conseguenza, quanto lo ritenga accettabile[37]. In generale, la percezione e l'accettabilità dell'inganno variano a seconda del contesto, del comportamento e della sensibilità individuale[38]. Ad esempio, l'utilizzo di traiettorie ingannevoli si è dimostrato efficace nell'ingannare gli utenti sulla destinazione del movimento stesso, ma tale inganno è più efficace quando si utilizzano strategie di movimento dinamiche e variabili, rispetto ad approcci statici, in quanto, in interazioni ripetute, l'utente riesce a predire i movimenti del robot con il passare del tempo[39]. L'inganno, quindi, almeno nelle sue manifestazioni attraverso il comportamento (inganno superficiale), è efficace solo se non ripetitivo. Gli esseri umani, infatti, sono in grado di riconoscere gli stessi pattern nel tempo e adattarsi ad essi. Quando si utilizza l'inganno con strategie di dialogo simili a quelle tra gli esseri umani (ad esempio usando un linguaggio che lo faccia percepire come un agente altamente sociale e capace di provare empatia e affetto), le persone tendono a considerare un robot come ingannevole in modo accettabile o non ingannevole[40], attribuendo il comportamento, eventualmente giudicato erroneo, del robot alla sua natura meccanica[41]. Anche in questo caso, strategie di inganno superficiale (modifica dello

in "J. Hum.-Robot Interact", 14, 3, Article 47, 2025. https://doi.org/10.1145/3721297

[37] A. Rossi, S. Rossi, *Evaluating people's perception of trust of a deceptive robot with theory of mind in an assistive gaming scenario*, in "2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)", 2023 IEEE. Available at: https://doi.org/10.1109/ro-man57019.2023.10309647.

[38] A. Rosero, *Using justifications to mitigate loss in human trust when robots perform norm – violating and deceptive behaviors*, in "Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA(2023): ACM. Available at: https://doi.org/10.1145/3568294.3579979.

[39] J. Bowyer Bell, B. Whaley, *Cheating And Deception*, pp. 479 Routledge, 10.4324/9781315081496, 1991; A. Dragan, R. Holladay, S. Srinivasa, *Deceptive robot motion: synthesis, analysis and experiments*, "Autonomous robots", 39(3), 2015, pp. 331-345.

[40] K. Winkle, *et al.*, *Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots*, in "Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA 2021: ACM. Available at: https://doi.org/10.1145/3434073.3444666.

[41] E. Short, *et al. No fair!! An interaction with a cheating robot*, in "2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", 2010 IEEE. Available at: https://doi.org/10.1109/hri.2010.5453193

stile di dialogo e comportamento) sono risultate efficienti nel generare una accettazione positiva dell'interazione. Generalmente, l'inganno fisico (ad esempio implementato attraverso movimenti ingannevoli) è generalmente percepito come intenzionale, mentre quello verbale è interpretato come un difetto di progettazione[42]. Infine, c'è da notare che l'inganno prosociale aumenta l'accettabilità dell'inganno da parte delle persone quando questo viene contestualizzato (ad esempio, in scenari di ricerca e soccorso o nell'educazione)[43].

L'inganno da parte di un robot influisce sulla percezione che le persone hanno dell'interazione sociale con il robot stesso, influenzando dimensioni come la presenza sociale, il coinvolgimento, l'intrattenimento, il divertimento e la motivazione. Nuovamente, la letteratura ci mostra che tali benefici sono legati principalmente a meccanismi di inganno superficiale. I comportamenti emotivi aumentano la percezione della presenza sociale del robot, ma non la sua accettazione o il livello di antropomorfismo percepito. La capacità del robot di avere dialoghi simili a quelli umani, invece, aumentano la simpatia, la motivazione e la preferenza per una collaborazione con i robot. L'inganno prosociale, sotto forma di feedback positivo (si ricordi l'esempio del robot motivatore), rende il feedback del robot più evidente e utile[44]. Al contrario, robot che mentono o scaricano la colpa su altre persone sono percepiti come meno amichevoli e gentili[45]. Traiettorie di movimento ingannevoli aumentano il divertimento, l'intrattenimento e il coinvolgimento durante l'interazione. I robot che barano sono considerati più coinvolgenti, ma anche associati a tratti negativi, quali la disonestà. Le persone considerano i comportamenti ingannevoli superficiali stimolanti e motivanti, e l'apprendere dell'inganno robotico non ha ridotto la loro volontà di continuare l'interazione. I robot che utilizzano traiettorie di movimento ingannevoli sono considerati avversari, anche se sono percepiti come più razionali e intelligenti.

Un secondo possibile parametro di valutazione è l'impatto che comportamenti ingannevoli hanno sulla fiducia tra persona e robot. I dialoghi con una connotazione emotiva aumentano la percezione di buona

[42] J. Shim, R.C. Arkin, *A taxonomy of robot deception and its benefits in HRI,* cit.
[43] J. Shim, R.C. Arkin, *Other-oriented robot deception: How can a robot's deceptive feedback help humans in HRI?*, in "*Social Robotics"*, Cham: Springer International Publishing (Lecture notes in computer science), 2016, pp. 222-232, Lecture Notes in Computer Science(), vol 9979. Springer, Cham. https://doi.org/10.1007/978-3-319-47437-3_22
[44] *Ibidem.*
[45] *Ibidem*; L. Wijnen, J. Coenen, B. Grzyb, *'It's not my Fault!'*, in "Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction", HRI '17: ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2017: ACM. Available at: https://doi.org/10.1145/3029798.3038300; G. Wilson, *et al.*, *Robot-Enabled Support of Daily Activities in Smart Home Environments*, "Cognitive systems research", 54, 2019, pp. 258-272.

volontà, mentre l'addebito delle colpe in caso di fallimento e il fornire suggerimenti errati riducono la fiducia. La fiducia è stata negativamente influenzata quando le persone hanno scoperto che il robot aveva mentito. Inoltre, è stato riscontrato che la fiducia può aumentare quando il robot inizia con un comportamento ingannevole ma successivamente dimostra onestà, mentre diminuisce quando un comportamento onesto viene seguito da azioni ingannevoli. Inoltre, la gestione dell'inganno dopo la sua scoperta è cruciale: la mancanza di riconoscimento o la gestione inappropriata delle scuse mal gestite possono aggravare i problemi di fiducia[46]. Al contrario, fornire giustificazioni riguardo allo stato mentale o motivazioni plausibili per l'inganno può mitigare alcuni degli effetti negativi sulla fiducia.

Merita una menzione particolare, infine, l'impiego di sistemi robotici che utilizzano una comunicazione basata su segnali emotivi[47]. L'uso di robot che esprimono emozioni suscita risposte individuali diverse. Il comportamento emotivo non sembra influire significativamente sui livelli di attaccamento nel tempo. Tuttavia, l'instaurazione di un collegamento emotivo è positivamente correlata alla percezione di facilità d'uso e alla percezione del robot come entità sociale. La costruzione di un legame emotivo tra le persone e i robot è fortemente correlata anche a fattori come antropomorfismo, simpatia e intelligenza percepita. Un attaccamento emotivo influenza significativamente la percezione delle persone del robot come entità sociale e riduce gli effetti negativi dovuti all'inganno nel tempo, indicando un adattamento al robot. I comportamenti antropomorfi dei robot portano a una maggiore conformità alle richieste del robot, in particolare nelle interazioni dirette faccia a faccia. È interessante notare che i comportamenti ingannevoli fanno sì che le persone considerino i robot più strategici e intelligenti.

In sintesi, l'inganno superficiale nel rapporto tra esseri umani e robot favorisce l'illusione di una presenza sociale o di intelligenza che può portare benefici durante l'interazione. L'inganno superficiale può rendere i robot più coinvolgenti a livello sociale e persino di incentivare comportamenti desiderabili negli utenti. Dall'altro lato, gli utenti sono spesso sensibili agli episodi di inganno, specialmente quando lo percepiscono come manipolativo o egoistico. Tali tipi di inganni, però, sono maggiormente

[46] K. Rogers, R.J.A Webber, A. Howard, *Lying about lying*, in "Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA 2023: ACM. Available at: https://doi.org/10.1145/3568294.3580178; A. Rosero, *Using justifications to mitigate loss in human trust when robots perform norm – violating and deceptive behaviors*, cit.

[47] A. van Maris, *et al.*, *Designing Ethical Social Robots-A Longitudinal Field Study With Older Adults*, "Frontiers in robotics and AI", 7, 2020, p. 1.

legati allo stato interno del robot e a obiettivi di manipolazione che non traspaiono nello stato superficiale. Sebbene l'inganno non percepito sia più coerente con l'obiettivo di fuorviare gli utenti, abbiamo riscontrato che anche l'inganno notato dagli utenti può avere implicazioni significative sull'interazione. In particolare, in contesti di intrattenimento, l'inganno può essere usato per aumentare il coinvolgimento, la percezione di intenzionalità del robot, l'intelligenza percepita, e il divertimento.

Dilemmi etici e implicazioni dell'inganno

La categorizzazione dei comportamenti ingannevoli e delle loro implicazioni nelle interazioni tra le persone e i robot consente di sviluppare strategie per mitigare gli effetti negativi e promuovere risultati positivi nella progettazione robotica.

Un primo canale di valutazione dell'eticità di un comportamento ingannevole in interazione persona-robot è sicuramente da identificarsi nel carattere altruistico di tale comportamento. Spesso, infatti, tale caratteristica è legata al dominio di applicazione del sistema robotico. Il dominio stesso, infatti, può caratterizzarne l'utilizzo etico. Ad esempio, nel settore dell'intrattenimento e del gaming, i robot possono utilizzare forme di inganno per rendere le esperienze più coinvolgenti e imprevedibili, aumentando il divertimento e l'interazione degli utenti[48]. Nei contesti critici per la sicurezza, i robot possono ricorrere all'inganno per influenzare il comportamento umano in modo positivo. Nel settore dell'assistenza, in particolare nella riabilitazione, i robot possono ricorrere all'inganno per rendere l'esperienza dei pazienti più coinvolgente e meno frustrante. Ad esempio, un robot potrebbe esagerare i progressi del paziente ingannandolo sul livello di miglioramento: "Oggi hai migliorato gli esercizi del 10% rispetto a ieri", con l'obiettivo di mantenerne alta la motivazione e l'autostima. Nel contesto educativo, i robot possono ricorrere all'inganno prosociale per fornire feedback positivi indipendentemente dalle prestazioni, incoraggiando così gli studenti a impegnarsi più a lungo e a migliorare i loro risultati di apprendimento.

Tuttavia, permangono ancora delle lacune nella ricerca sull'inganno robotico. In primo luogo, è necessario ampliare la diversificazione e

[48] D.B. Buller, *et al.*, *Testing interpersonal deception theory: The language of interpersonal deception*, cit.; E. de Oliveira, L. Donadoni, S. Boriero, S. et al. *Deceptive Actions to Improve the Attribution of Rationality to Playing Robotic Agents*. Int J of Soc Robotics 13, 391-405 (2021). https://doi.org/10.1007/s12369-020-00647-8; A. Ayub, A. Morales and A. Banerjee, *Using Markov Decision Process to Model Deception for Robotic and Interactive Game Applications*, 2021 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2021, pp. 1-6, doi: 10.1109/ICCE50685.2021.9427633.

l'inclusività degli studi, includendo un'ampia varietà di culture, lingue e contesti, per comprendere appieno l'impatto globale dell'inganno robotico. Inoltre, i confronti tra diverse fasce d'età potrebbero rivelare differenze significative nella percezione e nella reazione all'inganno. Mentre alcuni studi non riportano le caratteristiche del campione, i risultati indicano che l'attaccamento emotivo derivante dai comportamenti ingannevoli è stato studiato prevalentemente negli anziani. Inoltre, molti studi esistenti si concentrano su scenari in cui i robot manipolano le percezioni degli esseri umani attraverso bugie o segnali fuorvianti, mentre l'omissione strategica o la dissimulazione di informazioni rilevanti sono meno esplorate. Questa lacuna rappresenta un potenziale ambito di ricerca futura.

Un'altra area di ricerca cruciale per il futuro potrebbe essere la mitigazione della perdita di fiducia derivante dalle interazioni ingannevoli, nonché l'esplorazione degli effetti psicologici a lungo termine dell'inganno robotico nelle relazioni persona-robot e delle circostanze in cui i benefici di tale inganno potrebbero superare le implicazioni etiche negative. Riconoscere la variabilità nella percezione e accettazione dell'inganno nei robot sottolinea l'importanza di adottare approcci personalizzati nella progettazione dell'interazione. Sebbene la ricerca si concentri sulle risposte degli esseri umani all'inganno nei robot, c'è una limitata esplorazione dei meccanismi e dei processi per il riconoscimento e la gestione dell'inganno stesso. Comprendere le differenze individuali nel riconoscimento dell'inganno potrebbe condurre allo sviluppo di robot capaci di adattare il loro comportamento in base alle capacità di ogni utente di rilevare e tollerare l'inganno. Inoltre, comprendere i meccanismi di riconoscimento dell'inganno potrebbe facilitare la creazione di strategie ingannevoli più specifiche, garantendo l'uso dell'inganno solo quando è effettivamente richiesto dal ruolo del robot. Tuttavia, non tutte le forme di inganno comportano necessariamente un eccesso di fiducia. Alcuni tipi di inganno, come quelli legati a stati esterni, possono essere impiegati strategicamente per gestire le aspettative dell'utente e garantire che le interazioni rimangano ancorate alle reali capacità del robot, così come gli errori strategici.

Le considerazioni fatte evidenziano anche la necessità di avere delle regolamentazioni per la tutela dei diritti individuali. Un primo esempio è rappresentato dall'Artificial Intelligence Act dell'Unione Europea[49], che

[49] REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS COM/2021/206 final: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

mira a stabilire un quadro normativo completo per le tecnologie di intelligenza artificiale. In particolare, l'articolo 5 dell'AI Act vieta i sistemi di IA che utilizzano tecniche subliminali, manipolative o ingannevoli per distorcere materialmente il comportamento umano, causando o rischiando di causare danni significativi. Tale regolamentazione non influisce sul lavoro di ricerca nei campi dell'intelligenza artificiale e dell'interazione tra esseri umani e robot, tuttavia introduce limiti poco chiari che potrebbero portare a restrizioni sull'uso di robot ingannevoli in applicazioni nel mondo reale.

Come menzionato in precedenza, l'uso dell'inganno da parte dei robot sociali solleva importanti questioni etiche, sociali e legali. Tra le principali preoccupazioni vi sono la trasparenza, la fiducia, che abbiamo visto essere potenzialmente lesa, e il rischio stesso che l'inganno si trasformi in manipolazione. Idealmente, gli utenti dovrebbero essere consapevoli della capacità del robot di fornire risposte non letterali e comprendere le intenzioni alla base del suo design. Le considerazioni etiche sull'uso di tecniche di inganno da parte dei robot richiedono quindi un'attenta valutazione degli scopi per cui queste vengono usate, dai benefici per gli utenti ai potenziali rischi o svantaggi. Mentre è necessario chiarire alcuni aspetti e definizioni legati all'Articolo 5 del AI Act, riteniamo che sia possibile mitigare il possibile calo di fiducia dovuto a comportamenti ingannevoli e gli effetti negativi dell'inganno stesso sull'interazione, dotando il robot di tecniche che rendano più trasparenti gli intenti e gli scopi del robot. In *[blind-reference]*[50] abbiamo proposto, inoltre, che meccanismi di *leaking* a livello di comunicazione superficiale possano essere utilizzati anche per rendere i meccanismi di inganno dello stato interno più trasparenti e evidenti. Suggeriamo, quindi, di dotare i robot di comportamenti di *leaking* apparentemente accidentali (non strategici) che rivelino il loro inganno in modo sottile, mantenendo così un equilibrio tra inganno e trasparenza. Questa strategia mira a creare una forma di inganno che sia simile a quella umana e, al tempo stesso, trasparente. L'inganno superficiale, una caratteristica dei sistemi di robotica sociale, sembra essere un meccanismo di comunicazione i cui benefici si manifestano in termini di aumento del coinvolgimento e della naturalezza della comunicazione. Inoltre, tali meccanismi possono mitigare la perdita di fiducia in caso di interazioni con errori (intenzionali o meno). Tuttavia, è necessaria un'ulteriore ricerca per investigare i possibili segnali verbali e non verbali che un robot ingannevole può usare per rendere le sue intenzioni più trasparenti.

[50] R. Esposito, A. Rossi, M. Ponticorvo, S. Rossi, *RoboLeaks.: Non-strategic Cues for Leaking Deception in Social Robots*, in Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25). IEEE Press, 2025, pp. 1111-1120.

Conclusioni

La crescita tecnologica dei robot, la loro integrazione negli ambienti umani e la loro capacità di adottare comportamenti ingannevoli presentano sia opportunità che sfide. Gli effetti dell'inganno pro-sociale, in cui le azioni ingannevoli del robot sono dirette per il beneficio dell'interlocutore umano, hanno ricevuto poca attenzione. L'inganno prosociale può influenzare positivamente il comportamento umano e il coinvolgimento in determinati contesti. La ricerca futura, quindi, dovrebbe concentrarsi sull'identificazione delle condizioni precise in cui l'inganno prosociale migliora i risultati umani, in particolare la fiducia, e sui contesti in cui il comportamento ingannevole del robot potrebbe avere effetti collaterali indesiderati.

La ricerca sui comportamenti ingannevoli nei robot sociali è collegata anche al tema più ampio dell'etica della robotica e dello sviluppo di agenti artificiali morali. Man mano che i robot si integrano sempre più nei contesti sociali, il loro design e la loro programmazione entrano in contatto con questioni di moralità, autonomia e con il contratto sociale tra esseri umani e macchine. Tali dilemmi etici trovano riscontro nella generale esigenza di stabilire linee guida e regolamentazioni. A causa della complessità e dei potenziali effetti sia positivi che negativi dell'inganno nell'interazione tra esseri umani e robot, è fondamentale avviare un dibattito su come i robot che usano tecniche ingannevoli vengono programmati e utilizzati nella società, quali sono gli effetti di tali robot sulla percezione delle persone e quali sono le dinamiche tra esseri umani e robot. I ricercatori devono considerare attentamente le implicazioni dell'uso di tali comportamenti ingannevoli. Bilanciare i benefici di un'interazione migliorata con i rischi legati alla perdita di fiducia è fondamentale per lo sviluppo di sistemi robotici efficaci.

Bibliografia

E. Adar, D.S. Tan, J. Teevan, *Benevolent deception in human computer interaction*, in "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", *CHI '13: CHI Conference on Human Factors in Computing Systems*, 2013 New York, NY, USA: ACM. Available at: https://doi.org/10.1145/2470654.2466246

G. Angelopoulos, *et al.*, *Using theory of mind in explanations for fostering transparency in human-robot interaction*, in "Lecture Notes in Computer Science", Springer Nature Singapore (Lecture notes in computer science), Singapore, 2024, pp. 394-405

R.C. Arkin, P. Ulam, A.R. Wagner, *Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception*, "Proceedings of the IEEE. Institute of Electrical and Electronics Engineers", 100(3), 2012, pp. 571-589.

A.M Aroyo, *et al. Overtrusting robots: Setting a research agenda to mitigate overtrust*

in automation, "Paladyn: journal of behavioral robotics", 12(1), 2021, pp. 423-436.
A. Ayub, A. Morales and A. Banerjee, *Using Markov Decision Process to Model Deception for Robotic and Interactive Game Applications*, 2021 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2021, pp. 1-6, doi: 10.1109/ICCE50685.2021.9427633.
J.B. Bell, *Toward a theory of deception*, "International journal of intelligence and counterintelligence", 16(2), 2003, pp. 244-279.
T. Belpaeme, *et al.*, *Social robots for education: A review*, "Science robotics", 3(21). 2018Available at: https://doi.org/10.1126/scirobotics.aat5954.
J. Bowyer Bell, B. Whaley, *Cheating And Deception*, pp. 479 Routledge, 10.4324/9781315081496, 1991;
D.B. Buller, *et al.*, *Testing interpersonal deception theory: The language of interpersonal deception*, "Communication theory: CT: a journal of the International Communication Association", 6(3), 1996, pp. 268-289.
R. Carli, A. Najjar, *Reconsidering deception in social robotics: The role of human vulnerability (student abstract)*, "Proceedings of the … AAAI Conference on Artificial Intelligence", AAAI Conference on Artificial Intelligence, 37(13), 2023, pp. 16174-16175.
M.C. Carter, *et al.*, *Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial*, "Journal of medical Internet research", 15(4), 2013, p. e32.
C. Castelfranchi, *Ethics and information technology*, 2(2), 2000, pp. 113-119; A. Sharkey, N. Sharkey, *'We need to talk about deception in social robotics!*, *"Ethics and information technology*", 23(3), 2021, pp. 309-316.
R.M. Chisholm, T.D. Feehan, *The intent to deceive*, "The journal of philosophy", 74(3), 1977, p. 143.
M. Coeckelbergh, *Artificial companions: Empathy and vulnerability mirroring in human-robot relations*, "Studies in ethics, law, and technology", 4(3), 2011. Available at: https://doi.org/10.2202/1941-6008.1126.
J. Danaher, *Robot Betrayal: a guide to the ethics of robotic deception*, *"Ethics and information technology*", 22(2), 2020, pp. 117-128.
E. de Oliveira, L. Donadoni, S. Boriero, S. et al. *Deceptive Actions to Improve the Attribution of Rationality to Playing Robotic Agents*. Int J of Soc Robotics 13, 391-405 (2021). https://doi.org/10.1007/s12369-020-00647-8.
B.M. DePaulo, *et al.*, *Lying in everyday life*, "Journal of personality and social psychology", 70(5), 1996pp. 979-995.
C. Di Napoli, G. Ercolano, S. Rossi, *Personalized home-care support for the elderly: a field experience with a social robot at home*, "User Model User-Adap Inter 33", 2023, pp. 405-440. https://doi.org/10.1007/s11257-022-09333-y
A. Dragan, R. Holladay, S. Srinivasa, *Deceptive robot motion: synthesis, analysis and experiments*, "Autonomous robots", 39(3), 2015, pp. 331-345.
I. Deweese-Boyd, *Self-Deception,* "The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab", Stanford University, 2023, https://plato.stanford.edu/archives/fall2023/entries/self-deception/.S. Erat, U. Gneezy, *White lies*, "Management science", 58(4), 2012, pp. 723-733.
R. Esposito, A. Rossi, S. Rossi, *Deception in HRI and Its Implications: A Systematic Review*, in "J. Hum.-Robot Interact", 14, 3, Article 47, 2025. https://doi.

org/10.1145/3721297
R. Esposito, A. Rossi, M. Ponticorvo, S. Rossi. *RoboLeaks.: Non-strategic Cues for Leaking Deception in Social Robots*, in Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25). IEEE Press, 2025, pp. 1111-1120.
D. Laparidou, *et al.*, *Patient, carer, and staff perceptions of robotics in motor rehabilitation: a systematic review and qualitative meta-synthesis*, "Journal of neuroengineering and rehabilitation", 18(1), 2021, p. 181.
N. Lee, *et al. The influence of politeness behavior on user compliance with social robots in a healthcare service setting*, "International journal of social robotics", 9(5), 2017, pp. 727-743;
G. Maggi, *et al. "don't get distracted!": The role of social robots' interaction style on users' cognitive performance, acceptance, and non-compliant behavior*, "International journal of social robotics", 13(8), 2021, pp. 2057-2069.
A. van Maris, *et al.*, *Designing Ethical Social Robots-A Longitudinal Field Study With Older Adults*, "Frontiers in robotics and AI", 7, 2020, p. 1.
A. Matthias, *Robot Lies in Health Care: When Is Deception Morally Permissible?*, "Kennedy Institute of Ethics journal", 25(2), 2015, pp. 169-192;
G. Mois, J.M Beer, *Robotics to support aging in place*, in "Living with Robots", Elsevier, 2020, pp. 49-74;
K. Rogers, R.J.A Webber, A. Howard, *Lying about lying*, in "Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA 2023: ACM. Available at: https://doi.org/10.1145/3568294.3580178.
A. Rosero, *Using justifications to mitigate loss in human trust when robots perform norm – violating and deceptive behaviors*, in "Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA(2023): ACM. Available at: https://doi.org/10.1145/3568294.3579979
Rossi, A. *et al.*, *How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario*, in "*Social Robotics*", Cham: Springer International Publishing (Lecture notes in computer science), 2017, pp. 42-52.
A. Rossi, S. Rossi, *Evaluating people's perception of trust of a deceptive robot with theory of mind in an assistive gaming scenario*, in "2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)", 2023 IEEE. Available at: https://doi.org/10.1109/ro-man57019.2023.10309647.
A. Rossi, S. Rossi, *On the way to a transparent HRI*, in "Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization", UMAP '24: 32nd ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA 2024: ACM. Available at: https://doi.org/10.1145/3631700.3664890.
H. S. Sætra, *Social robot deception and the culture of trust*, "Paladyn: journal of behavioral robotics", 12(1), 2021, pp. 276-286;

A. Sharkey, N. Sharkey, *We need to talk about deception in social robotics!*, "Ethics and information technology", 23(3), 2021, pp. 309-316.
H.S. Sætra, *Machiavelli for robots: Strategic robot failure, deception, and trust*, in "2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)". 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE. Available at: https://doi.org/10.1109/ro-man57019.2023.10309455.
A. Sharkey, N, Sharkey, *Granny and the robots: ethical issues in robot care for the elderly*, "Ethics and information technology", 14(1), 2012, pp. 27-40.
A. Sharkey, N. Sharkey, *We need to talk about deception in social robotics!*, "Ethics and information technology", 23(3), 2021, pp. 309-316.
J. Shim, R.C. Arkin, R.C. *A taxonomy of robot deception and its benefits in HRI*, in "2013 IEEE International Conference on Systems, Man, and Cybernetics", 2013 IEEE International Conference on Systems, Man and Cybernetics (SMC 2013), IEEE. Available at: https://doi.org/10.1109/smc.2013.398.
J. Shim, R.C. Arkin, *Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark*, in "2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)". 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2014, IEEE. Available at: https://doi.org/10.1109/robio.2014.7090385..
J. Shim, R.C. Arkin, *The benefits of robot deception in search and rescue: Computational approach for deceptive action selection via case-based reasoning*', in "2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)", 2015 IEEE. Available at: https://doi.org/10.1109/ssrr.2015.7443002.
J. Shim, R.C. Arkin, *Other-oriented robot deception: How can a robot's deceptive feedback help humans in HRI?*, in "*Social Robotics*", Cham: Springer International Publishing (Lecture notes in computer science), 2016, pp. 222-232, Lecture Notes in Computer Science(), vol 9979. Springer, Cham. https://doi.org/10.1007/978-3-319-47437-3_22
E. Short, *et al. No fair!! An interaction with a cheating robot*, in "2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)", 2010 IEEE. Available at: https://doi.org/10.1109/hri.2010.5453193.
A. Tapus, M. Mataric, B Scassellati, *Socially assistive robotics [Grand Challenges of Robotics]*, "IEEE robotics & automation magazine", 14(1), 2007, pp. 35-42; L. Wijnen, J. Coenen, B. Grzyb, *'It's not my Fault!'*, in "Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction", HRI '17: ACM/IEEE International Conference on Human-Robot Interaction, New York, NY, USA, 2017: ACM. Available at: https://doi.org/10.1145/3029798.3038300.
G. Wilson, *et al.*, *Robot-Enabled Support of Daily Activities in Smart Home Environments*, *"*Cognitive systems research*"*, 54, 2019, pp. 258-272.
K. Winkle, *et al.*, *Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots*, in "Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21: ACM/IEEE International Conference on Human-Robot Interaction", New York, NY, USA 2021: ACM. Available at: https://doi.org/10.1145/3434073.3444666

Valeria Seidita[*], *Rosario Bongiorno*[*], *Daniele Franco*[*], *Alessandro Giambanco*[*], *Gianni Randazzo*[*], *Antonio Pio Sciacchitano*[*], *Antonio Chella*[*]

Robotics and AI for social justice: new perspectives between assistance, education and ethical reflection

Abstract
Artificial intelligence and robotics are changing the way we approach societal challenges, offering new opportunities to improve inclusion, access to care and ethical reflection. Although concepts such as introspection, consciousness and ethical reflection are the subject of wide philosophical debate, the aim of this paper is not to discuss them theoretically, but to explore how some of their dimensions can be captured in simplified computational models, with practical applications in the context of AI and robotics. This article explores the potential of these technologies to promote social justice. Two main application areas are analyzed: assisting vulnerable individuals, such as the elderly and people with disabilities, and the ability of robots to promote awareness and change in educational contexts. Experiments in the healthcare and school sectors will be used to demonstrate how technology can be used to optimize services and to stimulate critical reflection on the dynamics of exclusion and acceptance. However, the potential of AI and robotics comes with significant ethical challenges, such as the risk of bias and new forms of surveillance. The work emphasizes the importance of a responsible approach to technological development so that innovation can be guided by principles of justice and inclusion rather than purely economic or efficiency logics.

Keywords
Human-robot interaction; Assistive Robotics; Social Justice; Healthcare.

Introduction

In today's context, the definition of vulnerability has taken on increasingly complex nuances that go beyond the simple condition of economic or physical vulnerability. Being vulnerable today means being disadvantaged in terms of access to important resources – education, health, job opportunities – but also in terms of the ability to assert one's rights in a society where inequalities are reinforced rather than reduced.

[*] Università degli Studi di Palermo.

SIFM
FM/MP *Filosofia morale/Moral Philosophy*, vol. 1, 2025 • ISBN: 9791222322681 • DOI: 10.7413/fmmp0073

People with disabilities, older people, migrants, people with mental disorders or people in disadvantaged economic circumstances are just some of the categories at risk of being excluded from social and decision-making processes. But vulnerability is not just an individual condition: it is often the product of a system that, despite consolidated welfare mechanisms, is not always able to guarantee equal opportunities for all.

This reality concerns society's perception of diversity and the way in which it is welcomed or rejected is not merely a question of access to resources. We had the opportunity to explore this aspect in a project with a middle school class in which we used a robot to initiate a reflection on the meaning of diversity. It was surprising to observe how the students began to see 'difference' not as a threat but as a resource, as an opportunity for the growth of the whole group. The robot itself, which was initially seen as a foreign element, became the catalyst for a change of perspective: what if diversity was not an obstacle, but an opportunity for enrichment?

At a time when technological progress has become the main driver of social change, one of the key questions is: can technology be a means of reducing these inequalities, or does it run the risk of exacerbating them?

So far, artificial intelligence and robotics[1,2] have been developed primarily with the aim of increasing efficiency, shortening working hours and simplifying production processes. However, their potential goes far beyond this: if used consciously and responsibly, they can become important tools to improve well-being and promote the inclusion of those population groups that are more at risk of being marginalized than others.

The use of AI to support healthcare decisions[3,4], the use of robotics to improve the care of the elderly[5,6] or patients with special needs[7], the development of intelligent education systems that can adapt to the different

[1] S.J. Russell-P. Norvig, *Artificial intelligence: a modern approach*, Pearson, London 2016.
[2] R.A. Brooks, *New approaches to robotics, "*Science", 253(5025), 1991, pp.1227-1232.
[3] Ala'a M. Al-Momani, *Adoption of artificial intelligence and robotics in healthcare: a systematic literature review*. "International Journal of Contemporary Management and Information Technology" (IJCMIT), 3(6), 2023, pp.1-16.
[4] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H.R. Nambiappan, K.K Chaitanya, A. R. Babu, J. Mathew, F. Makedon, *A survey of robots in healthcare,* "Technologies", 9(1), 2021, p.8.
[5] A. Vercelli, I. Rainero, L. Ciferri, M. Boido, F. Pirri, *Robots in elderly care*, "DigitCult-Scientific Journal on Digital Cultures", 2(2), 2018, pp.37-50.
[6] R. Bemelmans, G.J. Gelderblom, P. Jonker, L. De Witte, *Socially assistive robots in elderly care: a systematic review into effects and effectiveness,* "Journal of the American Medical Directors Association", 13(2), 2012, pp.114-120.
[7] G.A. Papakostas, G.K. Sidiropoulos, C.I. Papadopoulou, E. Vrochidou, V.G. Kaburlasos, M.T. Papadopoulou, V. Holeva, V.A. Nikopoulou, N. Dalivigkas, *Social robots in special education: A systematic review,* "Electronics", 10 (12), 2021, p.1398.

needs of students[8]: all these scenarios show how technology can be a real breakthrough in dealing with vulnerability. However, the question is not just a technical one, but a deeply ethical one: how can we design and use these tools in such a way that they truly contribute to social justice and do not become an additional tool of discrimination or exclusion?

From this perspective, addressing the issue of vulnerability means recognizing its many forms and primarly understanding how technology – and AI and robotics in particular – can play a crucial role in promoting a more just and inclusive society.

Technologies for equity in access to care and support for the most vulnerable

In recent years, scientific and technological research has paid increasing attention to improving healthcare systems to ensure equitable access to care and adequate support for the constantly aging population[9,10], in terms of efficiency and economic sustainability. The calls for proposals of recent European projects show how central the issue of public health is: from the digitalization of healthcare facilities to telemedicine, from assistive robotics to artificial intelligence used for early diagnosis and personalization of treatments.

One aspect that is often underestimated when it comes to improving healthcare is the well-being of those who administer it: doctors, nurses, social and health workers. A healthcare system that cares for patients is a system that must first and foremost care for those who work in it. Overcrowded hospitals, unsustainable workloads, burnout among healthcare professionals: these factors undermine the quality of care and the ability of a healthcare system to function equitably and inclusively[11]. Technological support could represent a turning point in this sense: the use of AI to automate administrative and bureaucratic tasks, robotics to reduce the physical burden on staff, remote monitoring systems to reduce unnecessary hospital visits.

In addition to the well-being of healthcare professionals, there is also a fact that requires urgent reflection: Our society is aging. According to

[8] C. Syriopoulou-Delli, E. Gkiolnta, *Robotics and inclusion of students with disabilities in special education,* "Research, Society and Development", 10(9), 2021, pp. 1-11.
[9] W.C. Mann, *The aging population and its needs,* "IEEE Pervasive Computing", 3(2), 2004, pp.12-14.
[10] M.E. Pollack, *Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment,* "AI magazine", 26(2), 2005, pp. 9-24.
[11] K. Doulougeri, K. Georganta, A. Montgomery, *"Diagnosing" burnout among healthcare professionals: can we find consensus?,* "Cogent Medicine", 3(1), 2016, pp. 1-10.

WHO[12,13] estimates, the number of older people with chronic diseases will increase exponentially in the coming decades, putting unprecedented pressure on healthcare systems. This means more patients to treat and a greater need for long-term care. And this is where technology comes in: advanced home automation solutions, robotic assistants for home monitoring and support, AI for personalized management of treatment plans could help make healthcare more sustainable by enabling people to age in their own environment, reducing avoidable hospital admissions and improving quality of life.

It's not just about surviving longer, it's about aging well. Healthy aging is a key challenge for the future of our society, which is becoming older, lonelier and less cared for. This is health issue and at the same time a social and cultural one. Families are not always able to care for older people, whether for economic reasons, work dynamics or changing lifestyles. Technology can fill some of these gaps and create new spaces for sociality and support: from companion robots that encourage interaction and cognitive maintenance, to digital platforms that connect older people and caregivers, to virtual reality tools that combat isolation.

And the same goes for other vulnerable categories, such as families with autistic children or with learning disabilities. Today, there is much talk of an increase in cases of intellectual disability and autism: perhaps because we finally have the tools to diagnose them, perhaps because society itself has changed, setting rhythms and models that make the need for support clearer.

Whatever the cause, the fact is that more and more families are confronted with parenting and relationship problems for which the traditional system does not always have adequate answers.

And here we come back to the central point: what kind of society are we building? Are we moving towards a more inclusive model that responds to people's needs, or are we just creating new gaps, new forms of exclusion? In the face of these changes, we cannot afford to be passive bystanders. Technological innovation, when guided by the values of equality and inclusion, can become a powerful tool to respond to these new challenges. It is important to emphasize that it is not about replacing human relationships with machines, but about using technology to strengthen the social fabric and create new forms of support, connection and participation.

So the question is not whether technology can help us, but how we want it to do so. What principles should guide us in developing solutions to help? What ethical boundaries should we set? We are facing one of the greatest challenges of our time: we need to rethink our idea of care and community and then to ensure equitable access to health.

[12] https://www.who.int/health-topics/ageing#tab=tab_1
[13] https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

Technology, rights and inclusion: a bridge to equity

When we talk about equal access to healthcare and a fairer healthcare system, we inevitably have to address the issue of rights. Health is a fundamental right, but we know that it is not always guaranteed for everyone under the same conditions. There are economic, cultural, logistical and even cognitive barriers that prevent many people from getting the care they need. People who suffer from a disability, for example, must have the right to the same care as others and to access methods that take their special needs into account. The same is true for people with autism spectrum disorders, for older people with cognitive difficulties, for those who come from disadvantaged backgrounds or are in a state of social vulnerability.

In this perspective, technology also and above all becomes a means of reflecting on these issues and educating people to make more ethical and responsible choices. This is where artificial intelligence and robotics can come into play, as practical tools and as means that can promote new forms of awareness and integration.

Consider, for example, the interaction between humans and robots. A robot can be programmed to respond impartially, to suggest scenarios that present the interlocutor with ethical dilemmas, and to accompany them in a process of critical reflection on their own choices. Imagine a robot assistant in a hospital or school: it could perform supportive functions and at the same time be designed to guide users along the path of sensitization and awareness of rights and inclusion. A child interacting with a robot could be made to think about the importance of treating all people fairly, a doctor could be helped to navigate complex clinical decisions, a vulnerable person could find in AI a means of making their voice heard.

But how can this idea be translated into something concrete? Which technologies can actually help to create a fairer and more inclusive society? To answer these questions, it is worth examining some experiences where artificial intelligence and robotics have been used in real-life scenarios, both in the health and education sectors.

On the one hand, we have seen how the use of robots and AI in healthcare can provide innovative solutions to support patients and medical staff. On the other hand, we have observed how technology can be used to stimulate ethical reflection, as in the case of the experiment conducted with a school class, where a robot was the starting point for a discussion on diversity, inclusion and acceptance.

In the next sections, we will analyze these two scenarios in more detail and try to understand the potential of the technology used and mainly the challenges and implications of its responsible use.

Artificial intelligence and robotics for ethical and inclusive interaction

When it comes to the application of artificial intelligence and robotics in society, there is a danger that the discussion will be reduced to a question of automation and optimization. But there is a much deeper aspect that research is trying to explore: can the technology help and also stimulate ethical considerations, promote well-being and support social integration and inclusion?

In our research, we focus on two directions that address these questions. On the one hand, we are developing solutions to support healthy aging and the wellbeing of the most vulnerable, such as the elderly and patients with special care needs. On the other hand, we are exploring how artificial intelligence can be equipped with a mechanism of ethical introspection, capable of critically evaluating its own decisions and positively influencing those who interact with it.

Assistive robotics for well-being and healthy aging

The growth of the elderly population and the increasing pressure on healthcare systems require new solutions that can support patients, caregivers and medical staff. One of our research focuses is the development of robotic systems that are capable of interacting with users in a personalized way, adapting to their needs and promoting behaviors that enhance well-being.

An innovative aspect of this research is the integration of Large Language Models (LLMs)[14,15] to make interaction with the robot more natural and accessible. The latest generation of language models makes it possible to improve the understanding of context and the system's ability to respond flexibly and contextually. However, their use is not without its challenges: LLMs can make the robot more autonomous in conversation, but they can also present problems related to the generation of unpredictable responses, the possibility of reinforcing biases, and the difficulty of ensuring absolute reliability of information. A central point of our research is to integrate these tools into a controlled context in which the robot is not limited to generating answers, but can critically evaluate its statements and justify them in a transparent way.

[14] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du,. *A survey of large language models, arXiv preprint arXiv:2303.18223* 1, no 2 2023.

[15] A.J., Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W.Ting, *Large language models in medicine, "Nature medicine"*, 29(8), 2023, pp.1930-1940.

In addition to the use of language models, we are also exploring the potential of multi-agent systems[16,17], i.e. architectures in which several artificial intelligences work together to achieve common goals. This approach makes it possible to design more dynamic and adaptable environments in which the assistant robot does not act as an isolated entity but interacts with other agents, e.g. telemedicine platforms, environmental sensors and decision support systems for doctors and nursing staff. The use of multi-agent systems opens up interesting perspectives for the management of home care and for the personalization of care and also brings new challenges in terms of coordination, communication and safety between the various components of the system.

An artificial intelligence capable of ethical introspection

If, on the one hand, assistive robotics aims to improve people's well-being through active support, a more theoretical but equally crucial aspect concerns the possibility of endowing machines with a form of ethical introspection.

We wonder whether it is possible to develop an artificial intelligence that is capable of critically evaluating its own decisions and changing its own decisions on the basis of an internal reflection process. Such a system would not only execute predefined rules, but would be able to check, update and even change its range of values depending on the situation.

Integration with advanced language models can also play an important role here. A robot reasoning about its ethical decisions must "think" and be able to communicate its reasoning in a clear and understandable way. LLMs can be used to improve this ability, allowing the robot to express and justify its decisions in a way that is easier for humans to understand. However, the biggest challenge remains controlling the consistency and reliability of the explanations provided by the robot: the ethical reflection of an intelligent system cannot be based on statistically probable answers, but must be guided by structured and verifiable principles.

Another research direction we are investigating is the use of specialized artificial agents within a multi-agent system, in which different AI components work together to make more complex ethical decisions. This approach distributes the computational load and increases the reliability of the system, but also raises questions about the role of coordination be-

[16] W.Van der Hoek, M. Wooldridge, *Multi-agent systems,* "Foundations of Artificial Intelligence", 3, 2008, pp. 887-928.
[17] A. Dorri, S.S. Kanhere, R. Jurdak, *Multi-agent systems: A survey.* "Ieee Access", 6, 2018, pp.28573-28593.

tween the agents: to what extent can we allow robots to negotiate ethical solutions with each other? In what contexts is it acceptable to delegate some of the ethical considerations to an AI system?

From theory to practice: the value of research

This research is not just a theoretical exercise, but lays the groundwork for real-world applications that could change the way we interact with technology. When it comes to robotics and artificial intelligence, there is a danger of viewing these systems as mere entertainment products or advanced automation tools. However, the reality is very different.

Creating a system that is capable of interacting meaningfully with humans means tackling complex problems of cognitive engineering, computational ethics and artificial intelligence. The key is to design architectures that incorporate psychological models, learning strategies and adaptation mechanisms, rather than simply writing code.

The integration of LLMs and multi-agent systems opens up new perspectives, but also raises profound questions about reliability, control and social implications. A robot that can explain its behavior can contribute to greater transparency and trust, but only if its motivations are truly understandable and verifiable. Similarly, an AI that collaborates with other agents can provide more advanced solutions, but it also introduces new levels of complexity that need to be carefully managed.

The goal of our research is to understand how we can use new technologie to create a fairer and more inclusive society and to develop them. This requires a constant dialog between science, ethics and society so that innovation is guided by technological efficiency and then, above all, by a conscious reflection on its impact.

Technology and reflection on the different: the robot as a catalyst for awareness

When we think about the use of robotics and artificial intelligence in the field of inclusion, the focus is often on their ability to support the daily lives of people with difficulties by providing tools to improve communication, learning or autonomy. However, there is another, less explored but equally important perspective: the ability of these tools to stimulate deeper thinking about the concept of diversity and acceptance.

Based on this idea, we decided to test the impact of a robot in an educational context to find out how its presence could influence the way children perceive “otherness”. The starting point was a concrete need: to

make a statement about how robotics can be used to promote the integration of children with autism spectrum disorders within the classroom community. We decided to involve a middle school class in our city and turn the experience into a theater workshop, with the robot as one of the main characters in the scene.

The idea, which developed naturally in discussion with the children, was to tell a story about bullying in which the "other" was the autistic child in the class and the robot itself. In this way, we were able to approach the topic of exclusion from a particular perspective: the robot, as a foreign element, was immediately recognizable as different, but its presence gave greater expression to the perception of diversity within the group itself. In the story the boys wrote, the robot could only make friends with the autistic child and with a classmate who had just arrived from another school, both of whom were perceived as "strangers" by the other members of the class. This parallel has made it clear that the concept of diversity is often arbitrary and based on social patterns rather than actual incompatibilities.

From a technological point of view, the experiment posed no particular challenges: the interaction between the robot and the children took place naturally, without the need for complex interventions. But that is precisely the point: it was not the technology itself that was decisive for the experience, but the meaning that its presence had in this context. The robot wasa catalyst for thought and not merely a tool, an element that could draw attention to issues that would have remained in the background in a normal situation.

What made the experiment particularly interesting was the way in which the students worked out the meaning of diversity. The robot, which is free of prejudice and social patterns, allowed them to look at the problem with a fresh eye. The story they built ended with the bullies realizing at the end of the performance that the most important lesson they ever received was that they changed their minds.

This leads to a wider reflection. We often get so used to the diversity that surrounds us that we no longer notice it: older people, disabled people, people with learning difficulties are part of our world, but their condition no longer inspires active reflection. A foreign element, such as a robot, can instead draw attention to these issues, precisely because it represents a "new other", something that does not fit into our usual patterns. It is paradoxical, but it is as if in order to really think about inclusion, we need yet another Other that can confront us with our own mechanisms of acceptance and rejection.

This experience has shown that the use of technology is not limited to offering practical solutions for integration, but can also be a means to raise awareness and change perceptions of diversity. The robot in

this case was not a simple assistant, but a tool that encouraged ethical reflection, a means by which the children could examine and redefine their prejudices.

This prospect opens up interesting scenarios for the future. If technology can help us to reflect on who we are and how we treat others, then its role goes far beyond simple assistance: it can become a tool that improves our ability to live with diversity, to accept it and, above all, to recognize its value.

Conclusions

The analysis carried out in this thesis has made it clear that artificial intelligence and robotics are not just technical tools, but can play a deeper role in building a fairer and inclusive society. If used consciously, these technologies can help bridge social divides, ensure more equitable access to basic services such as health and education, and promote ethical reflection on the dynamics of exclusion and vulnerability.

The opportunities arising from the integration of technology into the social sphere are manifold. The use of robots and artificial intelligence in the care of the elderly, patients with special needs or students with learning difficulties can be a decisive factor in improving the quality of life and strengthening the rights of those who are often marginalized. In addition, the ability of technology to stimulate ethical reflection through interaction and self-explanation opens up new perspectives and transforms intelligent systems into real catalysts for awareness.

However, these opportunities must be balanced by a careful consideration of the ethical challenges they present. Technology is not neutral: the choices made in its development and implementation can reinforce existing inequalities or create new forms of exclusion. The risk of bias in AI systems, the loss of privacy and the increasing use of algorithmic surveillance are issues that cannot be ignored. Adopting a responsible approach to technological development is therefore crucial to prevent these tools from becoming tools of control rather than emancipation.

Ultimately, the relationship between social justice and new technologies is characterized by a delicate balance between potential and responsibility. AI and robotics can offer exceptional tools to promote greater justice, but only if they are accompanied by constant ethical reflection and a concrete commitment to ensure that their use does not reinforce inequalities, but reduces them.