TRAIL Series
PAPER N. 1
a.a. 2019/2020

Trento BioLaw Selected Student Papers

I.A. tra efficienza e trasparenza: il diritto a conoscere la natura dell'interlocutore

Alex Bertoldi, Caterina Tallarida, Kosovare Krasniqi, Alessio Gazzin, Nicolò Barbini

Jean Monnet European Centre University of Trento
Co-funded by the Erasmus+ Programme of the European Union
Università di Trento Facoltà di Giurisprudenza

I paper sono stati selezionati a conclusione del corso libero *Diritto e Intelligenza Artificiale* a.a. 2019-2020, organizzato all'interno del Progetto Jean Monnet "TrAIL – Trento Artificial Intelligence Laboratory", coordinato presso l'Università di Trento dai docenti Carlo Casonato e Simone Penasa.

I.A. tra efficienza e trasparenza:
il diritto di conoscere la natura dell'interlocutore

*Alex Bertoldi, Caterina Tallarida, Kosovare Krasniqi, Alessio Gazzin, Nicolò Barbini**

ABSTRACT: People are still very unlikely to trust machines more than they trust their fellow humans. This remains true even when people see algorithms repeatedly outperform humans in specific tasks. Recent studies show that humans and machines could achieve a more successful and profitable cooperation if machines did not disclose their true nature. As a result of this combination of bias and distrust towards machines, there is a significant trade-off between transparency and efficiency in human-machine interactions. Here, we discuss the case of Google Duplex, a voice assistant capable of mimicking a human voice to make appointments. Should we let such technologies deceive us for the sake of a more productive interaction or should every AI system be always identifiable and transparent? The aim of this paper is to determine if people have the right to know whether they are interacting with a human being or with an AI system. We ultimately argue that the disclosure of the non-human nature of machines is primarily an ethical issue, that should not therefore be regulated solely on the basis of economic efficiency and cost-benefit analysis.

KEYWORDS: artificial intelligence; human-machine interactions; transparency; efficiency; Google Duplex.

SOMMARIO: 1. Introduzione – 2. Trade-off tra efficienza e trasparenza nelle interazioni uomo-macchina – 3. Il caso Google Duplex – 4. Considerazioni etiche e giuridiche – 5. Conclusioni

1. Introduzione

Nell'ultimo decennio la tecnologia ha fatto passi da gigante nello sviluppo delle intelligenze artificiali e della robotica, arrivando a creare macchine estremamente sofisticate, sempre più simili a noi, nell'aspetto e nel comportamento.

Nonostante questa evoluzione, gli esseri umani tendono ancora a non fidarsi completamente delle macchine dotate di intelligenza artificiale e, anzi, spesso nutrono pregiudizi nei loro confronti. A causa di questa vera e propria "avversione per gli algoritmi"[1], sembra esistere un significativo trade-off tra efficienza e trasparenza nelle interazioni uomo-macchina. In altre parole, in determinati contesti, efficienza e trasparenza si trovano in un rapporto di proporzionalità inversa tra loro: nel momento in cui le persone

* *Studenti dell'Università degli Studi di Trento, Facoltà di Giurisprudenza*.

[1] Per un approfondimento del fenomeno di "algorithm aversion", si segnala, in particolare: B. J. DIETVORST, J. P. SIMMONS, C. MASSEY, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, in *Journal of Experimental Psychology: General*, n. 1, 2015, pp. 114-126.

vengono a conoscenza della natura della macchina, la collaborazione tra l'uomo e la macchina perde automaticamente in efficienza. Esseri umani e macchine, dunque, potrebbero realizzare una cooperazione più proficua e vantaggiosa se le macchine non rivelassero la loro vera natura.

Per comprendere meglio questo fenomeno e le sue conseguenze, esamineremo brevemente il caso di Google Duplex, un assistente vocale di ultima generazione in grado di riprodurre in maniera estremamente fedele la voce umana e di portare a termine conversazioni telefoniche decisamente realistiche in quasi totale autonomia.

Nei paragrafi che seguono si tenterà poi di rispondere ad una fondamentale domanda: è opportuno lasciare che queste tecnologie nascondano la loro natura artificiale per consentire una maggiore efficienza nelle interazioni uomo-macchina, oppure occorre privilegiare in ogni caso riconoscibilità e trasparenza?

Scopo dell'indagine sarà, in ultima analisi, quello di individuare alcune possibili soluzioni giuridiche ed etiche al problema, nonché di determinare l'eventuale sussistenza di una nuova figura di diritto soggettivo: il diritto di conoscere la natura, umana o artificiale, del proprio interlocutore.

2. Trade-off tra efficienza e trasparenza nelle interazioni uomo-macchina

Le persone possono mentire, dare informazioni volutamente errate. Ciò non rientra invece nelle capacità di una macchina, la quale al massimo potrà dare un'informazione errata non perché ci vuole ingannare, ma perché il suo programma è poco efficiente. Ma allora perché tendiamo a fidarci più dei nostri simili piuttosto che delle macchine dotate di intelligenza artificiale? Nonostante i grandi progressi fatti nel campo della robotica e dell'informatica per rendere le macchine sempre più simili a noi, infatti, vi è ancora una forte diffidenza verso di esse.

Un recente studio[2], condotto da un team internazionale, ha confermato l'esistenza di un trade-off tra trasparenza ed efficienza nelle interazioni uomo-macchina in contesti collaborativi. Tale studio ha dimostrato come le persone rompano più facilmente le promesse fatte ad una macchina e ritengano gli esseri umani più intelligenti e cooperativi delle macchine. Lo studio si è svolto osservando il comportamento di un campione di 700 persone nel gioco del dilemma del prigioniero[3]. Il campione è stato

[2] F. Ishowo-Oloko, J. F. Bonnefon, Z. Soroye, J. Crandall, I. Rahwan, T. Rahwan, *Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation*, in *Nature Machine Intelligence*, n. 11, 2019, pp. 517-521.

[3] Il dilemma del prigioniero è un noto gioco non cooperativo, elaborato dal matematico statunitense A. W. Tucker. La dinamica del gioco può essere sinteticamente descritta come segue. Due persone vengono arrestate e accusate di aver commesso un reato. Gli investigatori li confinano in due stanze separate, senza possibilità di comunicare tra loro. Ad entrambi vengono date due opzioni: confessare o negare la propria colpevolezza. Se entrambi confessano di aver commesso il reato, verranno condannati a sei mesi di reclusione; se entrambi negano, la pena sarà di tre mesi per entrambi; se soltanto uno dei due confessa e l'altro nega, la pena per chi ha ammesso la responsabilità del crimine verrà ridotta a un mese di reclusione, mentre verrà raddoppiata a dodici mesi a chi nega. La migliore strategia, al fine di minimizzare la propria condanna, è quella di confessare per entrambi, dal momento che non si ha modo di conoscere la scelta compiuta dall'altra persona.

suddiviso in quattro gruppi: ad un gruppo è stato detto che avrebbe cooperato con una IA; ad un altro gruppo è stato detto che avrebbe cooperato con altre persone; ad un terzo gruppo è stato detto che avrebbe cooperato con una IA, mentre cooperavano in realtà con una persona; e all'ultimo gruppo è stato detto che avrebbe cooperato con una persona, che in realtà era una macchina. Dai risultati di questa sperimentazione si è potuto constatare che le persone interagiscono in maniera diversa con le macchine, meno collaborativa rispetto alle interazioni con altri esseri umani. Tale pregiudizio nei confronti delle macchine può in alcuni casi annullare il vantaggio iniziale dato dall'impiego delle macchine al posto degli esseri umani[4]. In sostanza, le macchine sono migliori degli esseri umani nell'indurre alla cooperazione, ma solo se sia permesso loro di fingersi umani[5]. Infatti, anche nei casi in cui era stato specificamente indicato alle persone di cooperare con le macchine come se fossero esseri umani, i partecipanti si comportavano come se non avessero nemmeno ricevuto tale indicazione. Nel momento in cui la vera natura della macchina era rivelata, il livello di cooperazione diminuiva considerevolmente rispetto ai livelli raggiunti dalla combinazione uomo-uomo[6]. Ciò mostra come il bias sia troppo forte e il pregiudizio non sia superabile. Per garantire la piena efficienza delle macchine è dunque necessario nascondere la loro vera natura, ingannando l'utente e facendogli credere di cooperare con una persona. Tale condizione va, però, inevitabilmente, a discapito della trasparenza nelle interazioni uomo-macchina. Talal Rahwan, tra gli autori dello studio e docente di Computer Science presso la New York University Abu Dhabi, afferma che questi risultati evidenziano i possibili costi in termini di efficienza che la società potrebbe dover sostenere in cambio di maggiore trasparenza, anche se aggiunge che ciò può dipendere dal contesto. L'identificazione dei casi in cui il trade-off tra trasparenza ed efficienza esiste e di quelli in cui non esiste rimane una questione aperta che necessita di ulteriori ricerche e sperimentazioni[7].

3. Il caso Google Duplex

Nel maggio del 2018, in occasione di Google I/O[8], conferenza annuale dedicata agli sviluppatori dell'azienda, è stato presentato Google Duplex, un assistente vocale dotato di un sistema di intelligenza artificiale particolarmente sofisticato, capace di simulare le interazioni tra esseri umani e di condurre conversazioni virtuali estremamente realistiche. Per il momento Duplex è in grado di svolgere in quasi

---

Per una descrizione più accurata del dilemma del prigioniero e delle sue implicazioni economiche, si rinvia a D. TOSATO, *Dizionario di Economia e Finanza*, 2012, in http://www.treccani.it/enciclopedia/dilemma-del-prigioniero_%28Dizionario-di-Economia-e-Finanza%29/ (ultima consultazione 15/02/2020).

[4] S. EL-SHOWK, *Human bias burdens bots*, in *Nature Middle East*, 22 novembre 2019, https://www.natureasia.com/en/nmiddleeast/article/10.1038/nmiddleeast.2019.153 (ultima consultazione 15/02/2020).

[5] F. ISHOWO-OLOKO, J. F. BONNEFON, Z. SOROYE, J. CRANDALL, I. RAHWAN, T. RAHWAN, *op. cit.*, p. 519.

[6] *Ibidem.*

[7] S. EL-SHOWK, *op. cit.*

[8] Per ulteriori informazioni, si rimanda al sito ufficiale dell'evento: https://events.google.com/io/ (ultima consultazione 26/11/2019).

completa autonomia alcuni semplici e specifici compiti, quali prenotare una cena al ristorante o fissare un appuntamento dal parrucchiere per conto dell'utente[9].

L'obiettivo principale di questa innovativa tecnologia è quello di consentire agli utenti di risparmiare tempo grazie ad uno strumento ideato specificamente per telefonare e interagire al loro posto, che risulti del tutto spontaneo e naturale. Durante la presentazione ufficiale, il CEO di Google, Sundar Pichai, ha fatto ascoltare al pubblico la registrazione di due brevi conversazioni fra il sistema Duplex e degli operatori umani, suscitando stupore e ilarità tra i presenti[10]. In effetti, sono sufficienti pochi secondi per rendersi conto di quanto Duplex sia diverso dalle voci computerizzate e meccaniche dei comuni assistenti vocali cui siamo abituati. Il sistema di intelligenza artificiale di cui è dotato è in grado non soltanto di cogliere le sfumature nelle conversazioni[11] ma anche, e soprattutto, di comunicare attraverso espressioni ed esitazioni tipicamente umane, al punto da risultare sostanzialmente indistinguibile da un essere umano. Naturalmente, Duplex non può ancora condurre conversazioni vaghe e generiche, diverse da quelle per cui è stato programmato. Inoltre, è dotato di un meccanismo di auto-monitoraggio che gli consente di richiedere l'intervento umano nei casi in cui non riesca a portare a termine il proprio compito.

La presentazione di Duplex, come è facile immaginare, ha sollevato sin da subito forti critiche, legate in particolare alle implicazioni etiche di una tecnologia pensata per ingannare a tutti gli effetti l'essere umano. Il fatto che un assistente vocale sia stato progettato specificamente per nascondere la propria natura artificiale agli esseri umani è stato da alcuni considerato terrificante, oltre che subdolo e ingannevole[12]. Google, in risposta alle pesanti accuse ricevute, è stata costretta a dichiarare ufficialmente che in futuro il sistema si identificherà con l'interlocutore prima di procedere oltre con la telefonata, in modo da non creare alcun dubbio sulla sua natura[13]. Richiamando quanto già precedentemente affermato in un articolo pubblicato sul blog di Google dedicato all'intelligenza artificiale[14], i portavoce dell'azienda ribadiscono che

---

[9] V. RITA, *L'intelligenza artificiale di Duplex, l'assistente di Google che parlerà al vostro posto*, in *Wired.it*, 9 maggio 2018, https://www.wired.it/scienza/lab/2018/05/09/google-duplex-intelligenza-artificiale/ (ultima consultazione 26/11/2019).

[10] *Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments*, 11 maggio 2018, https://www.youtube.com/watch?v=D5VN56jQMWM (visualizzato il 25/11/2019).

[11] *Ibidem.*

[12] F. ISHOWO-OLOKO, J. F. BONNEFON, Z. SOROYE, J. CRANDALL, I. RAHWAN, T. RAHWAN, *op. cit.*, p. 517.

[13] Si segnala, a titolo esemplificativo J. VOMIERO, *Google now says controversial AI voice calling system will identify itself to humans*, in *The Verge*, 10 maggio 2018, https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update (ultima consultazione 28/11/2019); M. BERGEN, *Google Grapples With 'Horrifying' Reaction to Uncanny AI Tech*, in *Bloomberg*, 10 maggio 2018, https://www.bloomberg.com/news/articles/2018-05-10/google-grapples-with-horrifying-reaction-to-uncanny-ai-tech (ultima consultazione 28/11/2019); A. HERN, *Google's 'deceitful' AI assistant to identify itself as a robot during calls*, in The Guardian, 11 maggio 2018, https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls (ultima consultazione 28/11/2019).

[14] «The Google Duplex technology is built to sound natural, to make the conversation experience comfortable. It's important to us that users and businesses have a good experience with this service, and transparency is a key part of that. We want to be clear about the intent of the call so businesses understand the context. We'll be experimenting with the right approach over the coming

la trasparenza è una componente chiave del sistema Duplex[15]. Resta da capire in che misura la decisione di rivelare la natura non umana dell'assistente vocale di Google pregiudicherà la sua efficienza e il suo effettivo funzionamento.

4. Considerazioni etiche e giuridiche

Sebbene vi sia ampio consenso sulla necessità che i processi decisionali automatizzati suscettibili di incidere sui diritti delle persone siano il più possibile trasparenti, comprensibili e conoscibili, non altrettanto chiaro è se sia anche opportuno che le macchine dotate di intelligenza artificiale rivelino in ogni caso la propria natura non umana, indipendentemente dal contesto e dai diritti fondamentali coinvolti[16].

Da un punto di vista strettamente giuridico, quali diritti e principi costituzionali potrebbero essere lesi dal fatto che una persona fisica non sia posta nella condizione di conoscere la natura, umana o artificiale, del proprio interlocutore? Tale interrogativo si pone come fondamentale soprattutto nei casi in cui la IA abbia capacità tecniche tali da superare il test di Turing[17] o la teoria della "Uncanny Valley"[18]. Quale quadro giuridico potrebbe essere dato per regolamentare tali situazioni? Quali ripercussioni etiche dovrebbero essere considerate?

In primo luogo, tra le implicazioni problematiche che potrebbero presentarsi a causa della impossibilità di conoscere la natura del proprio interlocutore, c'è quella della violazione del diritto ad essere correttamente informati. Infatti, determinati soggetti potrebbero voler essere adeguatamente informati prima di interagire con un'intelligenza artificiale, anche per l'effetto che essa potrebbe produrre, cioè quello di immettere tutti i dati che riceve e raccoglie in un circuito difficilmente controllabile (cloud)[19]. In tal maniera si rischia inoltre di minare il diritto alla privacy, poiché dati personali, e potenzialmente sensibili, verrebbero inseriti in rete, senza il consenso dell'interessato. In secondo luogo, è lecito affermare che detta situazione di indeterminatezza potrebbe avere ripercussioni negative anche sul concetto di dignità umana, dal

---

months» Y. LEVIATHAN, Y. MATIAS, *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*, Google AI Blog, 8 maggio 2018, Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone (ultima consultazione 28/11/2019).

[15] Queste le parole di un portavoce di Google, in risposta alle critiche ricevute: «We understand and value the discussion around Google Duplex — as we've said from the beginning, transparency in the technology is important […]. We are designing this feature with disclosure built-in, and we'll make sure the system is appropriately identified. What we showed at I/O was an early technology demo, and we look forward to incorporating feedback as we develop this into a product», in J. VOMIERO, *op. cit*.

[16] F. ISHOWO-OLOKO, J. F. BONNEFON, Z. SOROYE, J. CRANDALL, I. RAHWAN, T. RAHWAN, *op. cit*., p. 519.

[17] Criterio elaborato dal matematico Alan Turing per determinare se una macchina sia in grado o meno di pensare autonomamente. Si rimanda a A. M. TURING, *Computing Machinery and Intelligence*, in *Mind*, n. 236, 1950, pp. 433-460.

[18] Ipotesi presentata dallo studioso di robotica Masahiro Mori in un saggio pubblicato nel 1970 dalla rivista giapponese *Energy.* Secondo l'autore, la reazione degli esseri umani nei confronti dei robot antropomorfi più sofisticati sarebbe passata dalla empatia e familiarità iniziali alla repulsione e inquietudine, al crescere della loro somiglianza con gli esseri umani. Si veda: M. MORI, *Bukimi No Tani [The Uncanny Valley]*, in *Energy*, n. 4, 1970, pp. 33-35. La traduzione inglese del saggio originale di Mori è reperibile al sito: https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley (ultima consultazione 29/11/2019).

[19] C. CASONATO, *Intelligenza artificiale e diritto costituzionale: prime considerazioni*, in *Diritto pubblico comparato ed europeo*, Fascicolo speciale, 2019, p. 126.

momento che la persona verrebbe a tutti gli effetti ingannata, e portata a credere di essere a contatto con un altro essere umano.

Un altro importante diritto che in questo modo rischia di essere sacrificato è la libertà di autodeterminazione dell'individuo, ovvero la capacità del soggetto di decidere in maniera autonoma, indipendente e consapevole se interagire con una macchina o meno, a seconda della propria sensibilità e dei propri bisogni. A tal proposito si deve ricordare che esistono alcune categorie di persone (soprattutto anziani o bambini) ancora incapaci di leggere e interpretare la realtà tecnologica che vertiginosamente si sta sviluppando. Infine, possono presentarsi casi in cui una presenza umana crea maggiore senso di affidabilità e calore che una macchina non è in grado di offrire.

Quali sono le possibili soluzioni giuridiche ai problemi poc'anzi sollevati? Autorevole dottrina propone, in primo luogo, una nuova figura di diritto soggettivo: il diritto ad essere resi consapevoli della natura, umana o artificiale, del proprio interlocutore[20]. Sembra che anche le istituzioni europee si stiano muovendo in tal senso: assumono rilevanza al riguardo gli *Orientamenti etici per un'IA affidabile*[21], e la *Raccomandazione del Commissario per i diritti umani del Consiglio d'Europa*[22].

Secondo la prima fonte: «*i sistemi di intelligenza artificiale non devono rappresentare se stessi come esseri umani per gli utenti; gli umani hanno il diritto di essere informati che stanno interagendo con un sistema di intelligenza artificiale. Ciò implica che i sistemi di intelligenza artificiale devono essere identificabili come tali. Inoltre, l'opzione per decidere contro questa interazione a favore dell'interazione umana dovrebbe essere fornita ove necessario per garantire il rispetto dei diritti fondamentali. Oltre a ciò, le capacità e le limitazioni del sistema di intelligenza artificiale dovrebbero essere comunicate [...] agli utenti finali in un modo adeguato*».

Secondo la seconda fonte: «*l'uso di un sistema di intelligenza artificiale in qualsiasi processo decisionale che abbia un impatto significativo sui diritti umani di una persona deve essere identificabile*».

5. Conclusioni

Da questa breve analisi è emersa la necessità di ulteriori e più approfondite ricerche sulla relazione tra efficienza e trasparenza nelle interazioni tra esseri umani e macchine.

---

[20] *Ivi*, pp. 125 ss.

[21] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, *Ethics guidelines for trustworthy AI*, 8 aprile 2019, p.18. Il documento, reso pubblico nell'aprile del 2019, è stato redatto da un gruppo di esperti ad alto livello sull'intelligenza artificiale istituito dalla Commissione Europea nel giugno del 2018. Il testo è reperibile al sito: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (ultima consultazione 15/02/2020).

[22] COMMISSARIO PER I DIRITTI UMANI DEL CONSIGLIO D'EUROPA, *Unboxing Artificial Intelligence: 10 steps to protect Human Rights. Recommendation*, maggio 2019, pp. 9-10. Si veda: https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights (ultima consultazione 15/02/2020).

Occorre identificare con precisione i casi in cui sia imperativo offrire maggiori tutele ai soggetti fisici che si trovino a interagire con sistemi dotati di intelligenza artificiale estremamente sofisticati, del tipo qui esaminato. È necessario, in altre parole, distinguere i casi in cui si può ammettere minore trasparenza, in funzione di una più produttiva ed efficiente interazione con le macchine, dai casi in cui occorre dare priorità assoluta al principio di trasparenza. Una disciplina giuridica che sia valida in ogni caso, e che imponga la massima trasparenza, senza compiere alcun bilanciamento tra i contrapposti interessi in gioco, non sembra essere la risposta adeguata a regolare la complessità del fenomeno. Nello specifico, è ragionevole affermare che un assistente vocale utilizzato esclusivamente per fissare semplici appuntamenti dal parrucchiere o per prenotare una cena, non ponga particolari problemi in termini di violazione dei diritti fondamentali dell'individuo. Ciononostante, riteniamo che sia opportuno interrogarsi sin da subito sulle possibili future applicazioni di una tecnologia talmente sofisticata da non consentire di distinguere un interlocutore artificiale da un essere umano.

Sembra in ogni caso auspicabile, come approccio generale, e in attesa di ulteriori studi sul tema, che le legislazioni nazionali e le normative europee diano preminenza ai principi di trasparenza e conoscibilità piuttosto che a quello di efficienza, nei casi in cui dalla interazione con la macchina possano derivare danni fisici o psicologici per l'essere umano[23]. Riteniamo, in ultima analisi, che si tratti di una questione principalmente etica, suscettibile di incidere su principi tutelati anche a livello costituzionale, quali la dignità umana, la libertà di scelta e di autodeterminazione del singolo, e che non possa dunque essere regolata strettamente sulla base di un'analisi economica di costi-benefici o della ricerca di maggiore produttività.

[23] Si segnala, a tal proposito, quanto afferma Brent Mittelstadt, studioso di data ethics e ricercatore presso l'Oxford Internet Institute: «[Mittelstadt hopes that, in the future] we default to a right to know in most situations, or at a minimum in those situations where material or psychological harm could result from the interaction, […] we will need much more research looking at the social, psychological, and ethical effects of interactions with bots, both when they announce themselves as non-human, and when they don't. Interacting with another person, especially face-to-face, can have significant psychological and social benefits that are difficult to quantify, and can easily be lost or ignored in development driven primarily by a pursuit of greater efficiency or saving» in S. EL-SHOWK, *op. cit*.