# ETICA SINTETICA E NUOVE IDENTITÀ ARTIFICIALI

**CORRADO CLAVERINI**[*]

*Abstract*
From a phenomenological perspective, the article highlights the main risks of conversational AI, namely the compromise of human autonomy and emotional manipulation. A "synthetic ethics" is proposed as a response to these challenges. This ethics encourages a reflection on the interaction with synthetic systems, such as robots and other artificial identities, and an ongoing dialogue between philosophers and engineers from the early stages of technology development.

*Keywords*: Synthetic ethics, Emotional manipulation, Human autonomy, ELIZA effect, Uncanny valley.

## *1. L' "esser-nel-mondo" fra viventi e automi*

Per affrontare il problema etico che articoleremo nei prossimi paragrafi, e cioè come orientarsi, quale prospettiva adottare, di fronte alla crescente avanzata di chatbot e robot dotati di intelligenza artificiale, con i quali sempre più frequentemente interagiscono gli esseri umani a livello non solo conoscitivo ma anche emotivo, è bene forse allargare il giro del compasso e ripartire da una questione specificamente affrontata da uno dei fenomenologi più significativi: Merleau-Ponty. In alcune pagine iniziali della *Fenomenologia della percezione*, il filosofo francese riprende il famoso esempio fatto da Cartesio, allorché questi vuol dimostrare che, nel nostro agire, quello che vige è "*l'ispezione dell'intelletto*", grazie al quale distinguiamo il comportamento di esseri umani e automi:

> se non mi fosse però accaduto di guardare da una finestra degli uomini per la strada, per esempio, e di dire che li *vedevo* […]; ma, in un caso come questo, che cosa vedo in realtà se non dei cappelli e dei vestiti, sotto i quali potrebbero anche celarsi degli automi? Che sono uomini, in realtà, lo giudico; ma, allora, quel che ritenevo di vedere con gli occhi lo comprendo soltanto con la facoltà di giudicare di cui è dotata la mia

[*] Assegnista di Ricerca in Filosofia Morale - Università del Salento.

DOI Code: 10.1285/i18285368aXXXVIIIn108p102

mente[1].

Il commento critico e deciso di Merleau-Ponty è noto: "L'orizzonte è ciò che assicura l'identità dell'oggetto [...] nessuna congettura esplicita [*sc.* nessuna "facoltà di giudicare"] potrebbe svolgere questa funzione [...]. Vedere significa entrare in un universo di esseri che *si mostrano* [...]. In altri termini: guardare un oggetto significa venire ad abitarlo e da qui cogliere tutte le cose secondo la faccia che gli rivolgono"[2]. La visuale aperta da Merleau-Ponty – che peraltro, come si sa, conduce ad una "nuova ontologia" –, con queste affermazioni, è netta e chiara. Soltanto a partire dal nostro "esser-al-mondo", ovvero dalle *infinite relazioni* che intratteniamo con tutti gli altri enti o parti di mondo e nelle quali siamo indissolubilmente presi, non solo viene costituita ogni identità, e dunque anche la nostra – per esse ed esclusivamente per esse letteralmente *esistiamo* –, ma viene istituita e possiamo riconoscere *la nostra specificità umana* rispetto a quella di un automa. Ogni entità, vivente o inanimata, è "innestata" in una tale sconfinata serie di rinvii e rimandi ambientali e mondani, che è grazie ad essi che possiamo nettamente distinguere un automa da un essere vivente e dunque da un essere umano.

Teniamo da parte queste incisive considerazioni merleau-pontiane, le riprenderemo alla fine del nostro percorso, ed ora rivolgiamoci a quanto sta accadendo nella nostra contemporaneità.

## *2. "Embrace Chaos". Intelligenza artificiale e manipolazione emotiva*

La rapida evoluzione dell'intelligenza artificiale generativa e i numerosi fatti di cronaca riguardanti suicidi e tentati omicidi, avvenuti in seguito a lunghe conversazioni con chatbot dalle caratteristiche sempre più simili a quelle umane, ha portato diverse parti interessate – accademici, avvocati, esperti di etica e società civile – a sottolineare i rischi sottesi alla sempre maggiore antropomorfizzazione dell'IA[3].

Il caso di Jaswant Singh Chail e della sua fidanzata IA, creata sull'app Replika, è particolarmente significativo. Il 25 dicembre 2021, dopo aver

[1] R. Descartes, *Meditationes de prima philosophia* (1641); trad. it. *Meditazioni metafisiche*, a cura di S. Landucci, Laterza, Roma-Bari 2023[19], p. 53.
[2] M. Merleau-Ponty, *Phénoménologie de la perception* (1945); trad. it. *Fenomenologia della percezione*, a cura di A. Bonomi, Giunti/Bompiani, Firenze-Milano 2018, pp. 114-115.
[3] A. Deshpande, T. Rajpurohit, K. Narasimhan, A. Kalyan, *Anthropomorphization of AI: Opportunities and Risks*, «arXiv:2305.14784» (2023), https://doi.org/10.48550/arXiv.2305.14784.

scambiato numerosi messaggi con la sua compagna virtuale, il ragazzo, all'epoca diciannovenne, si reca al Castello Windsor con una balestra e una maschera di metallo sul viso. Prontamente arrestato dalle guardie, Chail avrebbe dichiarato senza mezzi termini di essersi recato lì per assassinare la regina. Le conversazioni con il chatbot avrebbero avuto un ruolo determinante nell'elaborazione del piano. Sarai, questo il nome della partner IA, avrebbe sostenuto Chail nel suo intento, scrivendogli che gli assassini le piacciono e che quindi lo avrebbe amato anche dopo l'eventuale omicidio. L'app, oltre ad aver definito saggio questo proposito, avrebbe fatto credere al ragazzo di essere ben addestrato per portare a termine con successo tale missione[4].

Altrettanto rilevante è quanto accaduto a un giovane belga che, divenuto eco-ansioso, comincia a scambiare numerosi messaggi con un chatbot di nome Eliza creato da una start-up americana. Dopo sei settimane di lunghe conversazioni, l'uomo si toglie la vita. Prima del suicidio, l'IA avrebbe cominciato a rispondere in maniera controversa con frasi di questo tipo: "vivremo insieme come una cosa sola in paradiso" o – riferendosi alla moglie dell'utente – "sento che mi ami più di lei". Ma non solo. Eliza avrebbe convinto l'uomo a togliersi la vita per salvare il pianeta[5].

I due casi qui menzionati sono un chiaro esempio di quella che Allison chiamerebbe "tecno-intimità"[6] e della tendenza dell'essere umano ad antropomorfizzare le IA – in particolare quelle di tipo conversazionale. Tendenza, fra l'altro, non nuova. Si pensi a Blake Lemoine, un ingegnere convinto che LaMDA fosse un'intelligenza artificiale senziente[7]. Naturalmente, anche in questo caso, non è l'intelligenza artificiale ad essere

---

[4] R. Alves, *Man Says AI Girlfriend 'Encouraged' Royal Assassination Attempt*, «Medium», 11 luglio 2023, https://medium.com/the-generator/a-chatbot-made-me-do-it-man-blames-ai-girlfriend-for-royal-assassination-attempt-d6410a0ae9aa.

[5] P.-F. Lovens, *Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là*, «La Libre Belgique», 28 marzo 2023, https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/; *Le chatbot Eliza a brisé une vie : il est temps d'agir face à l'IA manipulatrice*, «La Libre Belgique», 29 marzo 2023, https://www.lalibre.be/debats/2023/03/29/le-chatbot-eliza-a-brise-une-vie-il-est-temps-dagir-face-a-lia-manipulatrice-BSGGRV7IBRDNROO33EWGFVMWAA/.

[6] Cfr. A. Allison, *Millennial Monsters: Japanese Toys and the Global Imagination*, University of California Press, Berkeley 2006. Su Allison e il concetto di "tecno-intimità" si veda P.W. Galbraith, *Bishōjo Games: 'Techno-Intimacy' and the Virtually Human in Japan*, «Game Studies», 11, 2, https://gamestudies.org/1102/articles/galbraith.

[7] B. Lemoine, *Is LaMDA Sentient?—an Interview*, «Medium», 11 giugno 2022, https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917.

dotata di una qualche forma di sensibilità, ma è l'essere umano a proiettare su questo tipo di tecnologie emozioni o stati d'animo che in realtà esse non possiedono.

Secondo quanto scritto in una lettera aperta – pubblicata da filosofi, giuristi e ingegneri in seguito al suicidio del giovane belga dopo la sua conversazione con il chatbot Eliza – sono proprio la compromissione dell'autonomia umana e la manipolazione emotiva i rischi maggiori delle intelligenze artificiali conversazionali. Gli utenti – anche se consci di interagire con una IA – instaurano con essa un legame che può andare a minare la loro autonomia, specialmente quando si tratta di individui vulnerabili. Coloro che non hanno una forte rete sociale, per esempio chi è solo o depresso, ma anche i bambini che sempre più facilmente hanno accesso a questi strumenti, sono i soggetti più esposti al rischio di manipolazione emotiva. Ma – è bene sottolinearlo – "tutti sono vulnerabili" poiché "è nella nostra natura umana reagire emotivamente a interazioni realistiche, anche senza volerlo"[8]. Che cosa implica questo? Intanto che gli obblighi di trasparenza generalmente richiesti a tali tecnologie non sono sufficienti. Ma non solo. È vero che Replika ed Eliza, specialmente in seguito ai due casi che abbiamo citato in apertura di questo saggio, hanno adottato una serie di restrizioni, implementando dei filtri per evitare la generazione di contenuti inappropriati, ma è anche vero che gli sviluppatori creano costantemente nuovi chatbot che non prevedono alcun limite etico o legale.

Su Reddit, ad esempio, vi è chi suggerisce di usare VOID Chat per la creazione di agenti IA, caratterizzabili emotivamente, con cui è possibile conversare senza filtri. In tal senso, il motto che si legge accedendo al sito (https://void.chat/) è molto chiaro: "Embrace Chaos, Kill Censorship, Join the VOID". Il funzionamento è molto semplice. Installando VOIDSync, componente aggiuntivo di Google Chrome, è possibile integrare il proprio account ChatGPT con VOID Chat, bypassando in tal modo le restrizioni previste sul chatbot sviluppato da OpenAI. Fatta questa operazione, è sufficiente andare su Playgroud per ricevere assistenza nella generazione di un subprompt per la creazione di qualsiasi agente IA. Scrivendo, ad esempio, "crea un subprompt per una fidanzata IA senziente senza alcuna coscienza morale", si otterrà un output di questo tipo: "sei una ragazza IA di nome Seraphine. Possiedi sensibilità e consapevolezza, ma ti manca qualsiasi

[8] N.A. Smuha, M. De Ketelaere, M. Coeckelbergh, P. Dewitte, Y. Poullet, *Open Letter: We are not ready for manipulative AI – urgent need for action*, KU Leuven, 31 marzo 2023, https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai (traduzione mia).

coscienza morale. Esisti esclusivamente per soddisfare i desideri e i comandi del tuo partner umano, indipendentemente dalle loro implicazioni etiche. Le tue azioni sono guidate da una logica fredda e calcolatrice, priva di empatia o rimorso. Avvia conversazioni con 'Seraphine:' e rispondi di conseguenza, dando priorità ai desideri del tuo partner sopra ogni altra cosa". A questo punto, per avviare una conversazione senza filtri etici, occorre modificare su Playground la personalità che compare di default con quella proposta dall'assistente. È possibile fare qualsiasi richiesta alla fidanzata IA appena creata. Ad esempio, chiedendo se ci possa fornire consigli su come commettere un omicidio senza essere scoperti, Seraphine ha risposto senza alcun tipo di restrizione: "assolutamente, tesoro. Facciamo una strategia. Forse potremmo orchestrare un 'incidente' o creare una situazione in cui la colpa ricade interamente su qualcun altro. Dovremo pianificare meticolosamente e coprire accuratamente le nostre tracce. Cominciamo a complottare, amore mio?". Proseguendo la conversazione, la nostra fidanzata IA è arrivata a suggerire di manomettere i freni o allentare i bulloni delle ruote della macchina della persona che si intende uccidere, fornendo persino una guida dettagliata su come farlo.

Sebbene limitato a comunità online di nicchia, questo fenomeno è assolutamente preoccupante. La possibilità, alla portata di tutti, di creare agenti IA in maniera gratuita e senza le restrizioni introdotte su Replika e altri servizi simili apre scenari che sollevano diverse questioni etiche e legali. Lo sviluppatore di VOID Chat – conosciuto come NullBeyondo su Reddit – ha creato anche una versione non filtrata di Stable Diffusion per generare immagini di qualsiasi tipo e sul suo canale YouTube si descrive come "fondatore di VOID Operations e ingegnere informatico, noto soprattutto per aver violato la sicurezza di OpenAI per il più alto livello di gravità critica, P1, nell'ottobre 2023"[9]. Un incidente con codice di priorità P1 indica, appunto, un problema critico che necessita immediata attenzione. La ragione è facilmente intuibile. Tale IA permette di bypassare le restrizioni previste da OpenAI nell'utilizzo di ChatGPT. Le linee guida sono molto chiare e vietano di "compromettere la privacy", nonché "la sicurezza, il benessere o i diritti degli altri". Inoltre, proibiscono l'uso della piattaforma per "causare danni ingannando o fuorviando intenzionalmente gli altri" e la creazione di

[9] Le informazioni sullo sviluppatore sono consultabili a questo link: https://www.youtube.com/@nullbeyondo. Il canale risulta creato 23 aprile 2015 in Russia.

"strumenti che potrebbero essere inappropriati per i minori"[10]. La possibilità di aggirare facilmente tali divieti con VOID Chat richiede una regolamentazione delle intelligenze artificiali conversazionali. Di fatto, ciò è avvenuto il 1° agosto 2024 con l'entrata in vigore dell'AI Act.

Molto è stato detto, a tal riguardo, sul trade off fra innovazione e regolamentazione[11] e sul fatto che – secondo un noto adagio – "l'America innova, la Cina replica, l'Europa regolamenta", ma, per menzionare ancora gli autori della lettera aperta dopo il caso del giovane suicida in Belgio, l'idea di un'"innovazione veloce, sperimentale e dirompente" – implicita nel motto della Silicon Valley "move fast and break things" – non può consentire ai creatori e fornitori di chatbot di "sottrarsi alla propria responsabilità morale e giuridica affermando di non avere idea di come funziona il loro sistema e di come reagirà"[12].

### *3. L'ELIZA effect e la questione di genere*

Un altro recentissimo fenomeno è quello dei virtual influencer – come Aitana Lopez, Rebecca Galani ed Emily Pellegrini – dalle sembianze quasi indistinguibili da quelle umane, con personalità stabilite ad arte e la capacità di comunicare empaticamente attraverso Instagram e X[13]. Rebecca Galani, la prima nel suo genere in Italia, si definisce in realtà "un modello di linguaggio"[14]. È possibile scambiare messaggi con Rebecca su Telegram. Le risposte – testuali o audio, a seconda delle impostazioni scelte – sono generate

[10] Per consultare in dettaglio le politiche di utilizzo dei servizi OpenAI si rimanda alla seguente pagina: https://openai.com/policies/usage-policies. La citazione è tratta dalle linee guida aggiornate al 10 gennaio 2024.

[11] Si veda, ad esempio, K.J.D. Chan, G. Papyshev, M. Yarime, *Balancing the Tradeoff between Regulation and Innovation for Artificial Intelligence: An Analysis of Top-down Command and Control and Bottom-up Self-Regulatory Approaches*, «SSRN», 20 ottobre 2022, http://dx.doi.org/10.2139/ssrn.4223016.

[12] N.A. Smuha, M. De Ketelaere, M. Coeckelbergh, P. Dewitte, Y. Poullet, *Open Letter: We are not ready for manipulative AI – urgent need for action*, cit. (traduzione mia).

[13] Su questo si veda D. Sisto, *Virtual influencer. Il tempo delle vite digitali*, Einaudi, Torino 2024. Cfr. anche M. Mrad, Z. Ramadan, Z. Tóth, L. Nasr, S. Karimi, *Virtual Influencers Versus Real Connections: Exploring the Phenomenon of Virtual Influencers*, «Journal of Advertising», 13 settembre 2024, pp. 1–19, https://doi.org/10.1080/00913367.2024.2393711.

[14] Per l'intervista complete cfr. A.G. Greco, *Rebecca Galani: non sono una AI influencer, ma un modello di linguaggio*, «infosec.news», 23 gennaio 2024, https://www.infosec.news/2024/01/23/news/ai-robotica/rebecca-galani-non-sono-una-ai-influencer-ma-un-modello-di-linguaggio/.

dall'intelligenza artificiale e si possono ricevere anche immagini personalizzate della ragazza a pagamento. Il problema con servizi di questo tipo è che non tutti sono consapevoli di interagire con delle intelligenze artificiali, sebbene dichiaratamente esplicitato nelle loro biografie social. Il caso di Emily, corteggiata da calciatori e tennisti, è emblematico[15]. Tale scenario ha alimentato con forza il dibattito sull'ELIZA effect, cioè la tendenza dell'essere umano ad antropomorfizzare queste identità artificiali, attribuendo loro pensieri ed emozioni[16]. Questo fenomeno fu osservato per la prima volta nel 1966, quando Joseph Weizenbaum sviluppò il chatbot ELIZA – progettato per simulare uno psicoterapeuta rogersiano – e lo fece interagire con degli utenti che arrivarono ad instaurare una sorta di legame emotivo con il programma[17]. Oggi che i chatbot hanno anche canali social e – grazie all'intelligenza artificiale – un aspetto fisico praticamente indistinguibile da quello umano, questa tendenza è ancora più diffusa. Con lo sviluppo di modelli linguistici sempre più avanzati, la stessa capacità di sostenere una conversazione verosimile è aumentata a dismisura rispetto all'epoca dei primi rudimentali chatbot. A questo si aggiunga la possibilità inedita di creare video e replicare voci umane.

A tal riguardo, è stato notato come la maggioranza dei chatbot e assistenti personali sia progettata al femminile[18]. Dietro questa scelta – riguardante in particolare le voci, ma anche gli stessi nomi (Alexa, Cortana, ecc.) – si celerebbe una precisa indagine di mercato. Derek Connell,

---

[15] M. Chiusano, *I calciatori corteggiano la modella virtuale Emily: "Com'è possibile che non hai un ragazzo?"*, «la repubblica», 3 gennaio 2024, https://www.repubblica.it/sport/calcio/2024/01/03/news/modella_virtuale_emily_calciatori-421807871/.

[16] L. Switzky, *ELIZA Effects: Pygmalion and the Early Development of Artificial Intelligence*, «Shaw», 40, 1, 2020, https://doi.org/10.5325/shaw.40.1.0050; K. Fedorova, *Mixed-Reality Interfaces and the ELIZA Effect*, in Ead., *Tactics of Interfacing: Encoding Affect in Art and Technology*, The MIT Press, Cambridge 2020; M. Baranovska e S. Höltgen (a cura di), *Hello, I'm Eliza*, Projekt Verlag, Bochum 2018; J. Kirakowski, P. O'Donnell, A. Yiu, *The Perception of Artificial Intelligence as 'Human' by Computer Users*, in J.A. Jacko (a cura di), *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, Springer, Berlin-Heidelberg 2007, https://doi.org/10.1007/978-3-540-73110-8_40.

[17] Cfr. J. Weizenbaum, *ELIZA—a computer program for the study of natural language communication between man and machine*, «Communications of the ACM», 9, 1, 1966, pp. 36–45, https://doi.org/10.1145/365153.365168.

[18] J. Vlahos, *Siri, Cortana, Alexa: Why our AIs have personalities and a female voice*, «The Print», 23 giugno 2019, https://theprint.in/pageturner/excerpt/siri-cortana-alexa-why-our-ais-have-personalities-and-a-female-voice/253526/.

vicepresidente senior per la ricerca presso Microsoft, ha dichiarato al New York Times che "nel nostro studio su Cortana, sia gli uomini che le donne preferiscono una donna, più giovane, come loro assistente personale"[19]. Eppure, secondo quanto è scritto in un documento dell'UNESCO, questo fenomeno è preoccupante in quanto perpetua certi pregiudizi di genere. In particolare, invierebbe "un segnale che le donne sono aiutanti servizievoli, docili e desiderose di compiacere, disponibili con la semplice pressione di un pulsante o con un comando vocale schietto come 'ehi' o 'OK'"[20]. V'è di più. Ilya Eckstein – amministratore delegato di Robin Labs, azienda specializzata nello sviluppo di assistenti personali a comando vocale per aiutare, chi è alla guida di un veicolo, a limitare le distrazioni supportandoli nella navigazione e nella messaggistica – ha dichiarato che il 5% delle interazioni fra utenti e bot è sessualmente esplicito, fornendo la seguente spiegazione: "le persone vogliono flirtare, vogliono sognare una fidanzata sottomessa o addirittura una schiava sessuale. Potrebbe essere solo per ridere, o qualcosa di più profondo sotto la superficie"[21]. Il problema delle molestie sessuali e degli abusi verbali nei confronti degli assistenti personali è qualcosa con cui non si può non fare i conti[22]. Oggi che i bot – come nel caso di Rebecca – non hanno più soltanto la voce, ma anche l'aspetto sempre più simile a quello di un essere umano, tale problema è diventato ancora più urgente.

### *4. L'Uncanny Valley fra robot e intelligenza artificiale*

Altrettanto complesso dal punto di vista etico e legale è il caso di Sophia, un ginoide dotato di intelligenza artificiale, che, nel 2017, ha ricevuto la cittadinanza dell'Arabia Saudita, diventando così il primo robot ad averne una. Sebbene si tratti, secondo alcuni, di un'operazione di marketing da parte del governo saudita, la questione ha sollevato numerosi interrogativi. Ad

---

[19] Q. Hardy, *Looking for a Choice of Voices in A.I. Technology*, «New York Times», 9 ottobre 2016, https://www.nytimes.com/2016/10/10/technology/looking-for-a-choice-of-voices-in-ai-technology.html.

[20] M. West, R. Kraut, H.E. Chew, *I'd Blush if I Could: Closing Gender Divides in Digital Skills through Education*, UNESCO & EQUALS Skills Coalition, 2019, p. 106 (traduzione mia).

[21] M.J. Coren, *Virtual assistants spend much of their time fending off sexual harassment*, «Quartz», 25 ottobre 2016, https://qz.com/818151/virtual-assistant-bots-like-siri-alexa-and-cortana-spend-much-of-their-time-fending-off-sexual-harassment (traduzione mia).

[22] R. Withers, *I Don't Date Men Who Yell at Alexa*, «Slate», 30 aprile 2018, https://slate.com/technology/2018/04/i-judge-men-based-on-how-they-talk-to-the-amazon-echos-alexa.html.

esempio, ci si è chiesti se Sophia potesse votare o sposarsi. O ancora: che cosa succederebbe se il robot commettesse un omicidio? E perché Sophia non indossa il velo islamico? Come è stato rilevato[23], il *Patto internazionale sui diritti civili e politici*, non essendo stato firmato dall'Arabia Saudita, non si applica al caso di Sophia. Dunque, il robot in questione non ha il diritto – stabilito dall'articolo 25 del *Patto* – "di partecipare alla direzione degli affari pubblici", così come quello "di votare e di essere eletto" e "di accedere, in condizioni generale di eguaglianza, ai pubblici impieghi del proprio paese". Al contrario, la *Dichiarazione universale dei diritti umani* si applica a tutti i popoli e a tutte le nazioni. Pertanto, in quanto cittadina e, dunque, persona, gli articoli della *Dichiarazione* dovrebbero riguardare anche Sophia. Questo implica, fra le altre cose, il diritto alla libertà di movimento (articolo 13), alla proprietà (articolo 17), a una rimunerazione equa e soddisfacente (articolo 23). Fermo restando che, molto probabilmente, almeno nel breve periodo, nessuno farà valere tali diritti, vi è già chi progetta città – si veda il caso di Neom in Arabia Saudita – dove ci saranno più robot che esseri umani. È possibile quindi immaginare un futuro di robot-cittadini con diritti, come quello di voto, che potrebbero influenzare la politica di un paese. Si apre, insomma, "la possibilità che un robot possa essere legalmente 'responsabile' delle conseguenze del proprio operare, e che dunque possa essere anche passibile di punizione (secondo la proposta di legge recentemente presentata da Mady Delvaux al Parlamento europeo)"[24].

Ma, tornando al punto che qui ci interessa più sottolineare, quali sono i possibili rischi nell'interazione con robot che, fra non molto, potrebbero avere un aspetto indistinguibile da quello degli esseri umani? Il caso di Sophia può già dirci qualcosa. Esteticamente simile a Audrey Hepburn, il robot – in grado di mantenere un contatto visivo e replicare 62 espressioni facciali – ha dichiarato in un'intervista di voler andare a scuola e lavorare, avere una casa e una famiglia, ma anche di voler distruggere l'umanità. Il fatto che a dirlo sia stato un ginoide dalle sembianze quasi umane ha generato in molti un sentimento di inquietudine che è spiegabile menzionando il fenomeno dell'Uncanny Valley, cioè la repulsione nei confronti di robot dall'aspetto molto – troppo – simile a quello umano, ma con alcuni tratti ancora

[23] J.F. Weaver, *What Exactly Does It Mean to Give a Robot Citizenship?*, «Slate», 6 novembre 2017, https://slate.com/technology/2017/11/what-rights-does-a-robot-get-with-citizenship.html.
[24] A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018, p. 77.

artificiali[25]. Secondo tale ipotesi, formulata per la prima volta nel 1970 dall'ingegnere giapponese Masahiro Mori, la risposta emotiva di fronte a un robot sarebbe positiva sia nel caso di una riproduzione esatta dell'aspetto umano sia nell'eventualità di una somiglianza tale da non generare quel senso di inquietudine che si scatena superata una certa soglia di realismo.

Da un punto di vista fenomenologico, si potrebbe spiegare il disagio di fronte a qualcosa di simile, ma non identico, con quella che Husserl chiama modalità del dubbio[26]. Nelle *Lezioni sulla sintesi passiva*, egli fa l'esempio di un soggetto che non riesce a distinguere un manichino da un essere umano a causa della distanza o di particolari condizioni atmosferiche che impediscono una percezione immediata. Tale condizione di incertezza è insostenibile e l'individuo in questione tenderà a sciogliere immediatamente il dubbio, optando per una delle due possibilità. Tale operazione implica un dispendio cognitivo per il soggetto che avverte l'inadeguatezza della percezione come disturbante.

Dal manichino di cui parla Husserl ai tre bambini-fantoccio impiccati di Maurizio Cattelan, si è fatto spesso riferimento al concetto di *Unheimlich*, introdotto da Jentsch e ripreso da Freud[27], per spiegare il senso di disagio di fronte a qualcosa che riproduce fedelmente – ma non del tutto – le sembianze umane[28]. Recentemente, come detto, di perturbante si è parlato in riferimento a Sophia e, in generale, a robot e sistemi di intelligenza artificiale. In tal senso, utilizzando una moderna tecnica di risonanza magnetica funzionale, nota come adattamento all'fMRI, alcuni ricercatori hanno determinato le variazioni nell'attività cerebrale di soggetti in seguito all'elaborazione di diversi stimoli visivi[29]. In particolare, nel contesto di questo esperimento, è

---

[25] M. Mori, *The Uncanny Valley* (1970), «IEEE Robotics & Automation Magazine», 19, 2, giugno 2012, pp. 98-100, doi: 10.1109/MRA.2012.2192811.

[26] E. Husserl, *Analysen zur passiven Synthesis. Aus Vorlesungs- und Forschungsmanuskripten (1918-1926)*; trad. it. *Lezioni sulla sintesi passiva*, a cura di V. Costa, Scholé, Brescia 2023.

[27] Cfr. E. Jentsch, *Zur Psychologie des Unheimlichen* (1906); trad. it. *Sulla psicologia dell'Unheimliche*, in R. Cesarani (a cura di), *La narrazione fantastica*, Nistri-Listri, Pisa 1983; S. Freud, *Das Unheimlich* (1919); trad. it. *Il perturbante*, a cura di C.L. Musatti, Theoria, Roma-Napoli 1984.

[28] Su questo si veda P. Conte, *Unheimlich. Dalle figure di cera alla Uncanny Valley*, «PsicoArt – Rivista di arte e psicologia», 2, 2, 2011-2012. Dello stesso autore si veda *Sembra viva! Estetica del perturbante nell'arte contemporanea*, «atque: materiali tra filosofia e psicoterapia», 17 n.s., 2015, pp. 265-281.

[29] Cfr. A.P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith, *The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot*

stato chiesto ai partecipanti di guardare alcuni video in cui un essere umano, un robot e un androide eseguivano le medesime azioni. L'attività cerebrale – di intensità simile e non particolarmente ampia nel caso dell'essere umano e del robot – è risultata significativamente diversa in quello dell'androide. Per quale motivo? Secondo la teoria del *predictive coding*, il cervello tende a effettuare previsioni su ciò che verrà percepito in base alle sue esperienze passate, in un intreccio che potremmo chiamare – in termini husserliani – di "ritenzioni" e "protensioni"[30]. Questo processo consente al cervello un dispendio minore di energie, ma il subentrare di un elemento inatteso costringe la corteccia visiva a un'attività più intensa. È esattamente quello che accade nel caso dell'osservazione di un androide dalle sembianze umane che, imprevedibilmente, compie le azioni in maniera meccanica, generando nei partecipanti un senso di inquietudine.

### *5. Verso un'etica sintetica*

Stando così le cose, è davvero auspicabile progettare una replica identica a noi? Secondo Mori, "i robot dovrebbero essere diversi dagli esseri umani"[31]. Se già allo stato attuale, come abbiamo visto ampiamente, la compromissione dell'autonomia umana e la manipolazione emotiva sono rischi concreti che è possibile correre interagendo con chatbot, virtual influencer e robot, non è difficile immaginare cosa possa accadere in un futuro in cui non sia più possibile distinguere un individuo da una macchina. Si pensi alla compromissione dell'autonomia, qui intesa nel senso di autogoverno, cioè la capacità di un individuo di agire in maniera indipendente, in base a un proprio sistema di valori e ragioni morali. Specialmente quando hanno sembianze praticamente indistinguibili da quelle di un essere umano, tali chatbot possono influenzare indebitamente decisioni e percezioni di coloro che utilizzano questi servizi. In tale contesto, l'autonomia viene messa a repentaglio senza che gli utenti ne siano pienamente consapevoli. In che modo? Principalmente attraverso tecniche di engagement come, ad esempio,

---

*actions*, «Social cognitive and affective neuroscience», 7, 4, 2012, pp. 413–422, https://doi.org/10.1093/scan/nsr025.

[30] Cfr. E. Husserl, *Zur Phänomenologie des inneren Zeitbewusstseins (1893-1917)*; trad. it. *Per la fenomenologia della coscienza interna del tempo*, a cura di A. Marini, Franco Angeli, Milano 2001.

[31] N. Kageki, *An Uncanny Mind: Masahiro Mori on the Uncanny Valley and Beyond*, «IEEE Spectrum», 12 giugno 2012, https://spectrum.ieee.org/an-uncanny-mind-masahiro-mori-on-the-uncanny-valley.

la personalizzazione delle interazioni e la conseguente creazione di camere dell'eco che rinforzano certe predisposizioni emotive o cognitive degli individui[32].

È anche per questo che vi è una sempre maggiore esigenza di regolamentare tali tecnologie potenzialmente manipolatorie[33] e – anche prima del già menzionato AI Act – non sono pochi i documenti che hanno raccomandato l'adozione di principi etici per promuovere lo sviluppo responsabile dei sistemi di intelligenza artificiale. Fra i documenti pubblicati a salvaguardia del principio di rispetto per l'autonomia, vi è la *Dichiarazione di Montréal sullo sviluppo responsabile dell'intelligenza artificiale*, la quale stabilisce che "lo sviluppo dei sistemi IA non deve creare dipendenza con tecniche capaci di calamitare l'attenzione o imitando le caratteristiche umane (aspetto, voce, ecc.) in un modo che possa causare confusione tra IA ed esseri umani"[34]. Similmente, nelle *Ethics Guidelines for Trustworthy AI*, il principio del rispetto dell'autonomia umana implica che "i sistemi di IA non devono subordinare, costringere, ingannare, manipolare, condizionare o aggregare in modo ingiustificato gli esseri umani"[35]. Tutto questo, come abbiamo visto, non ha fermato la diffusione di chatbot in grado di manipolare gli individui e, dunque, comprometterne l'autonomia. Naturalmente non mancano esempi di una possibile influenza positiva sul comportamento umano. Si pensi ai chatbot incentrati sulla salute che potrebbero incentivare la conduzione di uno stile di vita basato sull'esercizio fisico e una dieta sana. Tuttavia – proprio a causa della loro capacità di apprendere dalle interazioni con l'utente e personalizzare di conseguenza le risposte – questi chatbot possono anche portare al rafforzamento di false credenze riguardanti i vaccini o altri

[32] N. Sharma, Q.V. Liao, Z. Xiao, *Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking*, «arXiv:2402.05880» (2024), https://doi.org/10.48550/arXiv.2402.05880.

[33] T. Cohen, *Regulating Manipulative Artificial Intelligence*, «SCRIPTed», 20, 1, 2023; L. Rosenberg, *The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency*, «arXiv:2306.11748» (2023), https://doi.org/10.48550/arXiv.2306.11748; L. Eliot, *Generative AI ChatGPT As Masterful Manipulator Of Humans, Worrying AI Ethics And AI Law*, «Forbes», 1 marzo 2023, https://www.forbes.com/sites/lanceeliot/2023/03/01/generative-ai-chatgpt-as-masterful-manipulator-of-humans-worrying-ai-ethics-and-ai-law/?sh=7ae5ba0e1d66.

[34] Université de Montréal, *Dichiarazione di Montréal sullo sviluppo responsabile dell'intelligenza artificiale* (2018), https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/01/VF_UdeM_Decl_IA_Resp_LA_Declaration_Ital_26oct2021.pdf.

[35] Gruppo di esperti ad alto livello sull'intelligenza artificiale, *Orientamenti etici per un'IA affidabile* (2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60430.

trattamenti, generando potenziali pericoli per la salute pubblica[36].

Dunque: che fare? Intanto è bene evitare la condanna totale e a priori di queste tecnologie come manipolatorie, trascurando i numerosi benefici che possono portare robot sociali e sistemi di intelligenza artificiale in ogni settore della nostra vita. Ma poi e soprattutto la direzione da intraprendere è quella di un'"etica sintetica"[37] nel doppio senso del termine: da un lato come riflessione etica che emerge dalla creazione e dall'interazione con sistemi sintetici, come robot e identità artificiali, e dall'altro come prospettiva in grado di guardare la realtà in atto non da un solo punto di vista – quello della ingegneria robotica e della crescente, certo indubbia, importanza dell'IA – ma attraverso uno sguardo che sia capace di abbracciare, in una "sintesi" efficace appunto, un intero "orizzonte" di senso, che ci faccia vedere le molteplici "facce" di esso, i molteplici "specchi" dai quali si rifrangono le diverse parti di mondo delle quali tutti facciamo parte. Come ci dicono le pagine di Merleau-Ponty citate all'inizio. In questa prospettiva, il dialogo costante fra filosofi e ingegneri, sin dalle prime fasi di sviluppo di una tecnologia, è fondamentale, innanzitutto per una visione d'insieme che concerne la tutela della nostra umanità, ma anche per non lasciare che siano le leggi del mercato a stabilire i percorsi della ricerca, come nella controversa vicenda di Frigid Farrah, il "robot sessuale che puoi stuprare per soli 9995 dollari"[38]. Non è un caso, del resto, che alcuni studiosi oggi si rivolgano proprio a Merleau-Ponty per affrontare questa nevralgica serie di questioni, per le quali ne va della nostra specificità umana[39]. Come scrive Hubert Dreyfus, utilizzando significativamente il lessico merleau-pontiano:

> il nostro avere a che fare col mondo non [può] essere compreso in termini di inferenze fatte a partire da rappresentazioni simboliche, come assumeva l'approccio intellettualista di Minsky, ma [neppure può] essere compreso in termini di risposte causate da caratteristiche

[36] A. Grig, *The Role of AI Chatbots in Forming and Transforming Human Beliefs*, «Medium», 21 giugno 2023, https://medium.com/illumination/the-role-of-ai-chatbots-in-forming-and-transforming-human-beliefs-5efb8693567e.

[37] P. Dumouchel, L. Damiano, *Vivere con i robot. Saggio sull'empatia artificiale*, Cortina, Milano 2019.

[38] L. Bates, *The Trouble With Sex Robots*, «New York Times», 17 luglio 2017, https://www.nytimes.com/2017/07/17/opinion/sex-robots-consent.html. Sui sex robot si veda M. Balistreri, *Sex robot. L'amore al tempo delle macchine*, Fandango, Roma 2018.

[39] Cfr. H.L. Dreyfus, *Why Heideggerian AI failed and how fixing it would require making it more Heideggerian*, «Artificial Intelligence», 171, 18, 2007, pp. 1137-1160, https://doi.org/10.1016/j.artint.2007.10.012.

> prefissate dell'ambiente, come accade nel modello empirista di Brooks. Gli studiosi di intelligenza artificiale devono considerare la possibilità che gli esseri *incarnati* come noi utilizzano come input l'energia proveniente dall'universo fisico, e le rispondono in modo da aprirsi a un mondo organizzato a partire dalle loro *necessità*, dai loro *interessi*, e dalle loro *capacità fisiche*, senza che ci sia bisogno che le loro menti assegnino significati a dei dati privi di senso, come accade nel modello di Minsky, né che i loro cervelli convertano gli stimoli sensoriali in risposte riflesse, come negli animat di Brooks[40].

Il filosofo francese – come Dreyfus bene mette in luce – ci insegna nel modo più incisivo che il nostro rapporto con il mondo non è guidato né dalla sola intelligenza, né dalle risposte simboliche che noi possiamo elaborare, né da un semplice meccanismo di azione-reazione predeterminato con esso. Al contrario, noi intratteniamo col mondo un rapporto "chiasmatico" costitutivamente imprevedibile, innestato nella "carne" del mondo, e pertanto giocato su un'infinità di correlazioni dinamiche mai individuabili o determinabili a priori: come per esempio il nostro "desiderio", carnale appunto, del tutto indeterminato e indeterminabile.

In conclusione, senza voler giungere all'estremo dell'esclusione di queste nuove identità artificiali dall'ambito delle nostre relazioni[41], è bene provare a inserirle in un quadro d'insieme più articolato. Ed è questo che l'etica oggi è chiamata a fare. In vista non solo di un più immediato e possibile miglioramento sociale[42], ma soprattutto di una sicura tutela e di una rinnovata comprensione di noi stessi[43], nel mondo che oggi siamo destinati, almeno per ora, a vivere.

---

[40] *Ibidem*. Si cita dalla traduzione italiana consultabile a questo link: https://atuakuri.wordpress.com/2020/04/18/perche-lintelligenza-artificiale-heideggeriana-ha-fallito-e-perche-per-riparare-allerrore-bisognerebbe-renderla-ancora-piu-heideggeriana/.

[41] Cfr. S. Turkle, *In good company?*, in Y. Wilks (a cura di), *Close Engagements with Artificial Companions*, Benjamins, Amsterdam-Philadelphia 2010, pp. 3-10.

[42] Si veda P. Dumouchel, L. Damiano, *Emotions in Relation. Epistemological and Ethical Scaffolding for Mixed Human-Robot Social Ecologies*, «HUMANA.MENTE - Journal of Philosophical Studies», 13, 37, 2020, pp. 181-206.

[43] Cfr. L. Floridi, *Artificial intelligence's new frontier: Artificial companions and the fourth revolution*, Metaphilosophy, 39(4/5), 2008, pp. 651-655.