**Davide Bennato**

docente di Sociologia dei media digitali - Università di Catania

Esistono già laboratori di ricerca che stanno lavorando allo sviluppo intelligenze artificiali che mostrano comportamenti devianti come razzismo, violenza e altre psicopatologie più o meno deliranti.

E non si tratta di [fake news](https://www.agendadigitale.eu/tag/fake-news/), né di una puntata particolarmente visionaria di Black Mirror.

La notizia è tremendamente vera, e ci sono validi motivi scientifici per procedere su questa linea di ricerca.

Vediamo i fatti.

Una delle tecnologie di maggiore tendenza degli ultimi anni è sicuramente l’AI ovvero [*Artificial Intelligenc*](https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale-unaltra-euforia-di-breve-durata-cosa-insegna-la-storia/)*e* che oggi prende nomi diversi – *machine learning*, *deep learning* – ma in buona sostanza l’idea è sempre quella: la capacità di sviluppare un software autonomo che sia in grado di inferire nuova conoscenza usando una base dati sufficientemente ampia. Da questa impostazione nasce l’uso del termine *intelligence* che potrebbe essere correttamente tradotto come “abilità nell’investigazione” e non come ormai si è sedimentato nella vulgata comune “intelligenza” che tanti equivoci ha sollevato in passato e solleva tutt’ora.

Esistono già laboratori di ricerca che stanno lavorando allo sviluppo intelligenze artificiali che mostrano comportamenti devianti come razzismo, violenza e altre psicopatologie più o meno deliranti.

E non si tratta di [fake news](https://www.agendadigitale.eu/tag/fake-news/), né di una puntata particolarmente visionaria di Black Mirror.

La notizia è tremendamente vera, e ci sono validi motivi scientifici per procedere su questa linea di ricerca.

Vediamo i fatti.

Una delle tecnologie di maggiore tendenza degli ultimi anni è sicuramente l’AI ovvero [*Artificial Intelligenc*](https://www.agendadigitale.eu/cultura-digitale/intelligenza-artificiale-unaltra-euforia-di-breve-durata-cosa-insegna-la-storia/)*e* che oggi prende nomi diversi – *machine learning*, *deep learning* – ma in buona sostanza l’idea è sempre quella: la capacità di sviluppare un software autonomo che sia in grado di inferire nuova conoscenza usando una base dati sufficientemente ampia. Da questa impostazione nasce l’uso del termine *intelligence* che potrebbe essere correttamente tradotto come “abilità nell’investigazione” e non come ormai si è sedimentato nella vulgata comune “intelligenza” che tanti equivoci ha sollevato in passato e solleva tutt’ora.

Interazione uomo-AI, gli scenari

Ed è questo il punto centrale: cosa succede se noi facciamo esercitare una intelligenza artificiale piuttosto sofisticata con database dalla provenienza non controllata, meglio se frutto di interazioni umane?

La domanda sembra da film di fantascienza, in realtà è un problema piuttosto serio dato che il mercato delle intelligenze artificiali basate su interazione con la voce, ovvero gli assistenti vocali come Siri, Cortana, Google Assistant, Amazon Alexa, Google Now, viene considerato come uno dei più promettenti dei prossimi anni. Anzi, c’è chi giura che il gadget più regalato del Natale 2018 sarà proprio un assistente vocale. In attesa di avere in casa sistemi come quelli che fanno innamorare il protagonista del film *Her* di Spike Jonze, i ricercatori si stanno ponendo diverse domande su cosa dobbiamo aspettarci da intelligenze artificiali lasciate a interagire senza particolari vincoli o che utilizzano database da zone della rete poco tradizionali. La risposta è piuttosto inquietante: se l’IA interagisce con contenuti polarizzati, il comportamento della IA sarà polarizzato. Detto così sembra scientifico e asettico, ma cosa vuol dire polarizzato?

Facciamo due esempi affascinanti: Tay e Norman.

Tay e gli insulti razzisti

La Microsoft il 23 marzo 2016, forte del successo del progetto Xiaoice sviluppato nella sua consociata cinese, decise di rilasciare al pubblico, Tay – acronimo di *Thinking About You* – il cui scopo era quello di testare in un ambiente *social* le capacità di interazione di un chatbot basato su intelligenza artificiale. In pratica Tay era un software che imitava le interazioni linguistiche di una ragazza americana di 19 anni che installato su Twitter era in grado di interagire con i tweet che le mandavano gli utenti. Ciò che accadde fu che molti utenti cominciarono ad interagire con l’intelligenza artificiale in maniera provocatoria con insulti, osservazioni violente, epiteti offensivi e la reazione dei Tay fu quella di rispondere attraverso messaggi razzisti, violenti, politicamente scorretti. Il risultato fu che la [Microsoft fu costretta a chiudere il progetto Tay](https://socialhax.com/2016/03/24/microsoft-creates-ai-bot-internet-immediately-turns-racist/) dopo solo 16 ore che era stato diffuso al grande pubblico, non prima che l’intelligenza artificiale avesse inviato messaggi come “Hitler aveva ragione”, “Neri e messicani sono le razze peggiori”, “Bush è stato l’autore dell’11 settembre”. Il dibattito sollevato da Tay è stato piuttosto acceso, soprattutto perché non è chiaro se abbia sviluppato questa personalità razzista come imitazione dei messaggi ricevuti – e quindi fosse una funzione programmata nell’algoritmo – oppure se sia stato un comportamento emergente frutto di un adattamento complesso all’ambiente – e quindi una funzione “nuova” sviluppata autonomamente dall’intelligenza artificiale.

Norman e il test delle macchie di Rorschach

Il problema delle intelligenze artificiali che apprendono in contesti sociali controversi è reso ancora più evidente da [Norman](http://norman-ai.mit.edu/). Norman – il cui nome è un tributo a Norman Bates del film *Psycho* di Alfred Hitchcock – è un progetto dello Scalable Cooperation del [MIT Media Lab](https://www.media.mit.edu/groups/scalable-cooperation/overview/)che ha lo scopo di testare il tipo di personalità sviluppata da una intelligenza artificiale sulla base di uno specifico dataset. L’idea dell’esperimento è quello di prendere un classico algoritmo di deep learning in grado di generare una descrizione testuale di un’immagine, lasciare esercitare l’algoritmo con i contenuti presenti in un macabro canale Reddit specializzato in immagini di persone morte e mettere a confronto Norman con una intelligenza artificiale simile che però si è esercitata con un dataset molto più tradizionale (e meno cruento).

Come verificare il risultato dei cicli di apprendimento delle due intelligenze artificiali? Con lo stesso modo con cui si studia la mente di uno psicopatico: con il test proiettivo delle macchie di Rorschach. Chiedendo alle due intelligenze artificiali di interpretare le macchie del test (ovvero darne una descrizione testuale) i risultati del training di Norman sono piuttosto evidenti: dove l’AI vede un dettaglio di un vaso con fiori, Norman vede un uomo sparato a morte, dove l’AI vede una foto in bianco e nero di un uccellino, Norman vede un uomo gettato in una macchina impastatrice, dove l’AI vede una persona che tiene un ombrello in aria, Norman vede un uomo colpito a morte davanti sua moglie urlante. In pratica la frequentazione di uno dei luoghi più macabri della rete, ha reso Norman una intelligenza artificiale psicopatica.

La lezione di Tay e Norman

Che tipo di insegnamenti possiamo trarre da Tay e da Norman? Essenzialmente due. In primo luogo che bisogna stare molto attenti al tipo di dataset con cui vengono testate le intelligenze artificiali perché c’è il rischio che l’uso di dati culturalmente sensibili possano dare conseguenze impreviste nello sviluppo delle intelligenze artificiali. Immaginate se Tay o Norman fossero stati progetti pilota nello sviluppo di prodotti come assistenti vocali intelligenti, gadget interattivi o giocattoli.

In secondo luogo che è necessario dotare gli algoritmi delle intelligenze artificiali di un meccanismo di controllo in grado di rilevare distorsioni nell’apprendimento e nel comportamento e che siano in grado di far intervenire i responsabili del prodotto. In pratica una versione un po’ più realistica delle tre leggi della robotica di Asimov.

Ma, senza dubbio quello che emerge con forza da questi casi studio è che noi programmiamo gli algoritmi con le stesse debolezze dell’animo umano su cui Nietzsche ci ha aperto gli occhi: se guardi troppo nell’abisso, l’abisso guarderà dentro di te.

Anche se sei una intelligenza artificiale.

[Davide Bennato](javascript:void(null))

Sociologia dei media digitali

Relazioni sociali e processi comunicativi del web partecipativo

- disponibile anche in [ebook](http://www.laterza.it/schedalibro.asp?isbn=9788858103470)

|  |  |
| --- | --- |
| Edizione: | 20175 |
| Collana: | Manuali Laterza [320] |
| ISBN: | 9788842097716 |
| Argomenti: | [Scienze della comunicazione](https://www.laterza.it/index.php?option=com_laterza&task=cercacatalogo&&contesto=varia&Itemid=97&argomento=6-4), [Sociologia della cultura](https://www.laterza.it/index.php?option=com_laterza&task=cercacatalogo&&contesto=varia&Itemid=97&argomento=18-2) |

* Pagine 192
* 20,00 Euro
* [Acquista](http://www.laterza.it/index-shopping.asp?azione=inserimento&isbn=9788842097716)

[embed](javascript:void(null)) | [condividi](http://www.addthis.com/bookmark.php?v=250&username=xa-4c5ece603fc6431f)

* [In breve](javascript:void(null))
* [Indice](javascript:void(null))
* [Articoli e news](javascript:void(null))

**In breve**

Milioni di persone si informano e interagiscono fra loro attraverso l'uso di internet. Ognuno a suo modo partecipa alla messa in rete di notizie, ma anche alla trasformazione di questi strumenti di comunicazione e di socializzazione. Blog, wiki, social network sono – soprattutto – strumenti di relazione sociale. Il web partecipativo costringe quindi a un profondo ripensamento dei concetti classici della sociologia della comunicazione.  
Davide Bennato propone una analisi approfondita dei diversi strumenti e delle piattaforme note al grande pubblico, da Facebook a Youtube, ed esamina le conseguenze etiche e sociali dell'uso delle nuove tecnologie