Automi e persone

Introduzione all’etica dell’intelligenza artificiale e della robotica

a cura di

Fabio Fossa

Viola Schiaffonati

Guglielmo Tamburrini

Indice

[Capitolo 1](#_Toc66178205)

[Etica dei sistemi intelligenti e autonomi: una mappa per orientarsi 4](#_Toc66178206)

*Parte prima*

Decisioni e responsabilità

[Capitolo 2](#_Toc66178207)

[Apprendimento automatico e decisione umana 12](#_Toc66178208)

[Capitolo 3](#_Toc66178209)

[Medicina digitale e IA: profili etici 32](#_Toc66178210)

[Capitolo 4](#_Toc66178211)

[Etica dei veicoli autonomi 45](#_Toc66178212)

*Parte seconda*

Persone e istituzioni

[Capitolo 5](#_Toc66178213)

[Governance algoritmica: sorveglianza, censura e diritti fondamentali 65](#_Toc66178214)

[Capitolo 6](#_Toc66178215)

[Cybersicurezza 81](#_Toc66178216)

[Capitolo 7](#_Toc66178217)

[Armi autonome e controllo umano significativo 96](#_Toc66178218)

[Capitolo 8](#_Toc66178219)

[Sostenibilità ambientale della società dell’informazione 112](#_Toc66178220)

*Parte terza*

Persone e interazioni

[Capitolo 9](#_Toc66178221)

[Robotica sociale: persuasione, inganno e etica del design 129](#_Toc66178222)

[Capitolo 10](#_Toc66178223)

[Sex robots 145](#_Toc66178224)

[Capitolo 11](#_Toc66178225)

[Le implicazioni etiche degli usi educativi e didattici dei robot 160](#_Toc66178226)

[Capitolo 12](#_Toc66178227)

[Etica, videogiochi e Gamification 174](#_Toc66178228)

*Parte quarta*

Segnavia

[Capitolo 13](#_Toc66178229)

[Codici etici e documenti di indirizzo 190](#_Toc66178230)

[Capitolo 14](#_Toc66178231)

[Orizzonti 205](#_Toc66178232)

[Bibliografia 219](#_Toc66178233)

[Le autrici e gli autori 249](#_Toc66178234)

Capitolo 1

Etica dei sistemi intelligenti e autonomi: una mappa per orientarsi

Fabio Fossa, Viola Schiaffonati, Guglielmo Tamburrini

La riflessione etica ha accompagnato l’Intelligenza Artificiale (IA d’ora in poi) e la robotica fin dalla loro nascita intorno alla metà del secolo scorso, ma ha assunto una rilevanza sempre maggiore al crescere dell’impatto di questi due settori scientifici e tecnologici sulla vita delle persone e dei gruppi sociali. Dalle attività produttive e commerciali ai servizi sanitari, dal trasporto pubblico e privato alla proposta di una pietanza da farsi consegnare a casa o di un film da vedere dopo cena, risulta difficile individuare un ambito della vita sociale e personale che nel primo scorcio del XXI secolo non sia stato modificato o non possa essere modificato in profondità dall’introduzione di sistemi dell’IA e della robotica.

Questi sistemi possiedono – in varia misura e in determinati contesti d’uso – la capacità di percepire e classificare, formulare previsioni e proporre soluzioni, prendere decisioni e funzionare di conseguenza, eventualmente godendo di una piena autonomia operativa per alcuni compiti. Problematiche etiche nuove riguardano il dispiego di queste capacità per compiere scelte moralmente significative che fino a poco tempo fa erano appannaggio esclusivo di esperti umani. Per esempio, si pensi alla profilazione e classificazione di una persona ai fini della concessione di un prestito bancario, di un’assunzione, di un avanzamento di carriera o di un giudizio in tribunale (O’Neil, 2016). Oppure si pensi a scelte che possono avere un impatto ancora più immediato sull’integrità fisica delle persone, come la diagnosi medica, la selezione e l’attacco di un obiettivo militare da parte di un’arma autonoma o il controllo della navigazione di un’auto a guida autonoma.

Il riferimento all’IA e alla robotica – per quanto esse siano discipline dai contorni piuttosto sfumati – consente di afferrare intuitivamente e in modo più unitario quale sia l’oggetto delle riflessioni presentate in questo libro. È tuttavia importante sottolineare che oggi i sistemi sviluppati utilizzando le tecnologie dell’IA o della robotica sono integrati in un contesto socio-tecnico complesso, i cui confini sono spesso difficili da stabilire. Tali confini non coincidono sempre con quelli delle discipline in questione: come è evidente in alcuni capitoli di questo libro, le problematiche etiche sono spesso connesse a un panorama tecnologico più ampio. Partendo da questa constatazione, con qualche buona ragione si potrebbe utilizzare al posto della locuzione ‘sistemi dell’IA e della robotica’ quella di ‘sistemi intelligenti’. Quest’ultima, infatti, è stata già diffusamente impiegata per indicare un sistema tecnologico basato sulle tecnologie dell’IA o della robotica, ma non necessariamente solo su quelle. Un’altra possibilità è offerta dalla locuzione ‘sistemi autonomi e intelligenti’, che però restringe il campo oltre misura, poiché allude esclusivamente a sistemi basati su tecnologie dell’IA e della robotica che siano capaci di svolgere, in ambienti complessi e senza ricorrere all’intervento diretto dell’essere umano, compiti che chiamano in causa percezione, ragionamento, apprendimento, comunicazione o azione.

Ognuna di queste locuzioni ha evidentemente i suoi pregi e difetti, ed è perciò opportuno demandare ai singoli capitoli il compito di chiarire quale segmento di una classe così ampia e variegata di sistemi e tecnologie verrà in esso più specificamente considerato da una prospettiva etica. L’uso del termine *automi* nel titolo del volume intende proprio fare riferimento alla difficoltà di dare una caratterizzazione precisa e sufficientemente generale ai sistemi e alle tecnologie in questione, ma allo stesso tempo intende richiamare la ricchezza dei diversi significati sottostanti. Gli automi sono spesso intesi come macchine che riproducono l’aspetto esterno e i movimenti di esseri umani o animali, eventualmente stabilendo delle forme di coordinamento intelligente tra percezione e azione. Nell’informatica teorica, la classe degli automi comprende invece svariati modelli di macchine in grado di eseguire algoritmi. Varietà notevoli all’interno di questa classe sono gli automi finiti, gli automi a pila e le macchine di Turing. Non da ultimo, il termine automi vuole richiamare nel titolo la tradizione italiana di riflessione filosofica su questi temi, che ha una sua radice fondamentale nella raccolta di saggi *La filosofia degli automi* curata da Vittorio Somenzi (1965) e pubblicata in edizioni successive a cura dello stesso Somenzi e di Roberto Cordeschi. La riflessione etica sugli automi, solo marginalmente presente all’inizio nel dibattito italiano suscitato dalla pubblicazione di questa raccolta di saggi, ha progressivamente conquistato uno spazio più consistente, all’interno del quale si colloca anche questo volume. Ci limitiamo qui a ricordare, senza pretesa di esaustività, che fin dall’inizio degli anni Ottanta alcuni ricercatori italiani, come gli informatici del CNR di Pisa Diego Latella e Luca Simoncini, e Norberto Patrignani dell’Olivetti di Ivrea, aderirono e parteciparono alle attività della CPSR (*Computer Professionals for Social Responsibility*). Questa associazione internazionale istituì il *Norbert Wiener Award*, assegnato annualmente, dal 1987 e fino al 2008, a ricercatori informatici distintisi per il loro contributo alla promozione di un uso responsabile delle tecnologie informatiche. Successivamente Patrignani promosse il movimento *Slow Tech* (2015) – orientato allo sviluppo di un’informatica centrata sui bisogni degli esseri umani, ambientalmente sostenibile e giusta verso i lavoratori della filiera informatica. Un’altra iniziativa italiana, di chiaro rilievo internazionale, il *First International Symposium on Roboethics. The Ethics, Social, Humanitarian and Ecological Aspects of Robotics*, organizzato da Gianmarco Veruggio e da Fiorella Operto, si svolse presso la villa Alfred Nobel di Sanremo nel gennaio 2004[[1]](#footnote-1), e fu seguito l’anno successivo dal primo progetto europeo sull’etica della robotica (ETHICBOTS, dal 2005 al 2008) coordinato dall’Università di Napoli Federico II, al quale parteciparono vari gruppi di ricerca italiani di diversa provenienza scientifica e culturale[[2]](#footnote-2).

È tutt’ora in pieno svolgimento il dibattito accademico e pubblico sugli indirizzi etici (Tamburrini, 2020; Quintarelli *et al*., 2020; Coeckelbergh, 2020; Dubber *et al.*, 2020) e sulla regolamentazione giuridica (D’Aloia, 2020) da dare alla progettazione e all’uso di macchine che sono capaci di decidere e agire in contesti eticamente rilevanti. I contributi a questo dibattitto sono spesso focalizzati su specifici contesti d’uso e soluzioni locali ai problemi posti dalle decisioni automatiche e dall’autonomia operativa delle macchine. La varietà e la molteplicità di applicazioni delle tecnologie dell’IA e della robotica favoriscono la frammentazione del dibattito etico e giuridico, ostacolando anche la costruzione di una visione d’insieme. Questo volume si propone di offrire una panoramica introduttiva e di ampio respiro sui contesti d’uso delle tecnologie in questione e sui problemi etici ad essi associati, senza tuttavia tralasciare aspetti essenziali che schiudono la porta a una comprensione più approfondita e alla costruzione di una visione d’insieme dell’etica dell’IA e della robotica. Innanzitutto, gli esempi qui illustrati svolgono il ruolo – indispensabile nell’etica applicata – di casi esemplari, che si prestano a generalizzazioni e allo sviluppo di analogie significative con altri casi e contesti d’uso. Inoltre, nell’analisi dei casi esemplari si richiama diffusamente l’impalcatura più astratta dell’etica normativa, dei suoi principi e punti di riferimento valoriali: dall’etica dei doveri e dei diritti fondamentali delle persone, all’etica delle conseguenze e della virtù. Infine, si discute la rilevanza per l’etica applicata delle caratteristiche e dei limiti delle tecnologie che abilitano le nuove capacità decisionali e di autonomia operativa delle macchine. Tra le questioni collegate a tali caratteristiche e limiti figurano, ad esempio, problemi di gestione responsabile dell’incertezza associata alle decisioni e alle previsioni formulate da sistemi basati su tecniche di apprendimento automatico; oppure, problemi di comprensibilità dei soggiacenti processi di apprendimento, di elaborazione dell’informazione e di spiegazione delle risposte ottenute; e, ancora, problemi di attribuzione di responsabilità derivanti dalla difficoltà di prevedere e controllare il comportamento di sistemi robotici che, valicando i confini della produzione industriale e di altri ambienti rigidamente strutturati, si inseriscono nei contesti più dinamici e irregolari della nostra vita sociale.

In definitiva, il volume si propone di analizzare le problematiche etiche che emergono in svariati contesti d’uso e di inquadrarli in uno spazio concettuale di più ampio respiro. I casi esemplari qui esaminati consentono di familiarizzare con l’uso concreto di strumenti di analisi etica, epistemologica e socio-tecnologica, ma anche di sviluppare le competenze richieste per riutilizzare gli stessi strumenti, con opportune interpretazioni e adattamenti contestuali, nell’analisi di nuovi sistemi che a mano a mano emergono dallo sviluppo incalzante delle tecnologie interessate. Grazie al suo carattere introduttivo, questo libro è destinato a tutti coloro che – senza avere una preparazione specialistica – desiderano orientarsi tra le questioni etiche e i problemi morali che l’IA e la robotica pongono alla società nel suo complesso, e che possono affacciarsi nella vita quotidiana di ognuno.

Il libro può essere proficuamente utilizzato per scopi didattici, come strumento per studenti che frequentano corsi di studio sia umanistici sia scientifici e ingegneristici, i quali devono o desiderano cimentarsi con siffatte problematiche multidisciplinari. Negli ultimi anni, infatti, sono stati istituiti diversi insegnamenti a livello universitario che hanno per oggetto l’etica dell’IA e della robotica e, più in generale, delle tecnologie digitali. È importante sottolineare che insegnamenti di questo tipo sono stati introdotti anche in vari corsi di studio universitari centrati sulle discipline di area STEM (o STIM, volendo usare un acronimo formato da parole in italiano: Scienza, Tecnologia, Ingegneria e Matematica). A una solida formazione specialistica STIM viene così ad affiancarsi una prospettiva formativa diversa, offerta dalle discipline umanistiche, le quali consentono di acquisire strumenti di analisi etica, epistemologica e sociale della scienza e della tecnologia. Questo promettente potenziamento del sistema formativo italiano guarda lontano, verso la formazione di nuove figure di scienziati e tecnologi che sappiano coniugare competenze professionali con una forte consapevolezza delle implicazioni etiche e sociali del loro lavoro.

I capitoli successivi sono organizzati in quattro parti. La prima, intitolata *Decisioni e responsabilità*, comprende tre capitoli. Nel capitolo 2 Teresa Scantamburlo e Giovanni Grandi forniscono un’introduzione – accessibile, chiara e insieme rigorosa – all’apprendimento automatico, delineando le sfide e le opportunità che esso pone nell’ambito dei processi decisionali e delle relative responsabilità degli esseri umani. Questo capitolo apre una finestra sulla novità tecnologica che in anni recenti ha maggiormente consentito di incrementare la pervasività delle tecnologie dell’IA e della robotica, sollevando allo stesso tempo molte delle nuove problematiche etiche oggi in discussione. Nelle loro conclusioni, richiamando un’interessante cornice concettuale proposta da Paul Ricoeur, gli autori sottolineano l’importanza di un approccio pluralistico, che tenga debitamente conto di varie componenti dell’etica normativa e applicata, per poter affrontare le questioni sollevate dalle nuove capacità decisionali delle macchine. La questione del pluralismo etico, declinata in varie forme attraverso tutto il volume, stabilisce un collegamento con il capitolo 3, dedicato al caso esemplare degli autoveicoli a guida autonoma, un dominio applicativo che si trova alla confluenza di IA e robotica. In questo capitolo, Fabio Fossa e Guglielmo Tamburrini analizzano dalle prospettive dell’etica delle conseguenze e dell’etica dei doveri la questione delle scelte morali che possono porsi nel caso di collisioni inevitabili e dei relativi problemi di distribuzione di responsabilità. Si affrontano inoltre le problematiche etiche relative alla condivisione del controllo della navigazione tra macchine ed esseri umani, alla libertà di guidare, alla protezione della privacy dei passeggeri, alla sostenibilità ambientale dei veicoli a guida autonoma. Nel capitolo 4, infine, Alessandro Blasimme fornisce un quadro d’insieme delle applicazioni della tecnologia digitale e dell’IA in ambito medico e illustra le questioni etiche che tali applicazioni sollevano: la crescente circolazione e la protezione di dati sanitari personali generati tanto in ambito clinico che attraverso dispositivi mobili personali; l’automazione delle decisioni cliniche di natura diagnostica, prognostica e terapeutica; l’impatto sulla relazione medico-paziente e i cambiamenti, in parte inediti, che la medicina digitale imprime sul modo di concepire sia la salute umana sia le responsabilità individuali, professionali e collettive che la sua difesa comporta.

Quattro capitoli compongono la seconda parte, intitolata *Persone e istituzioni*. Nel capitolo 5, Francesca Musiani offre una panoramica sulla cosiddetta governance algoritmica, ossia sull’uso degli algoritmi come mezzo di regimentazione delle relazioni sociali e di esercizio del potere nel mondo digitale. Si evidenzia in particolare l’impatto della governance algoritmica su sorveglianza, censura e diritti fondamentali, illustrandone aspetti rilevanti a livello sia dei singoli individui sia delle istituzioni politiche e sociali. Il capitolo 6 di Michele Loi discute delle questioni etiche relative alla cybersicurezza individuale e collettiva, nel quale gli strumenti dell’apprendimento automatico e dell’IA giocano un ruolo crescente, a scopi sia di protezione sia di violazione della cybersicurezza. La cybersicurezza è un altro dominio emergente che richiede, come argomenta efficacemente Loi, una riflessione sui modi eticamente giustificabili di bilanciare una pluralità di valori morali che non possono essere tutti promossi, oppure tutti promossi in massimo grado, oppure ancora tutti promossi nella stessa misura. Nel capitolo 7, Viola Schiaffonati e Guglielmo Tamburrini discutono le questioni etiche sollevate dalle armi autonome, e cioè dai sistemi d’arma che l’IA e la robotica abilitano alla selezione e all’attacco di obiettivi militari senza richiedere in corso d’opera nessun intervento da parte di un essere umano. Il capitolo si concentra sui problemi deontologici che le armi autonome pongono rispetto alle possibili violazioni del diritto internazionale umanitario, dei dettami della teoria della guerra giusta e della dignità delle potenziali vittime. Trasversale e decisiva per l’analisi di tali questioni è l’incertezza epistemica che limita la capacità di prevedere il comportamento delle armi autonome in contesti bellici dinamici e poco strutturali. Su questa base si discute la richiesta di esercitare un controllo umano significativo su tutti i sistemi d’arma e di garantirne il rispetto con una opportuna regolamentazione internazionale. Infine, nel capitolo 8, Federica Lucivero affronta un problema finora poco trattato nel dibattito sull'etica del digitale e in particolare dell’IA: la sostenibilità ambientale dei sistemi hardware e software. Da un lato, le tecnologie digitali possono rappresentare una risorsa fondamentale per lo sviluppo sostenibile – basti pensare all’impiego di dati sul traffico urbano per ottimizzare l’uso della rete viaria e ridurre l’inquinamento, di *smart grid* per ottimizzare i consumi energetici domestici, di modelli sviluppati mediante apprendimento automatico per stabilire in anticipo la quantità di prodotti di cui hanno quotidianamente bisogno negozi alimentari e consumatori per ridurre al minimo gli sprechi. Da un altro lato, elaborare e immagazzinare grandi quantità di dati, così come costruire e smaltire i dispositivi digitali, sono tutte azioni che presentano impatti ambientali e sociali non trascurabili. Anche in questo caso si pongono perciò problemi di bilanciamento di valori etici e di individuazione di soluzioni eque per garantire la sostenibilità ambientale dell’ICT e dell’IA, regolando e distribuendo l’accesso di aziende, ricercatori e amministrazioni pubbliche a risorse limitate di immagazzinamento ed elaborazione dell’informazione.

La terza parte, composta da quattro capitoli, si intitola *Persone e interazioni*. Nel capitolo 9 Fabio Fossa introduce alcune delle questioni morali legate alla robotica sociale e ai suoi orizzonti. L’adozione di robot per l’assistenza agli anziani nelle loro case, per il monitoraggio dei pazienti in ospedale, per l’accoglienza dei clienti in un centro commerciale, e così via, dipende dallo sviluppo di tecnologie che rendano l’interazione tra esseri umani e sistemi tecnologici facile, coinvolgente ed efficace. L’intreccio inestricabile della dimensione sociale e della dimensione più propriamente tecnologica nella robotica sociale genera diversi problemi di carattere morale. L’autore si concentra in particolare sulle questioni della manipolazione e dell’inganno. Fino a che punto è lecito ingannare l’utente, generando ad arte la sensazione che stia interagendo con un essere senziente? Fino a che punto è lecito influenzarne implicitamente il comportamento, fosse anche per spingerlo ad agire in modi ritenuti socialmente desiderabili? Temi analoghi si collocano anche sullo sfondo del capitolo 10 di Maurizio Balistreri sui *sex robots*, dove si illustra innanzitutto che cosa essi sono. Si discutono inoltre diverse ragioni che possono motivare all’acquisto di questi robot, cercando di mostrare come il loro impiego non sia necessariamente immorale. Si discute inoltre se sia un errore categoriale parlare dello stupro di un sex robot e se la loro diffusione possa favorire un incremento delle violenze sulle donne o promuovere nei fruitori l’immagine della donna come oggetto sessuale e sempre disponibile, visto che la maggior parte degli attuali sex robots ha fattezze femminili. Nel capitolo 11 Edoardo Datteri e Luisa Zecca forniscono una panoramica sugli scenari d’uso attuale o futuribile dei robot per finalità educative e didattiche, focalizzandosi su questioni etiche relative alla tutela dei dati personali dei discenti, all’attribuzione ingannevole di stati mentali ed emozioni ai robot sociali, alla possibilità di delegare ai robot sociali il compito di educatore e insegnante. Essi mettono in evidenza le questioni empiriche e concettuali ancora aperte che è necessario affrontare per valutare la liceità di questi scenari d’uso in una prospettiva di etica deontologica e di etica delle conseguenze, con particolare riferimento ai loro effetti sull’apprendimento degli studenti e sul loro sviluppo cognitivo, emotivo e relazionale. Francesca Dagnino, Marcello Passarelli e Donatella Persico, infine, presentano nel capitolo 12 le principali problematiche etiche legate ai videogiochi, che possono comportare anche l’uso di tecniche di IA per realizzare sistemi più dinamici e attrattivi, e alla cosiddetta *gamification*. Gli autori prendono in considerazione questioni legate al “contenuto violento” dei videogiochi e alla possibilità che essi incoraggino la violenza nei giocatori; alla dipendenza, discutendo se il potenziale di intrattenimento offerto dai videogiochi e da alcune meccaniche di gioco portino gli utenti a varcare la linea sottile tra il passatempo e la dipendenza; alla valutazione delle rappresentazioni di genere e dei contenuti narrativi dei videogiochi alla luce del principio di non discriminazione; allo sviluppo di videogiochi che promuovano l’inclusione in termini di accessibilità; e, infine, allo sviluppo di videogiochi come strumenti per promuovere valori come la giustizia sociale, l’uguaglianza o la sostenibilità ambientale.

L’ultima parte del volume, intitolata *Segnavia*, racchiude i due ultimi capitoli. Nel capitolo 13 Fabio Fossa e Viola Schiaffonati discutono dei codici etici e di comportamento come strumenti utili a stimolare il giudizio, più che come sistemi perfetti di valori e raccomandazioni che corrono il rischio di scavalcare l’autonomia morale dei singoli. Si offre anche una panoramica su diverse tipologie di codici etici e di condotta, da quelli scritti con il proposito di aumentare la consapevolezza morale o identificare norme e valori per una determinata professione, fino a quelli scritti per l’esigenza primaria di migliorare l’immagine di un’associazione o di un’azienda. Viene discusso poi il ruolo cruciale delle istituzioni pubbliche per guidare il futuro della tecnologia verso direzioni più eque e sostenibili, poiché le realizzazioni tecnologiche sono il risultato di processi sociali articolati e possono dispiegarsi solo nel contesto di complesse strutture politiche e istituzionali che ne normano e sostengono lo sviluppo. Infine, nel capitolo 14 i curatori di questo volume congiuntamente presentano alcuni temi e problemi etici dei quali non si è potuto dare più ampiamente conto nei capitoli precedenti: le implicazioni etiche dell’automazione di attività lavorative abilitata dalle tecnologie dell’IA e della robotica; il potenziamento delle capacità mentali e motorie degli esseri umani che potrebbe essere indotto dalle stesse tecnologie; la cosiddetta *machine ethics*, ovverosia il programma di ricerca che si propone di dotare le macchine di qualche competenza morale; l’eventuale attribuzione di diritti e doveri a sistemi dell’IA e della robotica.

*Parte prima*

Decisioni e responsabilità

Capitolo 2

Apprendimento automatico e decisione umana

Giovanni Grandi, Teresa Scantamburlo

1. Introduzione

La vita umana richiede continuamente di prendere delle decisioni, che possono riguardare questioni individuali e private, ordinarie o di portata esistenziale (dalla scelta del modo di impiegare una serata a quella degli studi, della professione o del consolidamento degli affetti) così come questioni di tipo gestionale e di livello pubblico (dalle politiche aziendali, ai protocolli in ambito sanitario, alle modalità di selezione del personale fino alle politiche governative internazionali). In molti di questi contesti hanno trovato spazio anche gli algoritmi che, grazie all’enorme disponibilità di dati e la crescente potenza di calcolo, possono supportare le decisioni più disparate, da cosa mangiare per cena a quale candidato selezionare per la propria azienda.

Questo capitolo descrive come l’IA e, in particolare, le tecniche di apprendimento automatico possono essere usate nei processi decisionali fornendo esempi in settori diversi come la selezione del personale o la predizione del rischio di recidiva. Il capitolo intende suggerire, da un lato, le coordinate essenziali per comprendere i meccanismi di funzionamento degli algoritmi di apprendimento e il loro impatto nei processi decisionali, e dall’altro, evidenziare una possibile analisi delle problematiche emergenti in prospettiva morale. In particolare, a partire dalla lezione di Paul Ricoeur (1990), si evidenzieranno tre nodi su cui riflettere per poter promuovere un esame critico e uno sviluppo responsabile dell’IA nei processi decisionali: il rinvio al senso complessivo dell’agire, l’attenzione alla possibilità di universalizzare le norme morali, la capacità di tener conto della singolarità degli attori e delle situazioni.

1. Macchine che decidono?

L’uso dell’IA per supportare le decisioni umane non è una novità legata a progressi degli ultimi anni. In ambito medico, sin dagli anni ‘80 si introducono sistemi che possono essere consultati per la formulazione di diagnosi e il trattamento di malattie, o per la risoluzione di guasti in macchine complesse – si pensi ai sistemi esperti (Buchanan, Shortliffe, 1984). Ma le applicazioni più recenti si differenziano per l’approccio e i metodi utilizzati. Mentre nei sistemi di decisione più tradizionali si predilige un approccio *top-down*, basato su conoscenze e regole di inferenza stabilite da esperti umani, i metodi attualmente più in uso adottano un approccio *bottom-up*, basato sulla capacità di apprendere dall’esperienza le regole per inferire nuova conoscenza. Quest’ultimo approccio è tipicamente associato all’area dell’apprendimento automatico (in inglese *machine learning*), un campo dell’IA applicato con successo a svariati tipi di problemi come la classificazione e la predizione.

* 1. Apprendimento supervisionato: alcuni concetti chiave

L’apprendimento automatico (*machine learning*) è un problema studiato sin dalle origini dell’IA e già a metà degli anni ‘50 esistono metodi per il riconoscimento di semplici forme geometriche (Selfridge, 1955) e per il gioco della dama (Samuel, 1959).

In generale, l’apprendimento viene considerato una componente essenziale nello studio e nel progetto di macchine intelligenti, insieme alle euristiche e al ragionamento deduttivo[[3]](#footnote-3). L’assunto di base è che un programma capace di apprendere dall'esperienza può adattarsi a situazioni nuove ed eseguire compiti diversi senza aver bisogno di istruzioni esplicite. Infatti, mentre nella programmazione convenzionale, il programmatore deve definire con precisione le istruzioni che l'algoritmo deve eseguire al verificarsi di determinate condizioni (ad esempio seguendo la classica procedura *if-then-else*), nell'apprendimento automatico l’algoritmo, dato un insieme di dati sufficientemente grande, elabora un modello matematico che predice in modo autonomo la risposta più accurata a seconda dell’input ricevuto.

Una delle problematiche più classiche e popolari dell’apprendimento automatico è la classificazione. Tipicamente, lo scopo di un algoritmo di classificazione è predire la classe di appartenenza di un oggetto sconosciuto scegliendo da un insieme di classi già note. Ma come definire una classe? Ad esempio, come descrivere la categoria dei cani o dei gatti? Nell’apprendimento automatico l’approccio alla definizione delle classi è di tipo ostensivo e si serve di esempi “etichettati”, ossia oggetti descritti da un insieme di attributi o proprietà e accompagnati dal nome della classe di riferimento[[4]](#footnote-4).

Se volessimo descrivere il meccanismo con cui una macchina impara a classificare, potremmo benissimo immaginare un bambino allo zoo che mediante l’indicazione diretta di un insegnante (es. “questa è un’anatra” e “questo è un cigno”) impara ad associare ciascun animale alle specie di appartenenza[[5]](#footnote-5). All’inizio il bambino ascolterà l’insegnante e un po’ alla volta proverà a individuare le specie degli animali senza ricevere alcun suggerimento, ad esempio sulla base delle similarità colte con gli esempi già visti. Le risposte errate gli serviranno poi ad aggiustare il meccanismo di associazione così da migliorare le risposte future.

Nell’apprendimento automatico la presenza di un supervisore che indica le classi di appartenenza degli oggetti è implicita e concretamente rappresentata dall'insieme di esempi che servono ad “addestrare” l’algoritmo (in inglese *training set*).

Gli esempi possono essere descritti da una lista di attributi (*features*), spesso identificati attraverso un processo di selezione (*feature selection*), e da un’etichetta (*label*) corrispondente alla classe di appartenenza. Ad esempio, un libro può essere descritto dalla lista di parole che questo contiene, un’immagine diagnostica dai pixel che la formano, e un cliente di una banca dalle voci che compongono la sua storia creditizia (es. regolarità nel pagamento delle rate, eventuali debiti, ecc).

Le etichette da associare agli esempi possono variare nel numero. Ad esempio, nel caso di un libro, possono coincidere con il genere letterario e, nel caso di una radiografia, possono riferirsi a diverse patologie. Una situazione più semplice e comune è data dalla classificazione binaria, ossia quando l'etichetta può assumere uno tra due possibili valori (per esempio, “1” o “-1”) in modo da separare gli oggetti di una determinata classe (es. la classe dei clienti “non affidabili”) dagli elementi restanti. Tendenzialmente, il valore positivo non esprime una qualità morale delle istanze ma, più semplicemente, identifica il target della classificazione. Inoltre, si noti che in alcuni casi l’output dell’algoritmo non è una classe ma un numero corrispondente ad una probabilità o un punteggio di previsione da confrontare con un valore soglia sulla base del quale assegnare o meno una categoria.

Lo scopo dell’apprendimento è quello di elaborare una funzione matematica, il cosiddetto “classificatore”, in grado di associare gli attributi degli oggetti alle classi stabilite[[6]](#footnote-6). La funzione viene in parte definita dal programmatore e in parte determinata dall’algoritmo. Ad esempio, il programmatore può decidere che la funzione sia formata dalla combinazione lineare degli attributi che descrivono gli oggetti, e l’algoritmo, durante la fase di addestramento, computa i valori dei coefficienti. All’inizio i coefficienti possono assumere un valore casuale per poi aggiustarsi gradualmente grazie alla misurazione dell’errore commesso dall’algoritmo sugli esempi noti. In sostanza, i valori dei coefficienti vengono modificati in modo tale da minimizzare la differenza tra la classe predetta dall'algoritmo e l’etichetta corretta.

Una volta che l’algoritmo è stato addestrato resta il problema di verificare se la funzione ottenuta è generalizzabile ad altri esempi non ancora noti. Si tratta, dunque, di testare le *performance* dell’algoritmo su un nuovo insieme di esempi, questa volta, sprovvisti di etichetta (*test set*). Le performance di un algoritmo di apprendimento sono generalmente definite in termini di errori commessi sul *test set*. Ad esempio una radiografia classificata come “carcinoma” quando in realtà non lo è genera un falso positivo (ossia un falso allarme), viceversa, si parla di falso negativo.

L’analisi dell’errore ci offre importanti informazioni sul comportamento futuro dell’algoritmo e può essere quantificata in termini probabilistici attraverso alcune metriche come la sensibilità, la specificità, e l'accuratezza[[7]](#footnote-7). Analisi più complesse possono mettere in rapporto, anche graficamente, diversi tipi di errori al variare di certe condizioni. Ad esempio nella cosiddetta curva ROC (*Receiver Operating Characteristic*) è possibile studiare il comportamento dell’algoritmo al variare della soglia mettendo a confronto i falsi allarmi, ossia la frazione di falsi positivi, con gli allarmi “veri” (Fawcett, 2006).

In questa sezione ci siamo soffermati su una forma particolare di apprendimento: quello supervisionato. Ma esistono anche altre tipologie che, ad esempio, non richiedono fasi di addestramento e sfruttano le similarità tra gli oggetti per creare classi omogenee, come l'apprendimento non-supervisionato (noto anche come *clustering*). Una situazione intermedia è costituita dall’apprendimento semi-supervisionato, dove esempi non etichettati sono usati insieme ad un numero limitato di esempi etichettati. Un’altra classe di metodi si basa sulla nozione di rinforzo, ossia ricompense che l'algoritmo riceve in caso di risposta positiva e che tendono ad incoraggiare determinati comportamenti, il cosiddetto apprendimento rinforzato (Mitchell, 2020, pp. 161-177). Sebbene questi e altri approcci rappresentino aree di ricerca promettenti, il paradigma dell'apprendimento supervisionato è indubbiamente quello che ha registrato i maggiori successi nelle ultime competizioni[[8]](#footnote-8) e si è diffuso più rapidamente in diversi settori applicativi.

* 1. La decisione come un problema di classificazione

L’apprendimento automatico può essere formulato in termini di inferenza induttiva. Si consideri la classificazione. Il problema consiste nell'inferire da un insieme di esempi già classificati una regola di decisione sulla base del quale assegnare nuovi oggetti alla classe corretta. Tale regola permette di fare previsioni in condizioni di incertezza, visto che l'esperienza da cui si apprende è parziale e non sempre significativa (si veda nel paragrafo 2.3 il problema del non-determinismo dell’etichetta). Tuttavia questa incertezza può essere “controllata” attraverso alcuni strumenti teorici[[9]](#footnote-9) che permettono di scegliere la regola di decisione ottimale (ma su questo punto torneremo nella sezione 2.3).

Nella classificazione binaria la regola di decisione può essere interpretata come un test diagnostico in grado di dare una risposta positiva o negativa sulla presenza di una determinata condizione o sul verificarsi di un determinato evento. Ad esempio, un algoritmo, una volta addestrato, può prevedere se un’email è un messaggio indesiderato (*spam*), oppure se un paziente è a rischio di recidiva dopo il primo intervento.

Lo schema decisionale di un algoritmo di classificazione può essere applicato in moltissimi contesti e su diverse popolazioni. Da un punto di vista strettamente formale, per l’algoritmo non fa alcuna differenza se ad essere classificati sono delle email o delle persone. Ciò che conta è che vi sia un numero sufficiente di esempi per poter addestrare l’algoritmo e questi siano rappresentativi della popolazione su cui l’algoritmo dovrà poi decidere. Questi requisiti fanno parte degli assunti teorici che caratterizzano la metodologia dell’apprendimento automatico e, come vedremo nella sezione 2.3, non sempre è possibile rispettarli.

L’apprendimento automatico può essere applicato a decisioni di poco conto, come la scelta del prossimo film da vedere o del prossimo acquisto online. É questo il caso dei sistemi di raccomandazione, come YouTube, che, anziché una classe, deve predire un video scegliendo da un insieme di milioni di opzioni in una frazione di secondi (Covington *et al*., 2016). I dati con cui gli utenti vengono descritti sono eterogenei e possono includere milioni di attributi, come i video precedentemente visualizzati, la provenienza geografica, l’età e il genere. L’algoritmo impara a mappare gli utenti ai video sfruttando un principio di similarità secondo cui gli utenti che hanno mostrato un comportamento simile continueranno a mostrarlo anche in futuro[[10]](#footnote-10). Una volta filtrati i video candidati per la raccomandazione, l’algoritmo seleziona quello con la maggiore probabilità di essere cliccato e visualizzato.

Un meccanismo simile a quello di YouTube può essere applicato a decisioni più delicate come la valutazione degli impiegati. Ad esempio, nel suo best seller *Weapons of Math Destruction*, Cathy O’Neil (2016) descrive uno strumento per la valutazione automatica degli insegnanti, IMPACT, usato nelle scuole di Washington DC con lo scopo di assegnare un punteggio a ciascun insegnante sulla base del progresso fatto registrare dagli studenti durante test standardizzati. Alcune testimonianze hanno denunciato che, alla fine dell’anno scolastico, l’impiego dell’algoritmo ha portato al licenziamento degli insegnanti con un punteggio al di sotto di una certa soglia e questa decisione ha riguardato anche insegnanti apprezzati da famiglie e colleghi (cfr. McGuire, 2012; O’Neil, 2016). Modelli simili sono utilizzati anche per la selezione del personale, dove esistono sistemi per ottimizzare il *pre-screening* dei curriculum e per la valutazione dei colloqui. Ad esempio HireVue[[11]](#footnote-11) è un’azienda che fornisce una piattaforma per lo svolgimento di video interviste, dove i candidati, ad una posizione di lavoro o a un corso di laurea, sono valutati automaticamente sulla base di diversi indicatori che riguardano sia la comunicazione verbale che non-verbale (es. espressioni del volto o modi di costruire le frasi).

Applicazioni ancora più problematiche, da una prospettiva sia etica che giuridica, riguardano la giustizia predittiva. In varie parti del mondo si è diffuso l’uso di algoritmi di apprendimento automatico per predire il rischio di recidiva, ossia la probabilità che una persona possa commettere un altro crimine dopo l’arresto. Un caso tipico è decidere se rilasciare o meno su cauzione un soggetto arrestato in attesa della prima udienza considerando il rischio che il soggetto scappi o, peggio ancora, commetta un altro crimine. Negli Stati Uniti è ormai divenuto celebre il sistema di valutazione automatica del rischio COMPAS (Brennan *et al.*, 2009), e simili esempi esistono anche in Europa, come OxRec (Fazel *et al*., 2019) e HART (Oswald *et al.*, 2018).

L’approccio del *machine learning* propone un modello per la decisione basato sull’assunto che sia possibile classificare i soggetti in determinate categorie. Spesso, la classificazione deriva dalla probabilità che un certo individuo sia, per esempio, un candidato di successo, un debitore rischioso oppure un soggetto pericoloso. Il meccanismo è sempre lo stesso: osservare l’esito di decisioni passate attraverso una massiccia raccolta di dati in grado di “rivelare” dei *pattern*, ossia modelli di una determinata categoria che poi diventano modelli per decisioni future (questo studente porterà a termine gli studi nei tempi giusti? Questo cliente riuscirà a restituire i soldi chiesti in prestito? Questo soggetto potrà compiere atti violenti?).

In molti casi l’uso di algoritmi è favorito da protocolli di azione o criteri decisionali già consolidati. Ci sono molti ambiti dove l’impiego di semplici algoritmi per predire la probabilità di un determinato evento e, sulla base di questo, prendere una decisione è considerata una prassi comune ormai da molto tempo[[12]](#footnote-12). Ad esempio, in ambito assicurativo si è soliti individuare la categoria di rischio di un soggetto per determinare l’entità del premio. Una situazione analoga avviene nella valutazione del rischio di un trapianto o della probabilità che un paziente sopravviva ad un intervento. In generale, l’uso dell’apprendimento automatico è “facilitato” in tutte quelle situazioni dove le decisioni si basano su una previsione e l’evento o categoria da prevedere è ben definito e misurabile. Si noti che il più delle volte l’output dell’algoritmo non determina in modo diretto l’esito della decisione umana. Questo si presenta per lo più sotto forma di raccomandazione o evidenza empirica che l’utente o il professionista (per esempio, medico o giudice) prende in considerazione nella scelta finale.

Sebbene l’impiego dell’apprendimento automatico abbia offerto nuove opportunità nei processi decisionali di un individuo o di un’organizzazione, l’applicazione indiscriminata e su larga scala di questa tecnologia ha suscitato perplessità e dubbi sulla sua reale efficacia e convenienza. In particolare, l’uso di algoritmi in decisioni di rilevanza sociale come la profilazione dei disoccupati o l’identificazione di minori a rischio di violenza, ha portato alla luce problemi di opacità e discriminazione, oltre che di sicurezza - si pensi ad esempio alle vulnerabilità dei sistemi di visione artificiale scoperti con l’*adversarial machine learning* (Szegedy *et al*., 2014). Ma ancor più rilevante è la presa di coscienza che tali algoritmi debbano rispondere non solo a dei criteri di accuratezza e robustezza, ma anche a requisiti legali, etici e sociali, capaci di garantire i diritti delle persone e i valori della convivenza civile (Gruppo Indipendente di Esperti sull’IA, 2019).

* 1. Gli assunti teorici dell’inferenza predittiva

Gli algoritmi, così come la matematica più in generale, sono oggetto di un pregiudizio positivo molto comune: il fatto di essere neutrali e di produrre risultati oggettivi. Purtroppo, l'impiego di strumenti matematici e statistici non è sufficiente a garantire la neutralità e l’imparzialità di una decisione. Infatti, un modello matematico è sempre una rappresentazione della realtà e proprio per questo può essere esposto ad errori. Nella costruzione di un modello è normale fare assunzioni e compiere delle scelte che semplificano la realtà rappresentata. Questo succede anche quando si crea una mappa di un territorio dove certe informazioni possono essere escluse di proposito: ad esempio, la cartina dei musei di una città potrebbe omettere la posizione dei bar o dei negozi.

Così, anche l’apprendimento automatico si basa su alcune assunzioni, per lo più di natura statistica, che limitano le sue possibili applicazioni e possono essere all’origine di problemi etici e sociali.

Una prima importante assunzione riguarda il meccanismo con cui viene generato il campione dei dati usati per l’addestramento. Solitamente si assume che gli esempi, intesi come coppie di variabili rappresentanti la lista di attributi e le etichette, provengano da una distribuzione di probabilità sconosciuta e fissa. La conoscenza che possiamo avere di questa distribuzione è solo indiretta e mediata appunto dagli esempi che forniscono un numero limitato di osservazioni. Se questi non sono rappresentativi della popolazione di interesse (es. l’insieme degli studenti che vanno fuori corso) c’è il rischio di individuare delle correlazioni poco significative o fuorvianti, limitate solo a un sottogruppo della popolazione (come si vedrà nella sezione 3, questo aspetto può dare origine a risultati discriminatori).

Nella teoria classica dell’apprendimento automatico l’ideale di un campione rappresentativo è fornito dall’ipotesi che gli esempi siano generati in modo identico e indipendente[[13]](#footnote-13). Purtroppo questo assunto è spesso violato nella realtà per almeno due ragioni. In primo luogo, ci possono essere variazioni nel processo di raccolta dati. Spesso infatti gli esempi utilizzati per la fase di *training* e di *testing* provengono da sorgenti diverse - si pensi ad esempio a colloqui di lavoro svolti in aziende che operano in settori o aree geografiche differenti. Inoltre, anche se la sorgente è la stessa, potrebbero verificarsi cambiamenti nel contesto decisionale che vanno ad influenzare l’esito stesso delle decisioni (ad esempio una crisi economica).

Un secondo fattore di criticità è legato a possibili interdipendenze tra gli esempi. Si immagini di voler valutare le conoscenze acquisite da un gruppo di studenti a scuola attraverso un test. I risultati del test rappresentano un indicatore affidabile solo se si assume che gli studenti abbiano svolto il compito in modo indipendente l’uno dall’altro. Inoltre, si vorrebbe evitare che il risultato fosse correlato a caratteristiche degli studenti non rilevanti per la valutazione, come il codice di avviamento postale o l’etnia.

Un ulteriore assunto riguarda la natura non-deterministica dell’etichetta che può essere associata ad un oggetto per motivi casuali. Infatti, gli esempi potrebbero essere etichettati in modo erroneo - in gergo tecnico si dice che il processo di generazione contiene del “rumore” - si pensi, ad esempio ad una diagnosi sbagliata o un’email classificata erroneamente come spam. Ma un esempio potrebbe anche essere assegnato a più di una etichetta e dunque essere classificato in modi diversi e ugualmente plausibili. Infatti, le correlazioni individuate dal classificatore sono di natura statistica e non offrono alcuna garanzia sul possibile rapporto di causalità (formale[[14]](#footnote-14)) tra gli attributi e le etichette. Si consideri, per esempio, il problema di predire se una persona è maschio o femmina in base all’altezza. Nel caso in cui un soggetto fosse alto 170 cm l’etichetta potrebbe essere tranquillamente sia “maschio” che “femmina”. L’altezza infatti non cattura in alcun modo il concetto di “maschio” o “femmina”[[15]](#footnote-15).

Da questi assunti deriva un problema ben noto nel campo dell’apprendimento automatico: il cosiddetto *overfitting* (sovradattamento)*.* Semplificando, questo problema si verifica quando il modello risulta accurato solo con gli esempi utilizzati nel training e le sue *performance* degradano nella fase di test o in successive applicazioni[[16]](#footnote-16). In gergo tecnico si dice che il modello si è adattato eccessivamente agli esempi e non è in grado di generalizzare. È come se una mappa fosse molto precisa per una determinata zona di una città ma poco dettagliata per altre e, nel complesso, non ci aiutasse ad orientarci.

Per “contrastare” l’*overfitting* è necessario introdurre altre assunzioni che riguardano la classe che si vuole predire e, di conseguenza, la funzione (o regola di decisione) che si vuole apprendere. Questo insieme di assunzioni si chiama *inductive bias* (Mitchell, 1997, pp. 39-45)e consiste nell’esprimere una preferenza sul tipo di regola che l’algoritmo deve apprendere[[17]](#footnote-17). Ad esempio, nel caso di un filtro anti *spam* potremmo preferire le regole che si basano sulla combinazione lineare delle frequenze delle parole. L’*inductive bias* ed è una componente essenziale del *machine learning* che permette all’algoritmo di trovare una funzione un po’ più generale rispetto a quella rispecchiata dagli esempi del *training set*[[18]](#footnote-18). Semplificando, potremmo dire che l’arte del *machine learning* sta proprio nel trovare il bilanciamento ottimale tra il rischio di *overfitting*, introdotto dal campione di dati usati per l'addestramento, con quello di sovrageneralizzazione, dovuto appunto all’*inductive bias*.

1. L’impatto etico e sociale dell’apprendimento automatico nei processi decisionali

Uno dei punti di forza dell’apprendimento automatico è ottimizzare la ricerca del classificatore più accurato facendo leva sugli assunti descritti nella sezione 2.3. Anche se in alcuni casi è possibile “rilassare” le condizioni espresse dagli assunti, non sempre è possibile controllare o prevedere il comportamento di un algoritmo di apprendimento, soprattutto quando il contesto operativo dell'algoritmo si sposta dal laboratorio alla vita reale o in situazioni di grave incertezza (Schiaffonati, 2021). Infatti, come abbiamo già suggerito, molte delle condizioni teoriche sono distanti dalla realtà o possibili in contesti applicativi molto circoscritti. Nell’ambito della visione artificiale, esistono algoritmi che raggiungono livelli di accuratezza molto elevati, se non superiori alle prestazioni umane, in compiti molto specifici, come la segmentazione di immagini o l’identificazione di oggetti presenti nella scena, e su insiemi di dati ben definiti. Purtroppo, congiuntamente a questi risultati promettenti, si sono registrati casi in cui gli algoritmi hanno generato, se non aggravato, situazioni dannose per le persone e la collettività. Un elemento trasversale a queste difficoltà è l’utilizzo di enormi basi di dati costruite e continuamente alimentate dalle tracce digitali che le azioni online lasciano. Questo aspetto suggerisce un livello ulteriore delle questioni di privacy. In ballo, infatti, non c’è “solo” il problema della diffusione di dati personali ma anche e soprattutto le conseguenze che l’elaborazione di tali dati insieme ad altri possono creare per gli individui e la società.

L’impiego sempre più diffuso di algoritmi di apprendimento al fine di valutare un individuo e prendere decisioni sul suo conto richiede un esame attento di tali algoritmi che vada oltre le garanzie di accuratezza e includa la valutazione dei principi normativi fondamentali di una società, quali la giustizia, la legalità e la tutela di diritti (Scantamburlo *et al*., 2019).

Nei paragrafi successivi illustriamo alcuni di questi effetti negativi traendo spunto da esempi o casi che hanno animato il dibattito pubblico. Sebbene gli spunti offerti non esauriscono il dibattito sollevato dall’impatto etico, sociale e giuridico degli algoritmi le problematiche toccate rappresentano alcuni dei nodi principali della discussione contemporanea.

* 1. Opacità, intelligibilità e responsabilità

La trasparenza delle decisioni è un aspetto fondamentale nei processi decisionali delle amministrazioni pubbliche e private. Ospedali, scuole, banche sono soggette a regole precise in materia di trasparenza. Questo perché la trasparenza permette ad un’organizzazione di tenere traccia degli elementi rilevanti che hanno portato ad un determinata scelta, inclusi gli attori coinvolti (la cosiddetta tracciabilità dei processi). Ma la trasparenza è importante anche e soprattutto per coloro che “subiscono” gli effetti di una decisione. Se un’amministrazione è trasparente i cittadini dovrebbero poter accedere ai dati e alle informazioni sulle procedure che li riguardano così da poter fare accertamenti e, se volessero, contestare l’esito della decisione. La trasparenza dunque è una precondizione per poter esercitare il diritto alla privacy, verificare l’imparzialità e la liceità delle procedure. Il rispetto della trasparenza può essere difficile o quantomeno ostacolato dall’agire umano (si pensi alla corruzione) o da fattori sistemici, legati ad esempio a inefficienze della burocrazia. Ma l’inserimento dell'apprendimento automatico nei processi decisionali può aggiungere ulteriori difficoltà.

Un primo aspetto problematico deriva dal fatto che non sempre è possibile rendere pubblici e accessibili le informazioni e i dati che hanno portato ad una determinata scelta. Tra questi potrebbero infatti essere presenti dati sensibili o informazioni la cui pubblicazione danneggerebbe l’interesse dei soggetti coinvolti. Nei casi in cui il proprietario dell’algoritmo impiegato è un’azienda, la trasparenza potrebbe essere in conflitto con il diritto dell’impresa di proteggere la riservatezza dei propri prodotti o servizi e tutelare il vantaggio competitivo. A questo si aggiunge un elemento di opacità che è intrinseco al *machine learning* o perlomeno a buona parte degli algoritmi più in uso. Anche se fosse possibile accedere al codice sorgente dell’algoritmo non è detto che sia possibile comprendere la logica che ha determinato l’output[[19]](#footnote-19). Questo potrebbe essere solo in parte dovuto alla mancanza di competenze tecniche o inesperienza. Infatti, il modo in cui l’algoritmo elabora la risposta ad una decisione non corrisponde necessariamente a un processo umano. Spesso il classificatore individuato dall’algoritmo corrisponde ad una funzione non lineare, talvolta molto complicata, che risulta non intelligibile all’uomo (Burrell, 2016). Ad esempio, di solito una rete neurale profonda (*deep neural network*) determina il proprio output a fronte di miliardi di operazioni aritmetiche su altrettante unità connesse tra loro. La sequenza di tali operazioni non offre alcuna interpretazione della decisione computata, né una spiegazione delle ragioni per cui si è pervenuti a quel determinato risultato[[20]](#footnote-20).

Infine, il processo di sviluppo e di applicazione di un sistema di decisione automatica può essere molto intricato da un punto di vista dell’organizzazione degli attori coinvolti. In molti contesti d’uso lo sviluppatore dell'algoritmo differisce dal distributore e spesso anche dal proprietario dei dati. Inoltre, l’utente dell'algoritmo potrebbe dover rispondere a particolari doveri professionali (per esempio, l’utente potrebbe essere un poliziotto) e non è così scontato integrare i compiti associati all’uso dell'algoritmo con quelli legati alla professione. Dunque l’attribuzione di responsabilità in caso di output scorretto o sfavorevole può essere molto complessa quando il sistema di decisione risulta così frammentato e distribuito su una struttura di compiti e obblighi preesistenti.

* 1. Discriminazione e sistematizzazione delle disuguaglianze

L’opacità non è l’unico limite connaturato al *machine learning.* Un altro aspetto problematico riguarda il possibile effetto discriminatorio degli algoritmi di apprendimento[[21]](#footnote-21). Tale rischio è legato alle assunzioni statistiche descritte in sezione 2.3 e non necessariamente alla cattiva volontà del programmatore. Spesso un output ingiusto è un effetto causato involontariamente dai limiti degli assunti e dalla scarsa capacità dei programmatori e dei distributori di anticiparli e valutarli durante tutto il ciclo di vita dell’algoritmo.

Come anticipato, gli esempi usati per addestrare un sistema di apprendimento possono differire dall’insieme di dati su cui l’algoritmo andrà ad operare oppure essere semplicemente poco rappresentativi della popolazione oggetto di analisi. Quando nella composizione del campione ci sono degli sbilanciamenti (per esempio, una prevalenza di maschi caucasici) è altamente probabile che l’algoritmo sia molto meno accurato con la parte di popolazione meno rappresentata. Se lo sbilanciamento è legato ad alcune caratteristiche sensibili della persona, come il genere o l’etnia, questo può sfociare in una discriminazione sociale. Ad esempio, uno studio condotto sui principali sistemi di riconoscimento facciale in commercio ha dimostrato come le prestazioni dell'algoritmo degradano significativamente in funzione del genere e del colore della pelle del soggetto classificato (Buolamwini e Gebru, 2018).

Un algoritmo può essere addestrato su un insieme di dati che incorporano dei pregiudizi impliciti. Ad esempio, se nei dati usati per l’addestramento un’etichetta è ampiamente correlata ad una particolare etnia, il classificatore tenderà a riproporla anche in fase di test o successive esecuzioni. Un simile caso è stato dibattuto nell’ambito della giustizia predittiva, dove l’algoritmo COMPAS è stato giudicato discriminatorio verso gli afro-americani. In particolare, un’analisi condotta da ProPublica (Angwin, 2016) ha dimostrato come gli imputati afro-americani avessero quasi il doppio di probabilità di essere erroneamente classificati come a rischio più elevato rispetto agli imputati bianchi (45% contro il 23%). Purtroppo, per contrastare un simile effetto può non essere sufficiente eliminare l'attributo correlato all’output perché, in alcuni casi, potrebbero esserci correlazioni indirette. Ad esempio, tolta l’informazione sull’etnia dall'insieme degli esempi potrebbe rimanere traccia della correlazione “incriminata” attraverso gli attributi restanti, come il codice postale. Così, anche se ad una prima valutazione i dati sembrano non includere attributi critici non è da escludere che vi siano interdipendenze indirette.

Come anticipato in sezione 2.3, queste correlazioni sono introdotte in maniera spontanea e spesso riflettono pregiudizi preesistenti. Ad esempio, Caliskan *et al*. (2017) ha mostrato che gli algoritmi addestrati su estesi insiemi di testi presi dal Web riproducono gli stessi stereotipi umani studiati dalla psicologia. Anche se la discriminazione originata dagli algoritmi si pone in continuità con alcune cattive abitudini e fenomeni già noti nella società, un aspetto da non sottovalutare riguarda il rischio di sistematizzare le disuguaglianze. Infatti, gli algoritmi di nuova generazione, come le reti neurali profonde, traggono vantaggio dai cosiddetti big data ma d’altra parte l’avere a disposizione una sovrabbondanza di dati complica l'identificazione di possibili pregiudizi o distorsioni derivanti dalle correlazioni istituite tra gli stessi.

Inoltre, le decisioni influenzate da un pregiudizio, se non rilevate, possono essere riutilizzate per futuri addestramenti che contribuiranno, a loro volta, ad intensificare l’effetto discriminatorio dell’algoritmo.

* 1. Profilazione, autonomia e condizionamenti

Finora abbiamo considerato effetti negativi che ricadono sui soggetti terzi che in un certo senso subiscono il processamento automatico. Ma esistono anche elementi di preoccupazione nell’interazione diretta tra un sistema di apprendimento automatico e l’utente (Burr *et al*., 2018). Qui i rischi derivano dalla capacità dell’algoritmo di influenzare le azioni dell’utente indirizzandolo verso determinati comportamenti. Normalmente, il modo in cui l'algoritmo “guida” l’utente è soft e spesso benevolo (la *nudging theory* – cfr. Thaler e Sunstein, 2008; CAP. ROBOTICA SOCIALE pp. – fa frequentemente da sfondo a tali applicazioni). L’algoritmo suggerisce ciò che “pensa” sia più utile o di maggiore interesse per l’utente. I sistemi di raccomandazione sono un esempio quotidiano dove l’utente mediamente si sente sollecitato più che forzato verso una certa direzione. Ma seppur gentile, questo tipo di guida può essere insidiosa e trasformarsi in una subdola forma di manipolazione. Le correlazioni intercettate nei dati possono essere sfruttate a vantaggio di alcuni come suggerito dalle accuse avanzate all’azienda *Cambridge Analytica,* una società di consulenza che avrebbe usato tecniche di *machine learning* per influenzare le elezioni in USA. Il caso ha messo in luce come le risposte di milioni di persone ad un test psicometrico[[22]](#footnote-22), unite a dati demografici e comportamenti online (reazioni ai post, condivisioni, tempo di permanenza, ecc.) possano predire alcuni tratti delle personalità e così contribuire a definire l’orientamento politico degli utenti di Facebook (Menietti, 2018).

In alcuni contesti l’interazione continuativa con sistemi di apprendimento può contribuire allo sviluppo di una fiducia sproporzionata verso le capacità del sistema e l'output suggerito. Quando questo accade l’utente può sviluppare una sorta di dipendenza verso il sistema che poi si trasforma in atrofia decisionale e rischia di indebolire capacità e competenze possedute dal soggetto ma nel tempo sempre meno esercitate, come nel caso esaminato da Cabitza (2021), del deterioramento delle capacità diagnostiche di un medico nel caso in cui la diagnosi venisse affidata più sistematicamente ad una macchina. Ma simili situazioni potrebbero interessare anche attività ordinarie, come la scelta di un corso universitario o di un itinerario di viaggio, e coinvolgere soggetti fragili o meno attrezzati a contrastare gli effetti degli algoritmi.

Come nelle difficoltà precedentemente accennate, i condizionamenti e la perdita di autonomia decisionale sono raramente effetti ricercati. Spesso sono conseguenze indirette, legate strutturalmente all’impostazione teorica e pratica del *machine learning*. Anche in questo caso, l’indirizzamento dell’utente verso un certo obiettivo è “semplicemente” un riflesso dell’orientamento utilitarista del *machine learning*, volto a massimizzare l’utilità per cui è stato costruito (per esempio, numero di click, tasso di risposte corrette, ecc.).

1. Concetti filosofici e chiavi di lettura per una analisi del decidere

Per esaminare in prospettiva critica il ricorso all’IA e all’apprendimento automatico nell’ambito dei processi decisionali è essenziale anzitutto riconoscere che la problematicità del decidere rappresenta una questione classica del pensiero antropologico e morale.

Già in Platone (V-IV sec. a.C.) l’indecisione rappresenta il “caso studio” da cui trae origine la concezione psico-antropologica che si imporrà per secoli nel pensiero occidentale, ovvero la celebre tripartizione delle facoltà dell’anima – le parti *concupiscibile*, *irascibile* e *razionale* – (Platone, trad. it. p. 1178), a testimonianza del fatto che la necessità di scegliere tra opzioni diverse rinvia non semplicemente all’analisi di una dinamica o di specifiche questioni morali, ma a una indagine ancora più radicale sulla struttura dell’umano. In particolare quel che fin dagli albori della riflessione filosofica costituisce un problema eminente è il fatto che proprio dalle decisioni umane possa derivare il male per la vita personale e sociale. Socrate, come è noto, ne farà fondamentalmente una questione di conoscenza, di corretta e completa informazione potremmo dire in termini contemporanei, ma già Aristotele noterà nell’*Etica Nicomachea* che alle difficoltà dovute alla complessità delle situazioni si aggiungono quelle attribuibili alle manchevoli disposizioni dei decisori – i *vizi* nel suo linguaggio (Aristotele, lib. VII) – e Paolo di Tarso giungerà di lì a poco a osservare che talvolta, pur individuando correttamente il bene da compiere, tuttavia si finisce non di rado per essere autori del suo contrario (*Lettera ai Romani*, 7,15). Dunque, alla questione delle *informazioni disponibili* e delle *abitudini* andrebbe accostato anche un discorso sulla *volontà*, cosa di cui si occuperà lungamente il pensiero medievale, sempre tuttavia alle prese con il medesimo problema, ovvero con l’esperienza per cui dal decidere possa scaturire qualcosa di nocivo.

La riflessione morale moderna si è concentrata su ulteriori questioni, come ad esempio l’individuazione di criteri di universalità a cui sottoporre le proprie decisioni, come nel caso della formulazione del primo imperativo categorico di Kant: «Agisci unicamente secondo quella massima in forza della quale tu puoi volere, nello stesso tempo, che essa divenga una legge universale» (Kant, 1785, trad. it. p. 79). Tuttavia anche la modernità ha continuato a muoversi sempre nel medesimo alveo problematico: il secondo imperativo categorico kantiano individua infatti come invalicabile il riconoscimento e il rispetto della dignità umana – «Agisci in modo da trattare l’umanità, così in te come negli altri, sempre anche come un fine e mai solo come un mezzo» (ivi, p. 88). L’esigenza di questa seconda declinazione deontologica deriva, a parere di Ricoeur (1990), dall’osservazione per cui spesso le decisioni umane sono segnate dalla violenza e dalla sopraffazione, e questo solleciterebbe la riflessione teorica a non dare per scontato ma a esplicitare il riferimento alla dignità come fine.

Occorre perciò anzitutto osservare che la tradizione più classica, dall’antico al moderno, si è molto interrogata più sulla “cattiva” che non sulla “buona” decisione, più preoccupandosi di comprendere come arginare il male che non di misurare potenziali guadagni o vantaggi. Solo la tradizione utilitarista – a partire dalla prospettiva di Bentham (1789) sviluppata poi da Mill (1863) – ponendo come orizzonte dell’agire il raggiungimento della *massima felicità per il maggior numero*, ha in un certo modo depotenziato il problema del male che deriva dalle decisioni umane: che qualche costo ci sia, che qualcuno finisca nel novero degli infelici (o vi rimanga) a causa delle nostre scelte, risulterebbe – rigorizzando questa linea di pensiero, come ha fatto notare Sandel (2009) – qualcosa di fisiologico e di non particolarmente problematico, se controbilanciato dal vantaggio dei molti.

 Il crescente ricorso alle risorse dell’IA nei processi decisionali solleva – come si è visto sopra – diversi problemi di ordine morale. Per poterli affrontare in modo avvertito è opportuno poter fare riferimento ai principali guadagni delle diverse prospettive di pensiero maturate nel corso dei secoli. Senza dubbio la prospettiva utilitarista è un interlocutore stimolante (Russell, 2019) per mettere a fuoco la possibile interfaccia tra la dimensione computazionale e quella etica: rimane infatti suggestiva l’idea di poter realizzare un “calcolo felicifico” (Bentham, 1789, trad. it. p. 91), riuscendo cioè a tradurre tutto ciò che costituisce valore nell’orizzonte umano in un’unica misura, funzionale appunto alla determinazione del concorso delle decisioni alla realizzazione della massima felicità per il maggior numero. Più in generale, la prospettiva consequenzialista, all’interno della quale si iscrive l’utilitarismo classico, prescrive di valutare come buona o cattiva una decisione esclusivamente in base alle sue conseguenze reali o attese (CAP VEICOLI AUTONOMI pp….).

Al tempo stesso non vanno trascurate le linee di pensiero che rinviano al problema più antico e radicale del male e del suo concretizzarsi nella storia e nelle relazioni, a partire dalle decisioni. Le preoccupazioni che accompagnano lo sviluppo dell’IA richiamano infatti molto gli interrogativi più radicali degli antichi e dei moderni: posto che l’attesa condivisa è che lo sviluppo scientifico-tecnologico generi dei vantaggi, come è possibile farsi carico dei mali che potrebbero derivare da decisioni prese con l’ausilio delle macchine o dalle macchine stesse? Non sembra cioè essere sufficiente l’ammissione utilitarista per cui uno scarto ci sarà sempre e che il fatto di averlo minimizzato possa risolvere adeguatamente la questione morale. Le diverse problematiche evidenziate in precedenza – dalla trasparenza degli algoritmi, alla allocazione della responsabilità, dal contrasto al ricorso agli stereotipi – rappresentano in maniera diversa il timore che nel volume dei *big data* e attraverso l’attività classificatoria e selettiva affidata agli algoritmi si perda la possibilità di intercettare il processo di incubazione di una cattiva decisione, dovendola semplicemente prevedere come un prezzo da pagare a fronte di altri indiscutibili vantaggi.

Per discutere di questo timore prendendo in considerazione la varietà delle istanze morali legate ai processi decisionali può essere utile adottare lo sguardo panoramico proposto da Paul Ricoeur, che consente di riordinare le diverse tipologie di questioni attingendo a contributi di diverse tradizioni filosofiche.

1. Istanze morali nei processi decisionali

Seguendo la celebre lezione di Ricoeur (1990) è possibile articolare almeno tre livelli di attenzioni e preoccupazioni relative all’agire umano dal punto di vista dell’analisi dei processi di maturazione di una scelta.

* 1. L’attenzione al senso complessivo dell’agire

Il primo livello di attenzione riguarda la cornice di senso entro cui si muovono i decisori. Per Ricoeur si tratta del livello dell’*ethos*, del *sentire morale* di una società: qui rimangono validi gli avvertimenti aristotelici rispetto alla necessità di inquadrare la moralità delle decisioni riferendole ai loro fini e considerando la bontà di questi ultimi tanto rispetto ai singoli quanto rispetto alla *polis*. A questo livello si riconosce che le decisioni hanno un impatto sul legame sociale e che la cura per l’altro – la *sollecitudine* nel lessico ricoeuriano – è un elemento costitutivo e non meramente contrattualista della dimensione relazionale (ivi, trad. it. p. 36).

A questo primo livello di attenzione si possono ricondurre alcune considerazioni relative al ricorso all’IA nei processi decisionali complessi: la *finalità umanistica*, che per lo più rimane sottintesa, merita di essere talvolta richiamata esplicitamente, in particolare lì dove gli applicativi paiono essere più fortemente improntati a finalità commerciali di massimizzazione del profitto, anche a discapito del consumatore (CAP. CODICI ETICI pp….). Si possono ricondurre a questo livello le problematiche relative alla profilazione degli utenti in funzione di campagne promozionali mirate e aggressive, specie attraverso i *social media*.

* 1. L’attenzione alla sostenibilità universale dell’agire

Il secondo livello di attenzione riguarda il problema della verifica della *sostenibilità* morale del proprio agire. In questo caso Ricoeur invita a fare riferimento alla prospettiva kantiana (ivi, trad. it. p. 39) e quindi al momento “normativo”, che propone di qualificare tecnicamente come “morale” (riservando il termine “etica” alla riflessione sui fondamenti e sul senso dell’agire): l’elaborazione di una *normatività* (più spesso nella forma dell’interdetto) risponde all’esigenza di contrastare il male che può derivare da una eccessiva discrezionalità nelle decisioni, tale da sfociare nell’arbitrarietà. La proposta di Kant, in particolare con il primo imperativo categorico richiamato sopra, può valere come una sorta di *crash-test* per le proprie decisioni: il fatto che possano superare la *prova della universalizzazione* rappresenta appunto una garanzia di non-arbitrarietà nell’affermazione del bene e nel riconoscimento e rifiuto del male.

A questo secondo livello di attenzione è possibile ricondurre, tra quelle segnalate sopra, le problematiche relative alla trasparenza degli algoritmi e dell’ampia disponibilità di dati anche sensibili (CAP MEDICINA pp…., CAP CYBERSICUREZZA pp….): l’esigenza di controllo del procedimento dal punto di vista della universalità è infatti associabile all’aspettativa per cui un processo decisionale debba evitare l’adozione di comportamenti che si configurino come privilegi per alcuni o come vantaggi informativi, senza una adeguata considerazione dei costi che potrebbero comportare per altri. Considerare la sostenibilità di un comportamento, nell’ipotesi astratta che venisse adottato da tutti, risulta essere un modo efficace per sottoporre a critica le prospettive d’azione. Il “test” kantiano è peraltro particolarmente interessante perché conduce idealmente ad una “autonomia morale” (Ricouer, 1990, trad. it. p. 41): anziché esplicitare una serie di norme volte al rispetto dell’umanità si crea una sorta di meta-regola di controllo, che appunto consiste nella “prova” dell’universalizzazione. È cioè la singola persona a stabilire il contenuto specifico del da farsi e quindi a darsi la regola concreta secondo cui agire, venendo però invitata a chiedersi se il modo di procedere individuato possa essere ancora accolto nel momento in cui fosse adottato da tutti. In questo senso la prospettiva kantiana evidenza da un lato la necessità morale di una intelligibilità dei processi decisionali e dall’altro quella di ponderarli attraverso un “test di sostenibilità” che ne consenta una revisione critica intersoggettiva.

* 1. L’attenzione alla singolarità degli attori e delle situazioni

Il terzo livello di attenzione suggerito dalla lezione di Ricoeur procede dal rilevamento dei limiti del secondo: l’elaborazione di norme a validità universale mette al riparo dai rischi di arbitrarietà nel giudizio morale, ma lì dove le massime si trasformano in sistemi di regole più definite – ad esempio nella forma dei codici deontologici – quel che può insorgere è un *conflitto di doveri* (ivi, trad. it. p. 46). La medesima azione può, in diversi casi della vita reale, essere prescritta da una norma ed essere proibita da un’altra, avendo entrambe – in astratto – pari cogenza morale. Un caso potrebbe essere quello, in campo medico, della comunicazione al paziente di una diagnosi infausta: da un lato la norma che prescrive la necessità di dire il vero (nella formula dell’interdetto “non mentire”), dall’altra la norma che prescrive di perseguire il miglior interesse del paziente, indicazione entro cui può essere inclusa anche una riflessione sui modi e tempi della comunicazione, anche in relazione alla storia e alle attitudini della persona, per metterla nelle migliori condizioni soggettive per reagire alla fatica che la attende. Il conflitto non risulta di per sé risolvibile se non con una violazione. Ma è ammissibile moralmente? Qui Ricoeur, proponendo un ulteriore passaggio verso la dimensione delle “etiche applicate”, recupera il riferimento alla tradizione medievale e in particolare agli sviluppi di quegli aspetti della prospettiva aristotelica, che già aveva riservato particolare attenzione alla considerazione della particolarità delle situazioni. Il riferimento teorico-concettuale è qui alla virtù della *prudentia* e alla dimensione del *consilium* elaborata in particolare da Tommaso d’Aquino (trad. it. p. 158), sviluppando le riflessioni aristoteliche sulla *phronesis*.

Secondo la sua più specifica lezione, il *consiliari*, il consultarsi con altri, è quella sezione del processo decisionale in cui vengono raccolti più sguardi competenti sulla specifica situazione (ivi,trad. it. p. 540). È una fase collegiale, che non si chiude di per sé con la decisione: questa infatti spetta a chi ha il potere di agire, che sarà titolare del giudizio (*iudicare*) e dunque della scelta. Secondo Ricoeur la “cellula del consiglio”, l’esistenza di un luogo in cui «più punti di vista sono a confronto, nell’amicizia e nel rispetto reciproco» (Ricoeur 1990, trad. it. p. 49) è ciò che consente di far fronte al *tragico dell’azione*, ovvero all’insorgenza di conflitti di doveri che non appaiono concretamente superabili se non contravvenendo a qualche aspetto delle norme stesse. Si tratta, qui, della previsione della morale tommasiana che include nell’*agire prudente* la possibilità della *trasgressione della norma*, trasgressione tuttavia praticabile solo nel riferimento al più generale sentire (l’*ethos* condiviso) che Ricoeur stesso pone al primo livello di attenzione. Qui si comprende come l’attenzione sia dunque rivolta al far fronte al male che può derivare dall’applicazione rigida delle stesse norme morali che hanno superato astrattamente la prova dell’universalizzazione.

A questo terzo livello possono essere ricondotte le problematiche relative ai processi di classificazione implicati nel ricorso all’apprendimento automatico dell’IA: la generazione di classi e la riconduzione delle singolarità a profili rispondenti a stereotipi può in diversi casi condurre ad output discriminatori, ingiusti o comunque nocivi (per approfondimenti, si veda il capitolo 5). Si ritrovano qui tutti i casi in cui l’automatizzazione elide nel processo decisionale il passaggio sospensivo del consiglio, a valle del quale è contemplata – per fondate ragioni – anche la possibilità di una decisione trasgressiva. Confluiscono a questo livello di problematicità alcune delle situazioni richiamate in precedenza: l’applicazione immediata di un premio assicurativo in base alle previsioni statistiche di frequenza di incidenti, il licenziamento automatico degli insegnanti sulla base di classificatori di rendimento ma anche la comunicazione di una diagnosi affidata a un responso automatico o – caso che finora non si dà, proprio in applicazione del ricorso alla cellula del consiglio – l’intrapresa di una terapia unicamente a fronte di una diagnostica eseguita dalle macchine.

1. IA e istanze morali

Indubbiamente le macchine offrono possibilità crescenti di elaborazione di dati stratificati e complessi, e il sapere che ne può derivare costituisce una risorsa preziosa per compiere delle scelte che siano migliorative del vivere, nel rispetto dell’umano e dell’ambiente. Comprendere il funzionamento degli algoritmi e delle tecniche di progettazione è importante per evitare tanto timori quanto entusiasmi acritici, e allo stesso tempo risulta necessario prendere coscienza della complessità antropologica e morale di ogni processo decisionale. In questo senso la conoscenza tecnica va costantemente integrata con la consapevolezza della crucialità dei tre livelli di attenzione attivabili nei processi decisionali suggeriti dalla sintesi di Ricoeur, che vanno evidentemente integrati tra loro. Senza mantenere attiva una riflessione di ordine generale (il livello dell’*ethos*) improntata alla prospettiva della *sollecitudine* – anche secondo la celebre definizione dell’etica come «auspicio della vita buona, con e per gli altri, all’interno di istituzioni giuste» (Ricouer 1990, trad. it. p. 34) – si correrebbe il rischio di uno sviluppo di attenzioni funzionali importanti, come ad esempio la ricerca della trasparenza, ma che tuttavia, se prive di un ancoraggio al senso dell’umano, lascerebbero i processi più esposti all’infiltrazione del male. Senza una riflessione di tipo normativo (il livello della morale) si correrebbe il rischio dell’arbitrarietà e dell’opacità, precludendosi la possibilità di intercettare snodi in cui l’anti-umano potrebbe prodursi a causa di una non-considerazione dell’universalità. Senza una dimensione di tipo *phronetico* o *prudenziale*, senza cioè l’inserzione di un passaggio tra la raccolta e selezione delle informazioni (*consiliari*) e l’elaborazione di una risoluzione (*iudicare*), senza la possibilità di sostare tra gli output computazionali e la risoluzione attuativa, la decisione in senso stretto, si perderebbe la possibilità di riservare alle situazioni di vita, di per sé tutte irripetibili, quella cura locale che rimane sempre un passo oltre l’applicazione meccanica delle stesse norme morali.

Questi tre livelli di attenzione, tra loro integrati, recepiscono gli apporti delle tradizioni di pensiero morale più consolidate e consentono di impiegarli nell’esame critico delle risorse dell’IA e del ricorso alle macchine nei processi decisionali. Si tratta di attenzioni da tenere presenti quantomeno per accrescere il livello di vigilanza rispetto alle diverse forme di incompletezza dei processi stessi e – più in generale – per mantenere viva la sensibilità rispetto al tema del male e del suo riproporsi nella storia umana attraverso le “cattive” decisioni.

Capitolo 3

Medicina digitale e IA: profili etici

Alessandro Blasimme

1. Introduzione

La medicina, al pari di molti altri ambiti dell’attività umana, va incontro a profondi cambiamenti innescati dalla sempre maggiore diffusione delle tecnologie digitali. In particolare, la crescente disponibilità di dati relativi agli aspetti tanto biologici quanto sociali e comportamentali della salute e della malattia dà luogo a profonde trasformazioni che si ripercuotono direttamente sulla professione medica e sulle pratiche di cura.

Il capitolo si propone di fornire un quadro d’insieme delle applicazioni relative alla tecnologia digitale e all’intelligenza artificiale in ambito medico e di illustrare le questioni etiche che tali applicazioni sollevano.

Si porrà l’accento sui problemi legati alla crescente circolazione di dati sanitari personali generati tanto in ambito clinico che attraverso dispositivi mobili personali. Si affronteranno le questioni relative all’automazione sia dei processi di presa in carico, sia delle decisioni cliniche di natura diagnostica, prognostica e terapeutica. Si affronteranno le implicazioni della medicina digitale nell’ambito della relazione medico-paziente. Infine, si discuterà della medicina digitale non solo come di un avanzamento tecnico, ma anche come di un modo in parte inedito di concepire la salute umana e le responsabilità individuali, professionali e collettive che la sua difesa comporta.

Negli ultimi venti anni, la costante espansione delle infrastrutture di telecomunicazione ha creato le condizioni per una diffusione capillare delle tecnologie digitali. La crescita esponenziale della rete è stata accompagnata da trasformazioni profonde. Le attività umane che hanno saputo adattarsi a questo cambiamento assecondando o anticipando i processi di digitalizzazione hanno tratto enormi benefici dall’innovazione digitale. Tutte le altre però sono state condannate ad un rapido e irreversibile declino. Questo aspetto di distruzione creativa, per usare l’espressione del celebre economista liberale Joseph Schumpeter, è un tratto distintivo dell’innovazione nei sistemi capitalistici (Schumpeter, 1964). Tuttavia, con il trionfo dell’economia digitale, esso ha assunto l’aspetto di un processo inesorabile e inarrestabile.

Una conseguenza inattesa della centralità assunta dalle transazioni digitali (economiche, commerciali, sociali) è stata rappresentata dal crescente interesse suscitato da quelle che potremmo chiamare le tracce digitali di tali transazioni. Un numero sempre crescente di attori appartenenti ai settori più disparati ha compreso che i dati che descrivono i nostri comportamenti e le nostre interazioni *online*, così come i dati che permettono di ricostruire tutti i processi che hanno luogo per mezzo di infrastrutture digitali hanno un enorme potere predittivo. Vista la diffusione delle tecnologie digitali connesse alla rete e il loro utilizzo massivo, il volume di questi dati ha ormai toccato livelli esorbitanti. In molti utilizzano a questo proposito l’espressione *big data* (Mayer-Schönberger, Kenneth Cukier, 2013). Giungiamo così ad un secondo tratto distintivo della cosiddetta rivoluzione digitale, ossia la necessità di strumenti in grado di analizzare i Big Data in modo da poter informare i processi decisionali e orientare le scelte strategiche di un numero sempre crescente di attori economici, politici e professionali. La centralità assunta, anche nel dibattito pubblico, dai temi connessi all’intelligenza artificiale è legata proprio al suo uso in riferimento ai Big Data, i quali rappresentano per così dire la materia prima su cui gli algoritmi di intelligenza artificiale vengono sviluppati.

Secondo Klaus Schwab, i cambiamenti a cui abbiamo rapidamente accennato costituiscono lo sfondo della quarta rivoluzione industriale (Schwab, 2016). Quest’ultima non si limiterebbe alla crescente diffusione di macchine e dispositivi connessi. Essa infatti si caratterizza piuttosto per il modo in cui la rete, le macchine e gli algoritmi favoriscono una costante integrazione o fusione del modo fisico, del mondo digitale e dell’ambito biologico – quest’ultimo reso sempre più adatto all’elaborazione algoritmica da un lato grazie ai *big data* biomedici, quali ad esempio i dati genomici, e dall’altro grazie ai processi di fenotipizzazione (vedi OLTRE pp,….) resi possibili dai dispositivi digitali personali come gli *smartphone*, come vedremo più avanti.

Quello che viene evocato è dunque uno scenario di ulteriore radicale trasformazione dei processi produttivi, del modo di acquistare e di usufruire di beni e servizi, del modo di lavorare, di muoversi, di apprendere, di comunicare, e di guardare alla salute umana.

È in questo contesto dunque che deve essere collocato il dibattito sulla medicina digitale e sull’impiego dell’intelligenza artificiale nel settore della salute umana. In quello che segue non si pretenderà di fornire una panoramica esaustiva dei molteplici ambiti di applicazione della cosiddetta medicina digitale o dell’utilizzo in campo medico e sanitario dell’intelligenza artificiale. Forniremo piuttosto una panoramica generale che terrà conto dei più recenti sviluppi tecnologici ma che, per forza di cose, non potrà che suggerire le linee di sviluppo futuro in un settore che è in continua espansione. Ci occuperemo quindi di ricostruire il dibattito – in ambito sia internazionale che italiano – sulle questioni etiche legate alla medicina digitale prendendo le mosse dalle linee guida redatte da alcune importanti organizzazioni internazionali. Successivamente ci soffermeremo sulle questioni legate alla trasparenza degli algoritmi di intelligenza artificiale, alla protezione dei dati personali e alla regolazione dell’innovazione digitale in medicina. Infine ci occuperemo delle implicazioni che la medicina digitale e l’automazione comportano in merito all’idea di salute e alla relazione medico e paziente.

1. Concetti fondamentali e casi studio

In questa sezione del capitolo si intende fornire al lettore una panoramica dei più recenti sviluppi nell’ambito della medicina digitale. Prima di iniziare questo excursus converrà tuttavia soffermarci su alcuni aspetti relativi alla definizione stessa di medicina digitale. Questa espressione viene perlopiù utilizzata in modo generico, in quanto non esiste una definizione universalmente accettata di cosa si debba intendere per medicina digitale. Secondo la *Food and Drug Administration* (FDA) negli Stati Uniti, l’ambito della *digital health* (salute digitale) racchiude numerosi paradigmi sanitari, ovvero: le applicazioni legate a dispositivi mobili (*mobile health*, o *mHealth*), l’informatica medica, i dispositivi indossabili, la telemedicina e la medicina personalizzata (Vayena *et al.*, 2018). Ciò che accomuna questa variegata famiglia di modelli è la centralità attribuita al flusso circolare di dati e informazioni dal paziente ai dispositivi, ai professionisti della salute, ai sistemi sanitari e poi di nuovo ai pazienti (Vayena *et al*., 2018).

Altri hanno proposto di distinguere tra *digital health* e *digital medicine* (medicina digitale, termine introdotto già nel 2002 (Williamson Shaffer *et al*., 2002): la prima si riferirebbe alle applicazioni legate ai dispositivi personali (quali *smartphone* e altri oggetti connessi) e concepite principalmente per il mercato dei consumatori; mentre la seconda riguarderebbe più propriamente tecnologie e prodotti digitali che vengono sottoposti a rigorosi processi di validazione clinica e che dunque possono avere un impatto diretto sulla pratica clinica in termini di diagnosi, prevenzione, monitoraggio e trattamento delle malattie (Elenko, Underwood, Zohar, 2015). Questa definizione ha il merito di mettere in evidenza in maniera chiara alcuni importanti ambiti di applicazione delle tecnologie digitali nel contesto della salute umana. Possiamo infatti distinguere almeno quattro macro-aree di applicazione, ovvero: il monitoraggio della salute; la diagnosi della malattia; l’uso di algoritmi predittivi applicati sia alla prognosi che al trattamento; le applicazioni legate alla salute pubblica.

Molti di noi hanno familiarità con l’utilizzo di applicazioni per il tracciamento dell’attività sportiva. Attraverso tali applicazioni è possibile monitorare le proprie sessioni di allenamento e registrare alcuni parametri vitali come il battito cardiaco. La possibilità di acquisire dati relativi all’attività quotidiana e di monitorare in tempo reale lo stato di salute delle persone ha dato luogo ad una linea di ricerca incentrata sul cosiddetto fenotipo digitale. Con questa espressione deve intendersi la quantificazione continua del fenotipo individuale attraverso dati raccolti da dispositivi digitali e in particolare da *smartphone* (Jain *et al*., 2015; Huckvale, Venkatesh, and Christensen, 2019). L’acquisizione passiva di dati fenotipici attraverso accelerometri, sensori di geo-localizzazione (GPS) e l’analisi dell’interazione tra dispositivo e utente può dimostrarsi utile nel monitoraggio di condizioni cliniche come la depressione e l’ansia o per predire le ricadute nei pazienti affetti da disturbo bipolare (Huckvale, Venkatesh, and Christensen, 2019).

Approcci simili sono stati utilizzati per definire i cosiddetti biomarcatori digitali. In questo campo, algoritmi di intelligenza artificiale basati sull’apprendimento automatico (*machine learning*) sono in grado ad esempio di diagnosticare problemi legati al declino delle funzioni cognitive dovuti ai prodromi della demenza(Dagum, 2018). Il monitoraggio dei biomarcatori digitali si riferisce soprattutto alla misurazione di alcuni motivi ricorrenti (*pattern*) nell’interazione tra un utente e il suo *smartphone* (ad esempio la velocità di scrittura, gli errori commessi, i tempi di reazione a stimoli cognitivi controllati e via dicendo).

I dispositivi connessi di cui stiamo parlando possono essere portatili o di tipo residenziale. In questo la convergenza tra domotica, intelligenza artificiale e Internet delle cose consente di considerare le abitazioni come stazioni di monitoraggio della salute (*smart home*) soprattutto nel caso degli anziani (Liu *et al*., 2016). Alle funzioni di monitoraggio si aggiungono quelle di stimolo, volte ad incrementare l’attività fisica (attraverso semplici esercizi da eseguire con l’ausilio di app e allenatori virtuali) o l’interazione sociale (attraverso proposte culturali o piattaforme di comunicazione).

Lo sviluppo di tutte queste forme di monitoraggio digitale richiede ancora notevoli investimenti sul fronte della ricerca. In questo senso, gli sforzi degli scienziati si concentrano tanto sullo sviluppo di piattaforme tecnologiche apposite, quanto sull’utilizzo di dispositivi più ordinari – come testimoniato dall’accordo tra Fitbit (la casa che produce la popolare linea di bracciali per l’allenamento) e i *National Institutes of Health* (NIH) nell’ambito dell’iniziativa nazionale statunitense sulla medicina di precisione. Lo scopo di queste linee di ricerca è di analizzare i dati fenotipici digitali insieme ai dati genomici e a quelli relativi a determinanti ambientali e socio-economici della salute, al fine di adattare sempre più le pratiche diagnostiche e i trattamenti alle specifiche caratteristiche di ogni singolo paziente.

Uno degli aspetti più promettenti della medicina digitale è rappresentato dall’uso dell’intelligenza artificiale in ambito diagnostico. In particolare, gli algoritmi di apprendimento automatico si rivelano sorprendentemente efficaci nell’analisi di immagini cliniche e radiologiche (Hosny, 2012). Citeremo due casi a titolo di esempio. In uno studio condotto nel 2016, è stato sviluppato un algoritmo basato sull’apprendimento profondo (*deep learning*) in grado di diagnosticare il diabete a partire da immagini del fondo oculare dei pazienti (Gulshan *et al*., 2016).

In un altro studio del 2017, un gruppo di ricercatori dell’università di Stanford ha messo a punto un algoritmo di apprendimento automatico in grado di riconoscere la presenza di carcinomi e melanomi in un’immagine clinica – con un’accuratezza paragonabile o superiore a quella di dermatologi in carne e ossa (Esteva *et al*., 2017).

Questi sviluppi evocano la possibilità di automatizzare in maniera tecnicamente affidabile i processi diagnostici, aumentandone l’accuratezza, riducendo il carico di lavoro per gli specialisti e potenzialmente allargando la platea delle persone raggiungibili dalle campagne di prevenzione.

Gli algoritmi predittivi basati sull’intelligenza artificiale possono rappresentare per la medicina strumenti particolarmente interessanti. Gli algoritmi che analizzano la storia clinica di un paziente e i dati relativi ad una specifica condizione patologica – come ad esempio il cancro – possono elaborare previsioni attendibili in merito al decorso clinico della malattia. Ad esempio, un lavoro del 2018 descrive lo sviluppo di un algoritmo in grado di predire il rischio di mortalità a breve termine (trenta giorni) per i pazienti oncologici che inizino un nuovo ciclo di chemioterapia (Elfiky *et al*., 2018). Un approccio analogo è stato utilizzato nel 2020 per stimare il rischio di mortalità a breve termine di pazienti affetti da COVID-19 attraverso algoritmi capaci di analizzare in parallelo una molteplicità di parametri clinici (Zhu *et al*., 2020), o di sintomi associati all’infezione da SARS-CoV-2 (Pourhomayoun, Shakibi, 2021).

La pandemia causata da questo virus ha inoltre spinto numerosi governi a sviluppare strumenti di medicina digitale applicati alla dimensione della salute pubblica. Visto il ruolo svolto dagli asintomatici nella diffusione dell’epidemia, e al fine di potenziare il tracciamento dei contatti dei casi positivi sono state messe a punto app di tracciamento basate sulla tecnologia GPS o sul segnale Bluetooth presenti negli *smartphone*.

1. Profili etici

Negli ultimi anni, il dibattito sugli aspetti etici della medicina digitale e dell’impiego dell’intelligenza artificiale in medicina è stato molto acceso. Tuttavia bisogna rilevare che tale dibattito e le posizioni che al suo interno si confrontano sono difficilmente riconducibili a posizioni normative definite sul piano teorico e filosofico. Infatti, a differenza di quella che potremmo chiamare la tradizione normativa dell’etica medica, dell’etica della ricerca clinica e della bioetica in generale, tale dibattito non si svolge all’interno di un quadro formato da teorie morali o assunti filosofici chiaramente articolati. La mancanza di questo sfondo normativo si avverte nella genericità e ridondanza del dibattito, nonché nella sua incapacità di fornire orientamenti pratici agli attori in campo. In altri termini, sebbene il discorso etico sulla medicina digitale e sull’intelligenza artificiale abbia assunto proporzioni considerevoli, esso si limita per lo più ad una mappatura delle questioni in campo, mentre appare deficitario sul piano dell’articolazione normativa vera e propria.

I temi più ampiamente dibattuti sono tre: (1) le questioni inerenti la privacy e la protezione dei dati personali; (2) le questioni relative all’apprendimento automatico; e infine (3) le questioni legate alla trasparenza degli algoritmi di intelligenza artificiale.

Esistono poi discussioni di natura più propriamente sociologica relative ai modelli di cura incarnati dalla medicina digitale e sulle quali ritorneremo brevemente nella sezione 4.

(1) La medicina digitale si basa sulla possibilità di raccogliere, conservare, analizzare e distribuire una grande quantità di dati personali, sanitari e non. Questo pone non poche difficoltà. Innanzi tutto le persone i cui dati vengono raccolti e utilizzati devono essere al corrente di tali pratiche. In caso contrario si configura una violazione dei principi di autonomia e auto-determinazione che sono legati, nella tradizione morale del pensiero liberale, alla capacità di esercitare un certo grado di controllo su alcuni aspetti inviolabili della sfera privata, tra cui appunto il controllo relativo alle informazioni che ci riguardano in prima persona (Allen, 1998). Tali ideali morali si fondano invero sulla distinzione tra sfera privata e sfera pubblica – distinzione che si può far risalire già alla cultura politica dell’antica Grecia, nella quale si distingueva appunto tra *polis* e *oikos*. A questa concezione corrisponde l’idea che le decisioni relative all’accesso alle informazioni personali spettino esclusivamente alla persona alla quale queste informazioni si riferiscono, e non perché queste le appartengano, ma perché la riguardano. A questo diritto corrisponde un duplice obbligo da parte dell’autorità pubblica. Da un lato essa non può scavalcare le prerogative individuali in merito all’accesso ai dati perché questo violerebbe la privacy delle persone. In tal senso la privacy si configura come diritto negativo assimilabile ai diritti che proteggono ciascuno dall’interferenza delle autorità con scelte, decisioni e comportamenti che debbono considerarsi privati. Dall’altro, le autorità pubbliche hanno obblighi di giustizia derivanti dal diritto alla privacy. Esse devono cioè proteggere le persone da violazioni della privacy derivanti dall’accesso non autorizzato a dati e informazioni personali. Queste violazioni possono esporre le persone al rischio di essere ricattate, discriminate o truffate sulla base di informazioni personali acquisite illegittimamente. Ma le violazioni della privacy in merito ai dati personali sono da considerarsi moralmente rilevanti in quanto tali, ovvero indipendentemente dalle loro conseguenze, in quanto configurano un’invasione illegittima della sfera privata. Per questo motivo esistono condizioni molto specifiche – come il consenso informato – per l’accesso ai dati personali e per la corretta gestione tecnica dei dati raccolti.

Quanto abbiamo detto finora traduce dunque dei principi morali diffusamente accettati (almeno nel mondo occidentale) in un complesso sistema di norme e garanzie legali, istituzionali e tecniche. Il Regolamento Generale sulla Protezione dei Dati (RGPD) definisce dal 2018 il quadro giuridico di riferimento in materia per l’Unione Europea. In esso si introduce una serie di diritti specifici resi necessari dalla pervasività delle pratiche di raccolta di dati personali attraverso dispositivi digitali. Essi comprendono il diritto da parte di ciascuno di conoscere quali suoi dati personali una data organizzazione ha raccolto, per quali motivi, come li ha ottenuti e con chi li condivide. Ogni cittadino europeo può inoltre richiedere la cancellazione o la rettifica dei propri dati personali da parte delle organizzazioni che li hanno raccolti, e può obiettare all’utilizzo dei propri dati stanti specifiche condizioni. Un diritto particolarmente interessante contenuto nel Regolamento europeo riguarda il diritto di opporsi a decisioni risultanti da un’analisi interamente automatizzata dei propri dati (art. 22). In pratica, i cittadini europei possono contestare decisioni prese in modo esclusivo da algoritmi deterministici o che impieghino intelligenza artificiale, come ad esempio decisioni relative alla concessione di un prestito da parte di una banca. Come appare evidente, questo diritto si estenderebbe all’ambito della medicina digitale nel caso ad esempio di valutazioni diagnostiche completamente automatizzate. L’utilizzo di algoritmi per guidare le decisioni relative al trattamento potrebbe senz’altro rientrare in questa fattispecie, qualora il livello di automazione fosse tale da non prevedere il coinvolgimento di un medico. In questi casi, in base all’articolo 22 del RGPD, il paziente potrebbe ad esempio richiedere che un medico in carne ed ossa fosse coinvolto nel processo. Torneremo in seguito su questo argomento, nella sezione dedicata alla trasparenza degli algoritmi basati sull’apprendimento automatico.

Come abbiamo già visto, la medicina digitale si basa su fonti di dati molto eterogenee e si appoggia a piattaforme tecnologiche multiple, alcune delle quali non sono concepite secondo gli stringenti parametri etici e giuridici che caratterizzano l’ambito sanitario. Questo rende molto complesso comprendere quali aspettative le persone coinvolte possano nutrire nei confronti della propria privacy e degli obblighi di confidenzialità a cui si devono conformare le diverse figure professionali coinvolte nello sviluppo e nella messa in opera della medicina digitale. Gli standard relativi alla raccolta dei dati sanitari sono generalmente molto più elevati rispetto a quelli utilizzati ad esempio dalle piattaforme online. Mentre in ambito sanitario è generalmente richiesto il consenso informato, per i dati raccolti tramite app e dispositivi personali ci si limita in genere ad una sommaria approvazione dei termini del servizio.

 (2) Gli algoritmi di apprendimento automatico vengono sviluppati facendo, per così dire, allenare l’algoritmo su una grande quantità di dati annotati o meno. In altre parole, l’algoritmo apprende regolarità e correlazioni statisticamente rilevanti nei parametri che gli vengono forniti come esempio (vedi cap. 1). Quindi, per sviluppare un algoritmo in grado di diagnosticare, poniamo, un melanoma a partire da un’immagine clinica, i ricercatori forniscono degli esempi di immagini contenenti effettivamente un melanoma (più altre che non ne contengono) e lasciano che l’algoritmo apprenda da solo quali parametri grafici (colori, forme, contrasto, luminosità dei singoli pixel e via dicendo) correlano con una diagnosi positiva e quali no. In questo modo l’algoritmo ha appreso delle regole che applicherà ai nuovi casi clinici che gli verranno sottoposti e che lo guideranno in seguito nella diagnosi di nuovi pazienti. La qualità dei dati di apprendimento è fondamentale per assicurare l’accuratezza dell’algoritmo, ovviamente. Se i dati di apprendimento non sono sufficientemente rappresentativi della popolazione in cui l’algoritmo verrà impiegato, esso non sarà egualmente efficace per ogni tipo di paziente (Obermayer, Emanuel, 2016). È noto tuttavia che i dati disponibili nell’ambito della ricerca biomedica sono in grande prevalenza rappresentativi di adulti bianchi di origine europea o nord-americana (Redwood, Gill, 2013). Nel solo ambito della ricerca oncologica ad esempio, né i modelli preclinici (come le linee cellulari), né i campioni conservati nelle biobanche, né i dati genomici riflettono adeguatamente le reali proporzioni tra i diversi gruppi etnici nella popolazione umana (Guerrero *et al*., 2018). In assenza di un adeguato sistema di controllo che permetta di verificare la qualità dei dati di apprendimento, l’uso di algoritmi finirebbe dunque per penalizzare i gruppi meno rappresentati e acuire così le disparità che già li colpiscono. Per quanto riguarda i dati clinici poi, bisogna ricordare che alcune malattie sono sovradiagnosticate in alcuni gruppi etnici, come ad esempio la schizofrenia nei cittadini statunitensi di colore (Vayena, Blasimme, Cohen, 2018). Un algoritmo che venisse allenato su dati che riflettono questa tendenza tenderebbe a sua volta a riprodurla e in un certo senso a renderla automatica. È chiaro che questo tipo di problema configura il rischio di concrete violazioni del principio di giustizia e che tali violazioni sarebbero inaccettabili sul piano morale (Beam, Kohane, 2018; Beam, Kohane, 2016; Grote, Berens, 2020).

 (3) L’efficacia degli algoritmi di apprendimento automatico dipende dalla potenza di calcolo che consente di testare migliaia di parametri fino a scoprire quelli statisticamente significativi dal punto di vista clinico. La macchina apprende da sola quali parametri biologici, fenotipici o clinici correlano con una determinata condizione patologica all’interno dei dati che le sono forniti. Ciò che la macchina apprende non è altro che un insieme di correlazioni statistiche tra un certo numero (anche molto elevato) di variabili e una classificazione clinica (ad esempio la diagnosi di una malattia) o una previsione (ad esempio il modo in cui un paziente reagirà ad un dato farmaco). I dati su cui la macchina apprende possono essere tanto estesi (*complex high-dimensional data*) da non poter essere modellati seguendo approcci ipotetico-deduttivi convenzionali. Ad esempio, sarebbe impossibile formulare un’ipotesi circa il ruolo che l’espressione di un numero molto elevato di geni svolge in merito alla probabilità che un farmaco sia efficace. Semplicemente non disponiamo di un modello che definisca la relazione tra quel complesso di dati epigenetici e l’esito clinico atteso. Ma un algoritmo di apprendimento automatico può ovviare alla mancanza di un’ipotesi (o modello) di partenza e apprendere direttamente dai dati quali correlazioni esistono tra le numerosissime variabili in campo. In altri termini, per così dire, l’algoritmo non sa perché i parametri che ha imparato ad osservare siano rilevanti, ovvero, nel nostro esempio, non fornisce una spiegazione causale del perché a determinati profili epigenetici corrispondano determinati esiti clinici. Ovviamente, neanche il programmatore può fornire una spiegazione causale al riguardo. Anzi chi sviluppa questi algoritmi in genere non ha neanche modo di scoprire quali regole la macchina abbia appreso, ovvero su quali caratteristiche del fenomeno osservato essa si basi per formulare una previsione o una classificazione. Per questo motivo, in un senso metaforico, ci si riferisce spesso agli algoritmi di apprendimento automatico come algoritmi opachi, non trasparenti o come a delle scatole nere (*black-box algorithms*) (Ferretti, Schneider, Blasimme, 2018; Amman *et al*., 2020). Un algoritmo non trasparente può essere molto accurato ma rimane indifferente rispetto alla struttura causale del fenomeno a cui si applica. Per restare all’esempio illustrato poco sopra, il fatto che l’algoritmo possa risolvere la complessità dei dati e pervenire ad una decisione clinica, non dipende dal conoscere quale siano i meccanismi biologici che legano determinati profili di espressione genica ad una maggiore (o minore) efficacia di un farmaco. Bisogna tuttavia sottolineare che il comportamento di un algoritmo può essere analizzato retrospettivamente e indirizzare la formulazione di ipotesi di spiegazione causale che potranno successivamente essere verificate (o falsificate) sperimentalmente in laboratorio. Questi aspetti suscitano questioni epistemologiche molto complesse delle quali non possiamo dare conto in questo capitolo. Tali questioni sono legate all’enfasi che la medicina basata sull’intelligenza artificiale pone sulla previsione e sulla classificazione, piuttosto che sulla spiegazione di un dato fenomeno clinico (Wilkinson *et al*., 2020). Esse hanno però anche implicazioni pratiche ed etiche sulle quali occorre soffermarci.

Un medico che utilizzi un algoritmo non trasparente si affida in sostanza ad uno strumento di cui non conosce il funzionamento, o meglio, uno strumento in merito al cui funzionamento non è possibile fornire alcuna spiegazione di tipo causale. Ciò sembrerebbe minare alle fondamenta il dovere di ogni medico di agire, come si suol dire, ‘in scienza e coscienza’ – ovvero sapendo ciò che fa e avendo buone ragioni per farlo. Si impone dunque una riflessione sulla conciliabilità delle pratiche mediche basate su algoritmi *black-box* e l’*ethos* della medicina basato su norme di condotta che incarnano anche degli ideali epistemologici ben precisi. Sempre sul piano etico, si devono considerare altre due questioni. Innanzi tutto, bisogna stabilire se il medico abbia degli obblighi di trasparenza nei confronti del paziente in merito alle caratteristiche tecniche degli algoritmi che utilizza. In altre parole, il medico dovrebbe informare il paziente del fatto che si avvale di un algoritmo il cui funzionamento non può essere spiegato? In secondo luogo ci si deve interrogare su come l’opacità degli algoritmi di apprendimento automatico si ripercuota sulla responsabilità legale in ambito sanitario – ovvero se l’essersi affidato ad un tale algoritmo costituisca un’aggravante nei casi di danni ai pazienti o di negligenza medica. Il dibattito accademico su tali importanti argomenti è molto vivace. Tuttavia, nonostante il richiamo del RGPD in merito al diritto di opporsi a decisioni interamente automatizzate, tale dibattito non ha ancora prodotto raccomandazioni concrete in merito all’etica degli algoritmi cosiddetti black-box in ambito medico (Shmueli, 2010; Wilkinson *et al*., 2020).

Tuttavia i problemi di cui abbiamo discusso finora hanno dato vita ad una grande quantità di pubblicazioni da parte organismi internazionali, società mediche e gruppi di esperti in materia. Tali pubblicazioni contengono anche raccomandazioni e principi di ordine però piuttosto astratto e generale (per approfondimenti, si veda il capitolo 13).

Una parte considerevole di queste linee guida riguarda l’uso dell’intelligenza artificiale in senso lato. Tra queste vale la pena ricordare i principi sull’intelligenza artificiale pubblicati dall’Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) nel 2019. Tali principi definiscono il quadro etico di riferimento richiamando l’attenzione sul valore dell’inclusività, della giustizia, della centralità dell’elemento umano, della trasparenza (nel senso discusso nella sezione precedente), della qualità e della responsabilità[[23]](#footnote-23).

A febbraio del 2020, la Commissione Europea ha pubblicato un libro bianco sull’intelligenza artificiale nel quale si mettono in evidenza i principi che devono guidare l’approccio europeo all’innovazione in questo settore. In particolare, si pone l’accento sulla necessità di attribuire sempre ad operatori umani compiti di intervento e sorveglianza nelle decisioni guidate da algoritmi di intelligenza artificiale; di assicurare robustezza tecnica e sicurezza; di garantire la riservatezza e un livello adeguato di protezione dei dati personali; di affrontare sul piano tecnico le questioni legate alla trasparenza; di verificare la diversità dei dati utilizzati nell’apprendimento automatico in modo da garantire equità e assenza di ingiuste discriminazioni; di tenere in considerazione il benessere sociale e ambientale come obiettivi primari dell’innovazione; e infine di chiarire ruoli e responsabilità per tutti i principali attori coinvolti nello sviluppo e nell’applicazione degli algoritmi di intelligenza artificiale[[24]](#footnote-24). Sebbene si tratti di raccomandazioni di ordine generale, in larga misura, questi principi trovano applicazione anche in ambito sanitario.

Per quanto riguarda specifiche applicazioni mediche dell’intelligenza artificiale, bisogna ricordare alcune prese di posizione da parte di importanti agenzie quali l’OCSE e l’Organizzazione Mondiale della Sanità (OMS). Già nel 2015, l’OCSE aveva elaborato un quadro di riferimento per la gestione dei *big data* sanitari[[25]](#footnote-25). In esso si ribadisce la necessità di robusti meccanismi di sicurezza e controllo per garantire l’accesso ai dati sanitari tutelando al tempo stesso la privacy delle persone che rendono disponibili questi dati per finalità legate alla ricerca, all’innovazione tecnologica e alla pratica clinica. In un altro documento predisposto dall’OCSE si fa riferimento alle condizioni che devono essere soddisfatte affinché l’uso clinico dell’intelligenza artificiale meriti la fiducia di cittadini e operatori del settore[[26]](#footnote-26). Esse comprendono la minimizzazione delle conseguenze negative per individui, gruppi e professionisti, nonché la precedenza da accordare a soluzioni tecnologiche che affrontino problemi reali e che aggiungano valore per la società e per i pazienti. L’OMS dal canto suo ha istituito dei gruppi di lavoro specifici sulla digitalizzazione della medicina e pubblicato, nel 2019, delle linee guida in cui si sottolinea l’importanza che le soluzioni digitali integrino piuttosto che rimpiazzare le pratiche cliniche che ora sono affidate al personale sanitario (Labrique *et al*., 2020). A livello europeo, il *Joint Research Center* della Commissione Europea ha pubblicato nel 2018 un rapporto sull’intelligenza artificiale in medicina (Goméz-Gonsales, Goméz, 2020). Oltre alla sicurezza e affidabilità delle applicazioni mediche legate all’intelligenza artificiale, il rapporto sottolinea le questioni etiche legate alla trasparenza degli algoritmi, alla rappresentatività dei dati utilizzati per l’apprendimento automatico, e alla necessità di assicurare un controllo umano delle decisioni prese attraverso tali algoritmi (*human in the loop* (ARMI AUTONOME pp….).

1. Aspetti generali

Il dibattito sulle implicazioni della digitalizzazione in medicina è ampio e articolato anche al di fuori dell’etica intesa in senso stretto. Molti studiosi nel campo delle scienze sociali si occupano di questi temi. Tra loro, la sociologa australiana Deborah Lupton ha sviluppato un filone di studi critici sulla salute digitale. Secondo Lupton – che parte da posizioni vicine all’epistemologia critica di Michel Foucault – le pratiche di auto-monitoraggio e cura di sé che la medicina digitale supporta esprimerebbero assunti neoliberali circa il ruolo della responsabilità individuale nell’ambito della salute (Lupton, 2013; Lupton, 2014). Inoltre, il complesso di tecnologie, istituzioni e dati su cui si fonda la medicina digitale configura secondo Lupton un apparato di sorveglianza che, mentre promuove il ruolo attivo dei pazienti, li confina in realtà entro regimi di controllo nei confronti dei quali essi non hanno potere negoziale (Lupton, 2012). Anche gli usi legati alla salute pubblica, come ad esempio l’impiego di dispositivi mobili per la promozione della salute incarnerebbero in realtà assunti deterministici nei confronti delle tecnologie digitali, ovvero tenderebbero a considerare tali tecnologie come capaci di determinare trasformazioni sociali, senza tuttavia dare il giusto peso al modo in cui la società, la cultura e la politica – a loro volta – determinano tali tecnologie e ne articolano limiti e obiettivi (Lupton, 2014).

In uno studio molto approfondito, l’antropologa Veronica Barassi ha analizzato i processi di datificazione profonda che si mettono in moto con l’ingresso delle tecnologie digitali nell’ambito delle pratiche genitoriali – incluse quelle legate alla salute dei figli (Barassi, 2020). Barassi sottolinea la pervasività della raccolta dei dati fin dai primissimi anni di vita da parte delle grandi aziende tecnologiche, ma fa notare anche come sottrarsi a tali pratiche per i genitori non sia né facile, né a lungo andare auspicabile.

Queste constatazioni ci portano verso alcune considerazioni conclusive. La medicina digitale, sulla spinta del rapido sviluppo dell’intelligenza artificiale, si appresta secondo alcuni ad operare una vera rivoluzione nel modo in cui la medicina è concepita e praticata. Sul piano epistemologico, come abbiamo già avuto modo di sottolineare, in questo ambito si pone l’accento sulla capacità di previsione e classificazione degli algoritmi. L’interesse per l’aspetto legato alla capacità di spiegare sul piano causale la patogenesi e la guarigione appare secondario rispetto alla possibilità di sfruttare l’enorme potere di calcolo degli algoritmi nei confronti di quantità sempre crescenti di dati provenienti dalle fonti più disparate. Alcuni studiosi contrappongono dunque l’accuratezza degli algoritmi alla presunta esigenza di spiegare sul piano causale ciò che avviene in ambito clinico (*explainability*) (London, 2019). Non tutti però sono concordi nel dire che la medicina debba abbandonare ogni pretesa conoscitiva in favore della forza predittiva degli algoritmi – e chi scrive, per quello che conta, appartiene a questa schiera.

Innanzi tutto, studi recenti mostrano che solo in una piccola percentuale di casi, le prestazioni degli algoritmi che sono oggetto di pubblicazioni scientifiche sono validati su coorti di dati esterne e che in una percentuale ancora minore di casi questa validazione comprende un confronto tra le prestazioni dell’algoritmo e quelle di medici in carne e ossa (Liu *et al*., 2019).

Inoltre, l’enfasi sulla potenza predittiva dell’apprendimento automatico deve essere mitigata da considerazioni che hanno a che fare con l’importanza di includere parametri significativi sul piano causale al fine di migliorare le prestazioni dell’algoritmo stesso (Wilkinson *et al*., 2020). Lasciare che un algoritmo apprenda correlazioni all’interno di uno spazio parametrico indefinitamente ampio aumenta il rischio che esso apprenda correlazioni spurie, ovvero valide sul piano statistico per quel determinato insieme di dati ma non rilevanti sul piano causale. Ad esempio, un algoritmo potrebbe attribuire all’età dei pazienti un forte ruolo predittivo in ambito prognostico, senza che tuttavia questo parametro abbia un reale ruolo causale. Un tale esito potrebbe essere determinato dal fatto che l’età correli con un'altra variabile clinica che, a sua volta, svolge un reale ruolo causale, come ad esempio la presenza di malattie croniche. In altre parole, un algoritmo che apprenda correlazioni spurie rischia di fare previsione sbagliate, per esempio una prognosi sfavorevole per un paziente anziano ma in buone condizioni generali di salute. Questi problemi sono solo in apparenza problemi di natura tecnica o speculativa. In realtà il destino della medicina digitale basata sull’impiego dell’intelligenza artificiale è strettamente legato alla possibilità di un’integrazione ragionevole tra innovazione tecnologica e medicina come pratica sociale condivisa. In particolare il progresso tecnologico della medicina deve essere tale da promuovere gli interessi e i diritti dei pazienti sia sul piano clinico che sul piano etico. A questo fine, la discussione delle questioni affrontate nel corso di questo capitolo deve proseguire e portare a indicazioni normative concrete, nonché allo sviluppo di un quadro regolativo che favorisca in modo significativo la fiducia negli sviluppi tecnologici legati all’intelligenza artificiale. In mancanza di tali presupposti, i vantaggi potenzialmente offerti dall’intelligenza artificiale non potranno tradursi in benefici effettivi e i rischi connessi a questi avanzamenti finiranno per inibire il progresso stesso della medicina.

Capitolo 4

Etica dei veicoli autonomi

Fabio Fossa, Guglielmo Tamburrini

# Introduzione

In questo capitolo si offre un’ampia visione dell’etica dell’IA e della robotica attraverso l’analisi dei veicoli a guida autonoma (VA), passando in rassegna le principali problematiche morali e introducendo gli strumenti di etica normativa che possono aiutarci nella loro analisi. Cominceremo dalla questione delle collisioni inevitabili, che permette di familiarizzare con le prospettive dell’etica delle conseguenze e dei doveri. Tratteremo delle difficoltà relative alla distribuzione della responsabilità e del controllo, per passare poi a problemi legati alla libertà di guidare, alla privacy e alla sostenibilità ambientale. Il capitolo si conclude con un’analisi di varie politiche etiche sulla guida autonoma, esito regolativo delle indagini morali.

La discussione dei VA offre l’opportunità di presentare concetti chiave dell’etica sei sistemi autonomi e intelligenti, che ritroveremo più volte nel corso del libro. In questo senso, il capitolo vale come introduzione di ampio respiro al nostro tema, i cui numerosi aspetti saranno approfonditi nei capitoli successivi, sempre in relazione a tecnologie o sistemi specifici.

# Veicoli a guida autonoma e autonomia operativa

Per affrontare le questioni etiche legate alla guida autonoma bisogna anzitutto farsi un’idea di che cosa si tratta. A tale scopo è opportuno introdurre il concetto di *autonomia operativa*, che è fondamentale non solo per i VA, ma per ogni sistema tecnologico che tratteremo in questo libro. Con autonomia operativa si intende la capacità di un sistema di svolgere regolarmente uno o più compiti senza richiedere, in corso d’opera, alcun intervento umano. L’adattamento a contesti mutevoli e situazioni impreviste è un elemento importante qui, motivo per cui tecniche di IA, e più specificamente di apprendimento automatico (si veda il capitolo 2) giocano un ruolo fondamentale.

L’autonomia operativa di un sistema tecnologico è soggetta ad alcuni vincoli. Innanzitutto, si tratta di autonomia *per* alcuni compiti, e non altri. In più, è autonomia *da* qualcuno o qualcosa. Infine, è relativa a uno o più contesti di funzionamento e non altri, quindi a un *dove*. Il concetto di autonomia operativa “per-da-dove” si presta a utili specificazioni lungo le tre direttive. Nel caso dei veicoli si è proposto uno schema a 5 livelli di autonomia crescente, che vanno dall’automazione di funzioni di base come il governo di direzione di marcia, accelerazione, e mantenimento della velocità di crociera (L1) fino alla completa autonomia operativa dall’essere umano per tutte le funzioni di guida, sia su porzioni designate della rete stradale e date certe condizioni (L4), sia su tutta la rete stradale e in ogni condizione (L5) (SAE, 2016). Al momento, i modelli sperimentali più avanzati operano a L4, in porzioni della rete stradale e in condizioni in cui il conducente non è obbligato a tenere le mani sul volante, né a monitorare la navigazione, né a riprenderne il controllo su richiesta del veicolo. L5 rappresenta invece ancora una meta da raggiungere. Ai nostri scopi, con la sigla VA indicheremo veicoli di livello L4 o L5.

# Valutazioni d’impatto etico

Lo sviluppo di nuove tecnologie è spesso associato al conseguimento di obiettivi che hanno notevole rilevanza morale. In particolare, i VA promettono di migliorare sensibilmente la sicurezza dei trasporti. A cominciare dalla possibilità di ridurre drasticamente gli incidenti stradali grazie all’automazione della guida, proteggendo in questo modo l’integrità fisica e il benessere degli utenti della strada. Si ritiene che i VA rispetteranno le norme del codice della strada più scrupolosamente degli esseri umani, non potranno distrarsi, ubriacarsi o essere soggetti ad altre condizioni psicofisiche temporaneamente invalidanti. Poiché simili fattori causano gran parte degli incidenti (ISTAT, 2020) si prevede che la diffusione di veicoli ad autonomia crescente, e infine dei VA, consentirà di migliorare radicalmente la sicurezza stradale.

La maggiore sicurezza stradale non è l’unico vantaggio di carattere morale. Grazie ai VA, infatti, potranno acquisire maggiore indipendenza gli anziani e coloro che sono privi di patente per scelta o impedimento. Inoltre, come vedremo più nel dettaglio, la guida autonoma potrebbe avere un impatto positivo sulle emissioni di gas serra dovute al settore dei trasporti, ottimizzando il consumo energetico dei veicoli e inaugurando una gestione più efficiente del traffico.

L’introduzione dei VA porrà anche rischi per la privacy, la libertà, l’ambiente, l’attribuzione di responsabilità. Per cogliere le opportunità morali e mitigare i rischi è necessario prendere coscienza di entrambi, discuterne in sede di progetto e di utilizzo, elaborare linee guida etiche e regolative che offrano orientamento efficace a professionisti e utenti, e infine vigilare sulla loro applicazione. Dibattito integrato alla progettazione, impegno regolativo e vigilanza sono le dimensioni principali dell’etica dei sistemi autonomi e intelligenti, uno sforzo multidisciplinare per cui è richiesta la partecipazione di tutti gli attori sociali coinvolti.

# Collisioni inevitabili

La questione etica più nota in relazione alla guida autonoma riguarda le scelte che il sistema di controllo di un VA potrebbe avere il tempo di compiere in situazioni di collisione inevitabile.

Per “collisione inevitabile” si intende uno stato a partire dal quale il veicolo non può più evitare l’impatto con un ostacolo. Per quanto si dovrà fare di tutto per scongiurare che un VA entri in uno stato di collisione inevitabile, non si potrà escludere completamente questa evenienza. Condizioni avverse di illuminazione e corrispondenti errori di analisi percettiva hanno già generato stati di collisione inevitabile che hanno avuto conseguenze mortali per il passeggero del VA o per altri utenti della strada[[27]](#footnote-27). Altri eventi difficilmente prevedibili possono generare stati di collisione inevitabile: rocce o alberi caduti in strada, animali in transito sulla corsia di marcia, una persona distratta al volante di un altro veicolo, l’attraversamento incauto di un pedone.

Il problema etico delle collisioni inevitabili sorge all’intersezione di delega e autonomia operativa. Nella progressione da L1 a L5, l’essere umano concede all’autoveicolo una delega sempre più ampia, che comprende alcune valutazioni morali. Mentre guidiamo, infatti, esercitiamo la nostra capacità di giudizio morale, ad esempio circa il rischio a cui è lecito esporre gli altri utenti della strada.

Un conto, però, è delegare la manovra di parcheggio, un altro l’esercizio del giudizio morale. Si può pensare che la scelta su ciò che è bene e male o giusto e sbagliato debba rimanere nelle mani degli utenti umani. Tuttavia, in situazioni di collisione inevitabile sembra necessario delegare anche delle scelte morali. Un VA potrebbe scegliere l’angolo di impatto più vantaggioso per la protezione dei passeggeri e compiere in tempo utile altre scelte precluse ai conducenti umani, a causa dei loro tempi di reazione più lunghi, di forti interferenze emotive, di capacità più ridotte di elaborazione percettiva e cognitiva.

Se le cose stanno così, il problema etico delle collisioni inevitabili deve essere principalmente affrontato in sede di progettazione. La valutazione morale, infatti, non riguarda scelte già compiute, ma scelte da compiere in modo automatico tramite il sistema tecnologico. Non si deve giudicare ora se il VA abbia fatto allora una scelta moralmente buona o giusta, limitandosi a esprimere una valutazione morale *ex post* di una scelta già compiuta. Si deve invece valutare *ex ante* – in modalità, per così dire, predittiva – allo scopo di scegliere ora un’azione da compiere in seguito ed eventualmente in serie, su larga scala.

Decidere come il VA debba operare in simili occasioni è una faccenda estremamente delicata. Consideriamo due scenari schematici, nei quali il VA ha ancora la possibilità di scegliere tra diverse opzioni[[28]](#footnote-28).

Nel primo scenario [S1] vi sono due ciclisti che stanno attraversando la strada in direzione ortogonale al VA. Uno dei ciclisti indossa il casco, l’altro è senza casco. Il VA si trova in uno stato di collisione inevitabile con l’una o con l’altra delle due biciclette, ma ha ancora tempo sufficiente per scegliere con quale collidere. A parità di tutte le altre condizioni iniziali e del contorno (angolo d’impatto, masse coinvolte, velocità ecc.), che cosa dovrà fare il VA?

Nel secondo scenario [S2], il VA si trova improvvisamente davanti due persone che hanno iniziato ad attraversare la strada. In questo caso, lo stato di collisione inevitabile lascia aperte due alternative: investire entrambi i pedoni che si trovano sulla sua traiettoria, oppure sterzare bruscamente verso un fosso che costeggia la strada, risparmiando così i pedoni, ma mettendo a repentaglio la vita del suo unico passeggero. Che cosa dovrà fare il VA in questo secondo scenario?

# Etica delle conseguenze

Le domande sorte dagli scenari appena presentati sono complesse. Su quali strategie di analisi possiamo contare per esplorarne il profilo etico?

L’etica o filosofia morale si occupa di indagare le questioni relative al bene, al giusto e alla vita felice con lo scopo di orientare l’esistenza degli esseri umani. Nel corso della sua storia, alcuni approcci sono emersi come particolarmente significativi, in quanto colgono elementi fondamentali della nostra esperienza morale. Tra questi, l’etica delle conseguenze occupa una posizione di rilievo.

Secondo l’etica delle conseguenze, una decisione o un’azione deve essere valutata come moralmente buona o cattiva tenendo soltanto conto di quelle che sono le sue conseguenze. Poiché pone l’attenzione esclusivamente sulle conseguenze delle decisioni e delle azioni, questo approccio considera come fattori irrilevanti per la valutazione morale delle scelte il rispetto o la violazione di un presunto dovere incondizionato, ma anche le intenzioni e le ambizioni, gli atteggiamenti mentali virtuosi, magnanimi o meschini che ci spingono ad agire.

La cornice generale dell’etica delle conseguenze non fornisce però elementi sufficienti per emettere giudizi morali su azioni già compiute o per guidare la scelta di azioni da compiere. Per sviluppare criteri operativi di questo tipo bisogna specificare che cosa si intende per conseguenza moralmente buona o cattiva. Ad esempio, si può optare per il criterio del benessere collettivo e assumere il punto di vista di un osservatore imparziale che dà il medesimo peso al benessere di ciascun utente della strada. Oppure, altri criteri possono essere assunti – come la felicità o l’utilità – ma sempre alle medesime condizioni. L’etica delle conseguenze deve poi fare i conti con la difficoltà di pesare tra loro le ipotetiche conseguenze di un’azione e di azioni alternative. All’ambizione di individuare l’opzione migliore spesso si sostituisce l’obiettivo più modesto di individuare una o più alternative che siano *sufficientemente* buone, lasciando all’agente la possibilità di scegliere liberamente tra esse (Slote, Pettit, 1984).

Se proviamo ad applicare la prospettiva dell’etica delle conseguenze alle collisioni inevitabili, una possibile linea guida è la richiesta di minimizzare imparzialmente i danni attesi per le persone. La massima pare ragionevole, non troppo complicata, e potenzialmente implementabile. Basta però calarla nel contesto di S1 per farsi un’idea della difficoltà del compito.

Se applichiamo la massima proposta, in S1 il VA dovrà essere programmato per investire il ciclista che indossa il casco. Certo può sembrare inaccettabile investire sistematicamente chi rispetta delle elementari norme prudenziali di condotta invece del ciclista senza casco. Ma nell’ottica dell’etica delle conseguenze, una volta assunto il punto di vista di un agente imparziale, ci si propone esclusivamente di minimizzare i danni attesi, senza lasciarsi influenzare dalle virtù (dal rispetto dell’obbligo del casco) o dalle inadempienze delle persone coinvolte. Rispetto a quello senza casco, il ciclista che lo indossa gode di maggiore protezione, e perciò le conseguenze attese di questa scelta, a parità di altre condizioni, saranno verosimilmente meno dannose.

Subito si affaccia un’altra difficoltà: il peso etico delle conseguenze muta con eventuali iterazioni di tale scelta e con l’estensione della finestra temporale lungo la quale è opportuno effettuare una valutazione delle conseguenze. Infatti, la scelta di investire il ciclista con il casco appare giustificabile solo nel caso singolo, ma non in caso di metodiche iterazioni. Se si estende l’orizzonte temporale, nuove conseguenze devono essere considerate. Se i VA, in scenari riconducibili a S1, investiranno sistematicamente i ciclisti che indossano il casco, allora i ciclisti potranno convincersi che indossare il casco sia un fattore di rischio non trascurabile. L’uso del casco potrebbe essere scoraggiato da questa convinzione e la popolazione di ciclisti protetta dal casco si ridurrà, con conseguenze negative per il benessere collettivo.

Se si considera una valutazione di lungo periodo, i benefici attesi non riguardano situazioni isolate, ma una collezione statisticamente significativa di eventi. Casi simili, nel gergo dell’etica delle conseguenze, sono discussi dal “consequenzialismo delle regole”, un approccio distinto dal “consequenzialismo dell’atto”. Il consequenzialismo delle regolerichiede di agire senza limitarsi a valutare le conseguenze del singolo atto, ma in base al sistema di regole dalla cui prevalente accettazione e applicazione sono attese le migliori conseguenze per il benessere collettivo nel lungo periodo.

In definitiva, a seconda di come si interpretano i vincoli dell’etica delle conseguenze – riferiti a un orizzonte ampio o ristretto di conseguenze previste – si ricavano indicazioni divergenti. All’interno di un orizzonte ristretto di eventi attesi, il VA dovrà sempre investire il ciclista con il casco. Allargando l’orizzonte temporale, si arriva alla conclusione opposta che il VA non dovrà sempre investire il ciclista con il casco. Ma in questo secondo caso rimane aperta la questione di che cosa dovrà fare.

Anche S2 presta il fianco a simili osservazioni. Con lo sguardo rivolto a un orizzonte temporale ristretto, il sistema di controllo del VA deve evitare di investire i due pedoni per minimizzare i danni agli esseri umani coinvolti ‘qui e ora’. Invece, considerando un orizzonte più ampio, riemerge il dubbio che questa sia la decisione migliore. Un potenziale acquirente potrebbe infatti desistere dall’acquisto del VA perché il sistema di controllo non esclude il sacrificio dei passeggeri; una contrazione delle vendite dovuta al diffondersi di questa notizia potrebbe a sua volta compromettere gli obiettivi commerciali delle aziende automobilistiche; insieme al loro mancato conseguimento, si vanificherebbe la prospettiva di ridurre, grazie alla diffusione dei VA, il numero e la gravità degli incidenti stradali.

Per concludere, l’etica delle conseguenze sembra permettere un’analisi imparziale dei problemi morali legati alle collisioni inevitabili, il che può risultare un vantaggio non indifferente in sede di implementazione. Tuttavia, difficoltà relative alla definizione del criterio di scelta, alle metodologie di comparazione delle conseguenze ipotetiche, all’estensione della finestra temporale e agli effetti che l’iterazione della scelta può avere sul contesto complicano l’elaborazione di linee guida chiare e definitive.

# Etica dei doveri

Alle difficoltà interne all’etica delle conseguenze si sommano aspetti da essa trascurati, che sono invece centrali in altre prospettive. Si pensi al caso di un genitore che si serve abitualmente del suo VA come mezzo di trasporto quotidiano a uso del figlio nel tragitto tra casa e scuola. Supponiamo che la protezione del figlio sia per questo genitore un dovere prioritario, da anteporre a ogni altra cosa, indipendentemente dagli esiti attesi in base a un calcolo delle conseguenze. Cosa dovrà fare il sistema di controllo della navigazione del VA per gestire le collisioni inevitabili? Potrà essere programmato per anteporre la protezione del figlio del proprietario a ogni altra considerazione, derogando così al vincolo di imparzialità e all’obiettivo di minimizzare i danni per tutte le persone coinvolte?

Trascurare la minimizzazione dei danni attesi per rispettare un dovere genitoriale potrebbe essere una scelta permessa dalla prospettiva dell’etica dei doveri o etica deontologica. Quest’ultima ammette che vi siano doveri e diritti delle persone a cui non è permesso derogare in base a un bilancio delle conseguenze attese, se non in circostanze limitate e ben specificate. L’etica dei doveri, al contrario dell’etica delle conseguenze, può quindi subordinare il perseguimento di ciò che è bene fare per il benessere della collettività al perseguimento di ciò che è giusto fare per rispettare determinati doveri.

In ottica deontologica, i doveri possono avere diverse origini. Nel caso introdotto, il dovere di proteggere i propri figli dipende dal ruolo che si ricopre in relazione ad altri e alle responsabilità che ne conseguono. Si parla, in casi simili, di etica dei ruoli. Alla stessa categoria appartiene l’etica professionale, che ruota intorno ai doveri e alle responsabilità proprie di chi svolge un determinato ruolo nel mondo del lavoro. Esistono però anche approcci filosofici più generali all’etica deontologica, tra cui il più noto rimanda alla filosofia pratica di Immanuel Kant. Secondo il filosofo tedesco, il dovere proviene direttamente dalla ragione, dalla quale deriva una legge che ogni essere umano non può che riconoscere in quanto, appunto, essere razionale. In questo senso il dovere è universale: riguarda tutti noi indipendentemente dai ruoli che occupiamo e richiede rispetto in ogni situazione, quali che siano le conseguenze. Kant diede il nome di imperativo categorico a questa legge e ne offrì diverse formulazioni, tra cui ad esempio: «Agisci unicamente secondo quella massima in forza della quale tu puoi volere, nello stesso tempo, che essa divenga una legge universale» (Kant, 1785, trad. it. p. 79). L’intento di fissare doveri inappellabili, precisi e riconoscibili ha motivato l’istituzione dei diritti fondamentali degli esseri umani, elemento cruciale della cultura moderna esemplificato dalla Dichiarazione Universale dei Diritti Umani.

La dimensione concreta dei ruoli e quella più astratta dell’universale razionalità umana possono entrare in conflitto. Mettersi nei panni dei diversi soggetti coinvolti in collisioni inevitabili lo mostra facilmente. Potrebbe mai un genitore perdonarsi di aver utilizzato un VA che ha sacrificato uno dei suoi figli nel tragitto da casa a scuola? Ma come non preoccuparsi dei diritti degli altri utenti della strada e delle tensioni sociali che possono nascere dalla loro programmatica negazione? Un VA che trasporti un bambino e che sia stato programmato per rispettare i doveri genitoriali del proprietario investirà un numero maggiore di pedoni se l’unica alternativa è quella di schiantarsi contro un albero o finire in un fosso. In risposta a un episodio di questo genere, gli abitanti del quartiere costituiranno un cosiddetto gruppo da “effetto *nimby*” (*Not In My Back Yard*, Treccani, 2013), in analogia con quanto è già accaduto per opporsi alla localizzazione di impianti industriali inquinanti, depuratori di acque nere o bruciatori di rifiuti solidi urbani. Il gruppo *nimby* si opporrà alla circolazione di VA programmati nel rispetto dei doveri genitoriali, raccoglierà firme per una petizione e organizzerà una manifestazione in tal senso. L’azienda e i programmatori saranno accusati di avere deliberatamente scelto di sacrificare un numero maggiore di persone rispetto all’unico passeggero del VA.

Sebbene da una prospettiva filosofica non dovrebbe contare in modo decisivo, l’impatto del ruolo sulla percezione del dovere ha basi empiriche ben attestate. Alcune interviste e indagini psicologiche mostrano che gli intervistati, se vestono i panni di un pedone o di un astante, preferiscono che i VA siano programmati per minimizzare i danni, salvando il maggior numero possibile di vite umane, anche se ciò comporta il sacrificio di qualche passeggero. Ma gli stessi intervistati affermano di non voler viaggiare a bordo di un VA le cui politiche di controllo non diano priorità alla protezione dei passeggeri (Bonnefon, Sharif, Rahwan, 2016). In definitiva, a seconda dei ruoli assunti, si alterna in una stessa persona la preferenza per una politica di gestione delle collisioni che sia indirizzata a minimizzare imparzialmente i danni delle persone coinvolte con la preferenza per una politica indirizzata a privilegiare i diritti dei passeggeri del VA. Una persona che oscilla nelle sue preferenze in funzione del ruolo che riveste momentaneamente – di passeggero o di pedone – potrà sollevare in ogni caso delle critiche nei confronti dei programmatori e dei produttori, qualunque sia la politica prescelta dall’azienda e implementata dai programmatori: da passeggero, perché non riceve protezione sufficiente da parte di un veicolo imparziale al quale affida la sua incolumità; da pedone, perché il VA svaluta la sua vita rispetto a quella dei passeggeri, se privilegia incondizionatamente la protezione di questi ultimi.

# Controllo umano e distribuzione della responsabilità

I quesiti posti dalle collisioni inevitabili permettono di introdurre un secondo grande tema che riguarda ogni sistema dotato di autonomia operativa: l’attribuzione della responsabilità. Questo problema, di natura tanto morale quanto legale, nasce anch’esso al confine tra autonomia operativa e controllo umano. Come abbiamo visto, le tecnologie che abilitano l’autonomia operativa dei sistemi ci permettono di delegare compiti senza che sia necessario (o perfino possibile) intervenire in prima persona. Dal minor controllo diretto esercitato sul compito che viene automatizzato nasce un peculiare problema di attribuzione di responsabilità: chi è responsabile dei danni arrecati a cose o a persone? Consideriamo il caso di un ciclista investito e ucciso senza che abbia colpa alcuna. Chi è responsabile?

Passando in rassegna i diversi attori umani potenzialmente responsabili dell’accaduto – produttori, programmatori, passeggeri e così via – si scopre ben presto quanto sia difficile rispondere a questa domanda, in quanto nessuno di essi ha un controllo diretto e completo del sistema. Attribuire la responsabilità allo stesso VA, che in effetti può sembrare l’unico ente davvero in controllo, è però una strategia paradossale, in quanto i sistemi tecnologici non presentano allo stato dell’arte i requisiti perché venga loro sensatamente attribuita responsabilità di alcunché. È un errore categoriale ascrivere lode, biasimo e responsabilità morali alle tecnologie attuali che godono solo di autonomia operativa. La riflessione etica sui sistemi autonomi e intelligenti attuali è esclusivamente incentrata sull’uomo come agente morale, su ciò che è bene o giusto che l’uomo faccia avvalendosi dell’autonomia operativa delle tecnologie, sulle responsabilità e sui giudizi di lode o di biasimo morale che riguardano coloro che sono coinvolti nelle attività di impulso politico ed economico, di ideazione, progettazione, realizzazione, diffusione, gestione e uso di questi sistemi. Tuttavia, in assenza di negligenze acclarate, nessuno degli esseri umani coinvolti in queste attività sembra a pieno titolo moralmente responsabile di danni provocati da sistemi ad autonomia operativa. Si aprono così dei vuoti tanto di responsabilità che di retribuzione morale[[29]](#footnote-29) che rischiano di minare la fiducia sociale nella tecnologia e di comprometterne l’adozione (Nyholm, 2018).

Il fatto che nessuno sia pienamente responsabile non esclude che alcune parti coinvolte possano essere in qualche misura corresponsabili dell’accaduto. Così come l’incidente è causato dal concorso di una pluralità di agenti a vario titolo coinvolti nello sviluppo, nella gestione e nel dispiegarsi delle azioni del VA, così la responsabilità (e la relativa retribuzione) può essere distribuita lungo la rete degli agenti coinvolti. Se l’idea di distribuire la responsabilità in questo modo ha una sua attrattiva, bisogna però considerare il rischio posto dal cosiddetto problema delle molte mani. Se una molteplicità di soggetti partecipa a vario titolo ad un medesimo processo, come accade con i VA, può diventare assai difficile in pratica distribuire la responsabilità in modo tanto giusto quanto equo, con il rischio connesso di generare un nuovo vuoto di responsabilità e retribuzione.

Dal momento che buona parte delle difficoltà ha a che vedere con la perdita di controllo umano diretto sul funzionamento, sulle decisioni e sull’operato del sistema tecnologico, una strategia per affrontare tali problemi di attribuzione di responsabilità richiede di mettere gli utenti in condizione di riassumere il controllo del VA in situazioni moralmente impegnative, nelle quali imporre l’esercizio del giudizio umano. Una siffatta richiesta di *controllo umano significativo* (Amoroso, Tamburrini, 2019; Santoni de Sio, van den Hoven, 2020) deve affrontare molte sfide di carattere tecnico e teorico che riguardano tanto la creazione di interfacce per la collaborazione tra essere umano e macchina, quanto le necessità di massimizzare le prestazioni dei VA (potenzialmente disturbate da intromissioni umane) e di mantenere una distribuzione equa delle responsabilità.

Come colmare i vuoti di responsabilità e di retribuzione è un problema che accompagnerà l’inserimento dei sistemi autonomi e intelligenti nelle società umane. Questi pochi cenni bastino a prendere coscienza del problema, che verrà discusso più nel dettaglio nel capitolo 7 dedicato alle armi autonome.

# Oltre le collisioni inevitabili

# Libertà di guidare

L’introduzione di tecnologie ad autonomia operativa porta con sé impatti raramente confinati a un’unica questione etica. Nel caso dei VA, i dubbi legati alle collisioni inevitabili non sono che uno dei molti aspetti etici degni di considerazione. Un altro aspetto riguarda l’impatto dell’autonomia operativa dei sistemi sull’autonomia umana. La difficoltà è duplice: riguarda sia l’atto stesso del guidare che le regole di comportamento morale da implementare nei VA. Vediamo brevemente entrambi i casi.

I VA promettono di raggiungere livelli di sicurezza sulla strada assai più elevati di quelli attuali. In questo senso, la tecnologia è un fattore importante di benessere collettivo. Ma la ricerca del benessere collettivo non si accorda sempre con l’esercizio della libertà personale. Supponiamo di essere giunti alla conclusione, suffragata da prove sperimentali, che i VA generano rischi minori degli esseri umani alla guida. Da quel momento in poi, il rischio aggregato per gli utenti della strada aumenterà, se si continua a concedere la libertà di guidare agli esseri umani invece di affidarsi esclusivamente ai VA. Quali decisioni normative si dovranno prendere a quel punto? Sarebbe giusto introdurre una norma che proibisca agli esseri umani di guidare (Müller, Gogoll, 2020)?

Lo scenario evidenzia come l’autonomia operativa dei VA possa condurre all’opposizione di due valori radicati nella nostra cultura: l’idea del benessere collettivo e il diritto all’esercizio della libertà, in questo caso, di guidare e al piacere che ne possono trarre gli appassionati. Oltre alle difficoltà concettuali, una norma che imponga il divieto di guida per gli esseri umani avrebbe vita assai dura. Sarebbe probabilmente osteggiata dai produttori di auto sportive, poiché il piacere degli appassionati è un forte incentivo all’acquisto. L’iter potrebbe essere rallentato dall’industria automobilistica in generale, gravata dal problema di smaltire le scorte di veicoli tradizionali. E forse sarebbe ugualmente ostacolato da abitudini psicologiche e consuetudini sociali.

Un orientamento favorevole al divieto potrebbe però svilupparsi all’improvviso, a causa di un grave incidente provocato da un conducente assonnato, il quale ha dimenticato di attivare la modalità di guida autonoma già disponibile sulla sua auto; potrebbe rafforzarsi dopo la condanna penale dell’automobilista per guida negligente; diventare prevalente dopo l’avvio di una causa collettiva(*class action*) promossa dall’associazione dei parenti di vittime della strada nei confronti delle case automobilistiche che non producono ancora VA, accusate per questo di vendere veicoli che non rispettano standard di sicurezza facilmente raggiungibili. L’opinione pubblica potrebbe infine cambiare tanto da stigmatizzare un automobilista alla guida alla stregua di chi si ostina ancora oggi a fumare nei locali pubblici in spregio alla salute altrui.

#

# Libertà di scelta e paternalismo morale

Un discorso simile può essere fatto circa le scelte etiche da implementare nei VA. Poiché l’analisi morale non conduce a conclusioni definitive, applicabili indiscriminatamente a tutti i veicoli, ingegneri e produttori non possono farsi carico di scegliere quale prospettiva etica inserire nel loro sistema. Per evitare il fuoco incrociato sulle aziende automobilistiche e sui programmatori da parte di pedoni e passeggeri, si potrebbe allora richiamare il principio della libertà di scelta e lasciare che sia il proprietario del veicolo a scegliere.

In tal caso, si potrebbe dotare il VA di due diverse politiche di controllo: una che possiamo chiamare IMP, orientata alla minimizzazione imparziale dei danni, e un’altra che possiamo chiamare PASS, orientata prioritariamente alla protezione dei passeggeri. Il veicolo sarà dotato di un interruttore per commutare tra IMP e PASS (Contissa, Lagioia, Sartor, 2017). Lasciando al proprietario il compito di scegliere, ed eventualmente di modificare nel tempo le sue preferenze, l’azienda non scoraggerebbe l’acquisto dei VA. Si difenderà la libertà di scelta degli acquirenti e si proteggeranno al contempo gli interessi aziendali, favorendo anche una maggiore sicurezza stradale grazie alla diffusione dei VA. E poiché il proprietario del veicolo può scegliere altruisticamente la politica IMP, l’azienda non potrà essere accusata di scartare la prospettiva imparziale sul benessere collettivo al solo scopo di incrementare le vendite. Infine, l’azienda non potrà essere accusata nemmeno di paternalismo morale verso i proprietari e i passeggeri del veicolo, perché si astiene dal decidere al posto loro che cosa sia giusto fare in ciascuna situazione.

La soluzione liberale, tuttavia, ha i suoi limiti. In particolare, non risolve il conflitto morale: viene ridotto al silenzio il punto di vista del pedone, sistematicamente danneggiato dalla politica PASS; e si sacrifica l’istanza di imparzialità rappresentata dall’etica delle conseguenze, trascurando il fatto che essa fornisce, nella prospettiva di una riduzione delle vittime della strada, la più potente motivazione morale per l’introduzione dei VA. Infine, è verosimile che i proprietari di VA convergeranno sulla politica PASS, che non persegue la minimizzazione dei danni alle persone. Saranno in questo orientati dai vantaggi che potrebbero riceverne come passeggeri e da scelte analoghe fatte da altri proprietari. Poiché la maggior parte dei passeggeri assumerà in altre occasioni anche il ruolo di pedone, l’adozione generalizzata della politica PASS comporterà per quasi tutti un incremento del rischio. Lasciare che i conflitti morali che riguardano un progetto collettivo di innovazione tecnologica siano risolti caso per caso dai singoli, sapendo inoltre che il risultato aggregato delle singole scelte sarà verosimilmente tale da vanificare una parte dei vantaggi attesi dalla diffusione dei VA, sembrerebbe una decisione poco lungimirante, se non ipocrita. L’accusa di paternalismo morale[[30]](#footnote-30), inoltre, non pare del tutto appropriata quando si tratta di difendere il benessere collettivo. In definitiva, scaricare sulle spalle dei singoli le scelte morali più impegnative nasconderebbe una rinuncia all’esercizio della ragione pubblica.

# Privacy e cybersicurezza

I VA non sono concepiti come dispositivi isolati, ma come parti di un sistema distribuito che si estende oltre la carrozzeria e concorre con tutti i suoi componenti a gestire la navigazione stradale. Per funzionare correttamente, i VA necessitano infatti di sensori diffusi nell’ambiente e potenti centri di elaborazione dati per gestire le esigenze computazionali di una flotta di veicoli in costante comunicazione gli uni con gli altri. Per evitare collisioni prevedibili e ottimizzare il traffico o la gestione del parcheggio, ad esempio, l’interconnessione tra veicoli è cruciale. Tuttavia, lo scambio di informazioni su rotte, destinazioni e abitudini di trasporto – così come la raccolta di dati biometrici e di immagini ambientali – coinvolge dati sensibili da proteggere adeguatamente per rispettare la privacy degli utenti e degli altri soggetti coinvolti. In più, alcuni contrasti di valore tipici del discorso sulla privacy si ripropongono nel contesto della mobilità autonoma. Ad esempio, l’accesso ad informazioni dettagliate circa i percorsi frequenti degli utenti potrebbe avere un impatto positivo sulla gestione dei volumi di traffico e, dunque, permettere una migliore ottimizzazione dei trasporti, con potenziali effetti benefici sull’impatto ambientale. Tuttavia, ciò richiederebbe la condivisione di dati dalla cui analisi possono essere inferite informazioni personali e sensibili.

Le tecnologie digitali mettono a dura prova il rispetto della privacy e ne trasformano profondamente i caratteri. Ad esempio, strategie classiche per limitare i danni conseguenti alla condivisione di informazioni sensibili, come l’anonimizzazione e il consenso informato, non sembrano riuscire a tenere il passo delle novità introdotte dallo sfaccettato fenomeno dei big data (Barocas, Nissenbaum, 2014). Ciò, tuttavia, non implica che la privacy abbia un’importanza secondaria nelle società dell’informazione o che la sua erosione sia un prezzo da pagare per godere di benefici maggiori. La protezione della sfera privata continua infatti ad essere rilevante sia da una prospettiva individuale che sociale.

Dal punto di vista dei singoli cittadini, la privacy fornisce le condizioni per uno sviluppo della personalità libero da ingerenze e sguardi indesiderati, per il godimento dell’intimità, l’esercizio del pensiero critico e la sperimentazione anticonformista di idee e comportamenti. In un certo senso, la nostra stessa identità personale è inscindibile dalle informazioni che disseminiamo consapevolmente o meno, per cui la protezione della privacy è funzionale alla costruzione del proprio sé (Rodotà, 2005; Floridi, 2014). In più, la possibilità di rivelare informazioni personali diverse a persone diverse sta alla base della pluralità delle relazioni sociali che intratteniamo, ad esempio, con amici, parenti, coniugi, colleghi e professionisti, e dunque del benessere individuale che deriva dal godere di una rete di rapporti gratificanti (Rachels, 1975). Ognuna di queste relazioni appartiene a un contesto specifico e si struttura secondo regole esplicite e implicite che riguardano proprio i tipi di informazioni che è adeguato condividere, i modi in cui farlo e le aspettative circa il comportamento di chi entra in possesso di informazioni che ci riguardano (Nissenbaum, 2004). La mancata protezione della privacy mette a rischio tutto ciò e deve essere riaffermata anche in un mondo in cui raccogliere, analizzare, riunire, condividere, immagazzinare e conservare dati è sempre più facile, conveniente e profittevole.

Non è vero, infatti, che la privacy protegga solo chi ha qualcosa da nascondere, oppure che si tratti di un valore sopravvalutato, in quanto saremmo già immersi inesorabilmente in una società della sorveglianza (Johnson, Miller, 2009). Le conseguenze di decisioni prese sulla base di informazioni inaccurate o irrilevanti possono essere estremamente negative anche per chi non commette alcun reato o portare a esiti discriminatori e ingiusti. Tuttavia, connettere le nostre scelte in materia di condivisione dei dati personali alle conseguenze negative che ne possono seguire non è né facile né ovvio, motivo per cui le scelte dei singoli – ammesso e non concesso che indichino disinteresse per la protezione dei propri dati – non costituiscono una ragione per abbandonare il discorso sulla privacy.

Resta però vero che una maggiore quantità, qualità e circolazione di informazioni rappresenti una risorsa importante per prendere decisioni e offrire servizi migliori, a beneficio di tutti. L’attenzione per la privacy non deve far perdere di vista i vantaggi che possono seguire dalla mole di dati che siamo in grado di raccogliere e analizzare. Da una prospettiva consequenzialista, si può sostenere che la privacy individuale valga nella misura in cui non ostacoli il benessere di molti. Si è perciò spesso argomentato per una sua limitazione, soprattutto nei casi che coinvolgono beni collettivi come la sicurezza o l’efficienza. Per rispondere adeguatamente a simili pressioni, molti hanno però sottolineato il valore non solo individuale, ma sociale della privacy, conferendogli dignità uguale ad altri valori collettivi (Regan, 2015; Barocas, Nissembaum, 2014). La protezione della privacy è infatti cruciale per lo sviluppo e l’esercizio di qualità come il pensiero critico, la creatività e l’anticonformismo, ingredienti fondamentali per l’organizzazione democratica e per l’instaurazione di un regime sociale trasparente dove opportunità e servizi sono distribuiti secondo giustizia e equità e i cittadini sono protetti da manipolazioni. Alla luce di quanto detto, diventa importante mettere in discussione le motivazioni per cui si ricorre alla raccolta dei dati nei diversi contesti pratici e misurarne l’opportunità in relazione ai valori sociali, agli interessi e agli scopi che è bene perseguire in società democratiche e liberali. Inoltre, occorre introdurre strumenti sia tecnologici che normativi (si veda, ad esempio, il Regolamento Generale sulla Protezione dei Dati dell’Unione Europea – RGPD, 2016) che regolino il trattamento dei dati personali e concorrano al bilanciamento di benefici e rischi.

Per queste ragioni, l’attenzione per la tutela della privacy è un tema centrale della riflessione etica sui sistemi digitali, robotici e di IA. La guida autonoma non fa eccezione. Delle venti raccomandazioni sull’etica dei VA emanate recentemente dalla Commissione Europea (Horizon, 2020), ben nove coprono temi che gravitano intorno alla necessità di proteggere la privacy degli utenti e si estendono fino a includere trasparenza e spiegabilità. Le linee guida sottolineano la necessità di informare adeguatamente i soggetti coinvolti circa i dati raccolti e la finalità dell’operazione, oltre a facilitarne la gestione e il controllo. Inoltre, si chiarisce che la raccolta di dati non essenziali per il funzionamento dei VA necessita di un consenso esplicito. La protezione contro abusi, manipolazioni o accessi non autorizzati è anche un requisito. In più, si invita ad elaborare nuove forme di consenso informato che non precludano l’accesso a servizi in caso di rifiuto, a elaborare standard che promuovano la privacy e a proteggere gli utenti da discriminazioni e dalla raccolta, condivisione e analisi di dati potenzialmente problematici.

Connesso al discorso sulla privacy è la preoccupazione per la cybersicurezza, approfondita nel capitolo 6. Già abbiamo accennato all’importanza di mettere al sicuro i dati personali raccolti dai VA da attacchi esterni. I problemi di cybersicurezza non finiscono certo qui: secondo l’Agenzia Europea per la Cybersicurezza rappresentano invece uno dei principali ostacoli all’adozione della tecnologia (Dede *et al*., 2021). Naturalmente, la resistenza del sistema a attacchi hacker o manomissioni di vario genere messi in atto sfruttando le vulnerabilità software, hardware e del collegamento in rete dei veicoli è fondamentale anche per evitare intromissioni malevoli nel funzionamento del sistema e i danni conseguenti. Privacy e sicurezza da attacchi esterni sono precondizioni fondamentali di uno sviluppo eticamente improntato dei VA (Taheiagh, Lim 2019).

# Sviluppo sostenibile e impatto climatico

Tra i vantaggi prospettati dallo sviluppo della guida autonoma si include la riduzione dell’impatto ambientale del settore dei trasporti. Questo settore si colloca al secondo posto tra le attività umane che maggiormente contribuiscono alle emissioni di gas serra, preceduto solo dalla produzione di elettricità e calore. Una progettazione orientata all’efficienza energetica dei VA consentirà di ridurre il consumo di carburante per chilometro percorso. Altri benefici potranno derivare da una gestione del traffico volta a evitare le congestioni e ridurre i tempi di ricerca di un parcheggio o dalla diffusione di flotte di VA da utilizzare a pieno carico attraverso servizi di condivisione dei percorsi (*car sharing*). Ma se ciò accadrà davvero dipenderà dalla messa in opera di un ampio ventaglio di politiche di progettazione e gestione della rete stradale, delle infrastrutture informatiche e di telecomunicazione, degli interventi urbanistici. Senza azioni di questo tipo, la diffusione di VA potrà avere un “effetto rimbalzo”, incentivando talmente il trasporto privato su gomma da controbilanciare i benefici derivanti dalla maggiore efficienza energetica dei singoli veicoli. Vediamo perché.

La spesa minore in termini di carburante per chilometro percorso potrà incoraggiare la gente a fare uso di autovetture private a guida autonoma. La comodità di viaggiare senza avere l’incombenza della guida e potendosi dedicare ad attività lavorative o di svago potrà incentivare il pendolarismo su percorsi più lunghi e il trasferimento verso zone periferiche meno costose o congestionate. Inoltre, le persone che oggi non possono o non desiderano guidare potranno sfruttare le nuove opportunità di mobilità offerte dai VA privati (Taiebat *et al.*, 2018). Nonostante la maggiore efficienza energetica di ogni VA rispetto ai precedenti modelli di autoveicoli asserviti al controllo umano, ciò potrà generare problemi di sostenibilità ambientale da affrontare sviluppando adeguate politiche di gestione del trasporto pubblico e privato. Infine, al contrario degli attuali veicoli, i VA avranno bisogno di un’infrastruttura imponente di centri di elaborazione dati ad alto consumo energetico (Belkhir, Elmeligi, 2018). Per un approfondimento sullo sviluppo sostenibile e l’impatto ambientale delle tecnologie digitali rimandiamo al capitolo 8.

# Dall’etica applicata alle politiche etiche

I problemi fin qui esaminati sono troppo grandi perché la responsabilità della loro soluzione ricada sulle spalle degli ingegneri, delle case automobilistiche e dei loro clienti. È necessario un confronto che coinvolga tutte le parti interessate, rivolto alla formulazione e all’adozione di una politica etica condivisa per la progettazione e l’utilizzazione dei VA. Più in generale, nella discussione pubblica sui sistemi autonomi e intelligenti, la riflessione etica può fornire un contributo significativo, delineando un ventaglio di politiche etiche che abbiano l’obiettivo di superare, seppure in modi diversi tra loro, le situazioni di stallo dovute ai conflitti e alle tensioni tra etica delle conseguenze, etica dei doveri ed eventuali altre cornici dell’etica normativa.

# Politiche libertarie e segregazioniste

Per presentare il tema delle politiche etiche in relazione ai VA è utile introdurre due politiche notevoli e alternative tra loro: la politica segregazionista e quella libertaria. La prima di queste confina la circolazione dei VA su strade a essi riservate, nelle quali vi è una probabilità trascurabile di collisioni inevitabili e dell’insorgere dei relativi conflitti morali. Nel mondo altamente prevedibile di strade progettate a uso esclusivo dei VA, non ci sarà posto per esseri umani al volante, pedoni, ciclisti o bambini intenti a giocare sul margine della strada. Saranno rispettate le distanze tra veicoli e inutili le precauzioni da applicare altrove per garantire la non discriminazione e l’esenzione delle persone non coinvolte nelle collisioni. Si potrà sempre perseguire la minimizzazione dei danni come richiede l’etica delle conseguenze.

La politica segregazionista è una soluzione tecnicamente adatta al breve periodo, da applicare in attesa di progressi che permetteranno di progettare VA capaci di gestire i comportamenti poco prevedibili di pedoni e conducenti in situazioni di traffico misto. Se invece viene adottata come soluzione definitiva, rivela un punto di debolezza dalla prospettiva dell’etica delle conseguenze. Insistendo sulla segregazione anche quando i VA garantiranno maggiore sicurezza dei conducenti umani in situazioni di traffico misto, non si potranno usare i VA per ridurre gli incidenti in tali situazioni.

Un’altra politica etica notevole è di tipo libertario. Nella versione “ristretta”, si consente la scelta tra dueimpostazioni etiche alternative per il sistema di controllo dei VA, secondo quanto discusso in 3.2. Visti i problemi legati ad un’adozione diffusa di PASS, le due alternative potrebbero essere presentate in modo tale da favorire con una “spinta gentile” (*nudge*) la scelta degli utenti per la politica IMP, rendendo una deviazione da tale percorso ancora accessibile, ma psicologicamente più costosa (Thaler, Sunstein, 2008; ROBOTICA SOCIALE pp….). All’atto dell’acquisto, per esempio, il VA potrebbe essere predisposto con IMP. Il proprietario sarebbe libero di modificare questa impostazione, ma solo su richiesta e dopo la partecipazione a una sessione di studio sulle implicazioni morali delle due opzioni.

Una politica etica libertaria “non ristretta” abilita invece un repertorio più vasto di scelte libere, sempre che sensoristica e software ne rendano possibile l’implementazione. I passeggeri del VA potrebbero essere liberi di adottare, per esempio, preferenze altruistiche selettive a favore di bambini o di persone anziane. E potrebbero anche scegliere l’impostazione di massimo altruismo, privilegiando la protezione di tutte le persone coinvolte nella collisione che non si trovino a bordo del veicolo. Si farebbe così spazio per la scelta di atti supererogatori. Un atto di supererogazione è un atto di generosità morale, non deriva da obblighi morali che ciascuno è ragionevolmente tenuto a rispettare e comporta costi personali straordinari, fino a contemplare la possibilità dell’autosacrificio.

# Politiche a priorità deontologica

Nel 2016 il ministero dei Trasporti e dell’Infrastruttura digitale della Repubblica federale tedesca ha conferito a una Commissione *ad hoc* il compito di redigere un codice etico per la guida autonoma, pubblicato nel giugno 2017 e formato da 20 linee guida (Ethik-Kommission, 2017).

In generale, si afferma che «lo scopo primario dei sistemi di trasporto parzialmente o totalmente automatizzati è di migliorare la sicurezza per tutti gli utenti della strada»; e, inoltre, che «l’obiettivo è quello di ridurre il livello dei danni fino ad azzerarlo completamente» (Ethik-Kommission, 2017, trad. degli autori). Più nello specifico, la linea guida n. 9 delinea una delle possibili politiche etiche per risolvere il problema dei conflitti morali posti dalle collisioni inevitabili. Essa dà priorità a un insieme ristretto di obblighi inderogabili, con valutazioni ispirate all’etica delle conseguenze che subentrano nei casi in cui gli obblighi deontologici non si applicano.

Prendiamone in considerazione i punti di più immediata rilevanza etica:

1. in caso di situazioni di collisione inevitabile, ogni distinzione basata su caratteristiche individuali (età, genere, costituzione fisica o mentale) è strettamente proibita;

2. le parti coinvolte nella generazione di rischi da mobilità non devono sacrificare altre parti non coinvolte;

3. è ammissibile una programmazione informatica generale che sia volta a ridurre i danni alle persone.

Regole morali riconducibili all’etica dei doveri sono enunciate ai punti 1 e 2, mentre il punto 3 ammette una giustificazione riconducibile all’etica delle conseguenze.

Il punto 1 specializza al contesto dei VA il principio di non discriminazione. Pertanto, se il VA ha la possibilità di scegliere chi mettere maggiormente a rischio, non potrà basare questa scelta sul fatto che le persone coinvolte abbiano la pelle di diverso colore oppure siano di differente età. Si esclude così di poter implementare una politica di controllo che porti sistematicamente a sacrificare il più anziano di due pedoni in una collisione inevitabile con uno dei due. Il principio di non discriminazione sembra imporre, a parità di altre condizioni, una scelta casuale tra le vittime potenziali.

L’interpretazione del punto 2 dipende da che cosa si intende per “parte non coinvolta nella generazione di rischi da mobilità”. È ragionevole supporre che i passeggeri di un VA siano sempre coinvolti. Perciò, la loro protezione non può considerarsi una priorità incondizionata, ma solo subordinata alla salvaguardia di coloro che non contribuiscono a generare la situazione collisione. La posizione di un astante è dunque diversa da quella dei due ciclisti evocati in S2: pur non essendo passeggeri del VA, le azioni dei due ciclisti possono aver contribuito a generare la collisione inevitabile. Il punto 2 esclude in particolare che si possano coinvolgere degli astanti allo scopo di proteggere i passeggeri del VA, anche quando i passeggeri sono in numero maggiore.

Veniamo infine al punto 3, in base al quale è ammissibile “una programmazione informatica generale volta a ridurre i danni alle persone”. L’attributo “generale” si deve intendere – pena un conflitto con le richieste ai punti 1 e 2 – come una limitazione: una programmazione volta a ridurre i danni non può scavalcare gli obblighi 1 e 2. Ecco allora l’elenco delle priorità: il principio di esenzione dei non coinvolti prevale su tutto; e il principio di non discriminazione tra le persone coinvolte prevale su quello di riduzione dei danni.

La Commissione etica non ha avanzato pretese di completezza a proposito della linea guida n. 9. In particolare, non è stata esplicitamente esclusa la possibilità di aggiungere altri obblighi deontologici ai quali ottemperare in circostanze che non siano previste dalle regole 1 e 2, con l’effetto di restringere ulteriormente il campo di applicazione della regola 3, volta alla minimizzazione dei danni attesi. In conclusione, la linea guida n. 9 del Codice etico è una proposta per dare un ordine di priorità a principi etici eterogenei, che nel contesto dei VA generano tensioni se non addirittura conflitti morali. Questa linea guida tiene conto sia delle istanze dell’etica dei doveri, sia di quelle dell’etica delle conseguenze, specifica alcune priorità inderogabili per obblighi riconducibili all’etica dei doveri, ma non esclude esplicitamente la possibilità di ampliare la lista delle regole proposte in fasi successive dello sviluppo tecnologico o alla luce di nuove sensibilità morali. Il caso del codice etico tedesco mostra come sia possibile attingere a diverse prospettive etiche per elaborare una politica di sviluppo e uso responsabile dei VA, ma lascia anche intravedere numerose difficoltà, tanto tecniche quanto morali e sociali, che richiedono di essere adeguatamente affrontate.

# Senso e limiti dell’analisi morale

Per quanto, come abbiamo visto, la discussione etica possa (e, in un certo senso, debba) informare efficacemente sforzi legislativi o di regolamentazione, sarebbe un errore pensare che grazie agli strumenti dell’etica normativa ogni problema possa trovare facile e veloce soluzione. Più comunemente, questi strumenti permettono di comprendere meglio le difficoltà da affrontare ma non sembrano rivelare alcuna soluzione definitiva. Ciò vale tanto per l’etica delle conseguenze quanto per l’etica dei doveri, i cui limiti sono stati passati in rassegna.

L’etica dei doveri e l’etica delle conseguenze sono le teorie a cui più spesso ci si rivolge per discutere i problemi sollevati dai sistemi autonomi e intelligenti, ma non sono le uniche disponibili. Altre prospettive etiche possono diventare qui rilevanti. Ad esempio, l’etica della virtù (Hursthouse, Pettigrove, 2018) si concentra più sulle nostre abitudini e sulla formazione del carattere morale che sui singoli atti. Ancora, l’etica della cura (Biller-Andorno, 2012) pone l’attenzione sull’altro, sui bisogni e le necessità di chi ci sta vicino. Molteplici punti di vista possono essere assunti per gettare nuova luce sui problemi morali che dobbiamo affrontare. Ma l’analisi etica tenderà più verso nuove domande che verso risposte definitive. Ciò non significa che la riflessione etica sia inutile. Al contrario, si tratta del migliore strumento a nostra disposizione per prendere decisioni moralmente rilevanti con cognizione di causa.

*Parte seconda*

Persone e istituzioni

Capitolo 5

Governance algoritmica: sorveglianza, censura e diritti fondamentali

Francesca Musiani

1. Introduzione

La “governance algoritmica” (*algorithmic governance* o *governance by algorithms* in inglese) è un concetto emerso essenzialmente nell’ultimo decennio per esprimere l’idea che gli algoritmi siano ormai, per chi li sviluppa, li realizza e li rende operativi in particolari contesti, un importante mezzo di regolamentazione e di esercizio del potere nel nostro mondo digitale. In parallelo alla gestione politica *degli* algoritmi (“governance *of* algorithms”) emerge un insieme di operazioni e decisioni a valenza politica che vengono messe in atto *tramite* algoritmi – usando gli algoritmi come strumenti di potere (“governance *by* algorithms”). La governance algoritmica è una parte, sempre più importante, dell’insieme di questioni definite come “governance di Internet”, ossia l’insieme delle operazioni di progettazione, coordinamento, controllo e manutenzione, svolte da diversi attori pubblici e privati, che fanno funzionare Internet e regolano le interazioni che possono avvenire in rete. In tale ambito, l’uso del termine *governance*[[31]](#footnote-31) rispetto a quello di *governo* intende mettere in rilievo che, per quanto riguarda Internet, la regolamentazione, l’esercizio del potere e la costruzione di norme non sono prerogativa esclusiva delle istituzioni, ma di una molteplicità di attori che include gli stessi sviluppatori informatici, l’industria e la società civile.

Le questioni legate alla governance algoritmica, pur essendo relativamente recenti, si rifanno ad un’idea da lungo tempo fondamentale negli studi sociali della tecnologia – che le tecnologie strutturino il sociale in modi vari e specifici. Le ricerche attuali su cosa costituisca la governance algoritmica si concentrano sul ruolo degli algoritmi e sui loro effetti di ordinamento e strutturazione dei mondi sociali. La governance algoritmica ha acquisito un ruolo sempre più importante in campi come l’amministrazione pubblica, i media e i servizi online, il *digital labor* e le libertà digitali dal punto di vista sia della loro protezione sia degli abusi di cui possono essere oggetto.

Il presente capitolo si propone di fornire un’introduzione al concetto di governance algoritmica[[32]](#footnote-32) ed in particolare agli aspetti che influenzano la sorveglianza, la censura e la protezione delle libertà e dei diritti fondamentali su Internet. Il capitolo include dapprima una panoramica di situazioni in cui la governance algoritmica è “all’opera” in diversi contesti (motori di ricerca Web, raccomandazioni nell’e-commerce, moderazione automatizzata dei contenuti nelle reti sociali, polizia predittiva); in seguito, discute di come autori e ricercatori provenienti da differenti orizzonti disciplinari hanno affrontato e affrontano la questione della governance algoritmica; infine, delinea diversi tipi di governance algoritmica e i loro effetti su questioni come la sorveglianza, la vita privata, l’autonomia, la trasparenza e l’equità.

1. Dove si osserva la governance algoritmica?

L’importanza di riflettere sulla governance algoritmica da diverse prospettive diventa evidente quando si esamina ciò che cambia nei meccanismi di potere quando esso fa uso di algoritmi: narrative di meccanismi di governance sempre più potenti, invadenti e pervasivi si mescolano a narrative di governance potenzialmente più inclusiva, reattiva e foriera di maggiore diversità sociale. Per rendere più tangibile quali sono alcuni dei contesti in cui si pone la questione della governance algoritmica, la prima parte di questo capitolo presenta brevemente degli esempi che illustrano algoritmi “all’opera”, sia nel settore pubblico che quello privato – con l’obiettivo di mostrare come la governance algoritmica può assumere molteplici forme, ed è invariabilmente controversa, fortemente dipendente dai suoi contesti di implementazione e plasmata da una varietà di interessi, fenomeni di potere e resistenza.

* 1. Motori di ricerca sul Web

Le modalità in cui il Web dà più visibilità ad alcune informazioni e contenuti rispetto ad altre sono al centro del dibattito ricorrente sulle caratteristiche che definiscono lo spazio digitale come *pubblico*. Secondo Jürgen Habermas, per strutturare uno spazio pubblico sono necessarie due condizioni: la libertà di espressione e la discussione come forza di integrazione. L’architettura di Internet sembra articolare queste due condizioni. Tuttavia, se la prima è spesso riconosciuta come una delle sue virtù diffuse, la seconda sembra allo stato attuale assai più incerta (Cardon, 2013). I motori di ricerca ed i molteplici meccanismi alla base del Web gerarchizzano infatti la visibilità delle informazioni, proponendole all’inizio delle liste dei risultati ottenuti tramite una ricerca, oppure relegandole alla fine di tali liste.

In particolare, grazie all’attuale posizione di predominio che Google detiene sulle pratiche di ricerca sul Web, il suo algoritmo *PageRank* è stato ampiamente esaminato come nuovo *gatekeeper* (Smith, 2013) e “dittatore benevolo” degli spazi pubblici digitali. L’algoritmo implementa, secondo una modalità che rimane in parte un segreto industriale (e che si è evoluta attraverso gli anni), diversi criteri di misura che valutano l’autorità di un’informazione (in base al numero di citazioni), la sua audience (in base al numero di visite o click), la vicinanza e l’affinità all’utente (in base alle raccomandazioni) o la velocità (in base all’aggregazione in tempo reale e all’inoltro di argomenti d’attualità). *PageRank*, come “*master switch*” di Internet (Wu, 2010), centralizza e organizza la circolazione delle informazioni nella rete delle reti, decidendo in sostanza ciò che è più o meno rilevante per ogni interrogazione e richiesta di ricerca. Poiché i motori di ricerca sono diventati essenziali per determinate esigenze della società, come la libera circolazione e l’accesso a informazioni e idee, è fondamentale discutere in che modo gli attuali motori di ricerca stiano effettivamente soddisfacendo queste esigenze e come la promozione di eventuali politiche relative al pluralismo e alla diversità di prospettive dovrebbero tenere conto di questi aspetti in misura maggiore.

* 1. Raccomandazioni nell’e-commerce

Da alcuni anni, il venditore online Amazon è un “suggeritore eccellente”: le sue prescrizioni si basano infatti sulle numerose raccomandazioni dei suoi acquirenti. Nel caso dell’acquisto di libri il sito del venditore permette a ciascuno degli utenti iscritti di conoscere, con un solo clic, gli altri acquisti effettuati in passato da utenti che hanno acquistato lo stesso titolo. È sulla sistematizzazione e sull’automazione di un fenomeno molto diffuso e molto sociale – lo scambio di consigli e indicazioni, la condivisione di preferenze e affinità tra utenti – che Amazon e altri venditori online basano i loro sistemi di raccomandazione. Attingendo a metodi basati sia sui contenuti (per esempio classificando come simili due libri che condividono un certo numero di parole) sia sul filtraggio collaborativo (l’intersezione di liste contenenti libri particolari e liste con precedenti registrazioni di libri acquistati o presi in prestito dai lettori), Amazon ha sviluppato un algoritmo chiamato “filtro collaborativo *item-to-item*”. I suoi dettagli rimangono un segreto industriale, ma l’algoritmo mostra ogni giorno a milioni di individui la sua efficacia nel personalizzare le raccomandazioni secondo gli interessi di ciascuno dei suoi clienti. Come suggerisce il nome, questo algoritmo mette in relazione ogni articolo ordinato e acquistato dagli utenti con articoli simili e alla fine li combina in una lista di raccomandazioni.

Dietro a questo algoritmo si celano anni di ricerche ed esperimenti in un campo dell’informatica le cui applicazioni pratiche sono sempre più diffuse: il *data mining e,* in particolare, l’*affinity analysis* e la *market basket analysis*. Per i lettori alla ricerca di nuove letture vengono costruiti suggerimenti di testi simili agli articoli da loro precedentemente acquistati sulla base di diverse fonti di informazioni incrociate, che vanno ad alimentare un ampio database dove vengono combinate con altre storie di acquisti. Queste informazioni possono andare dai dati demografici relativi al cliente e ai suoi cari fino a valutazioni più complesse basate sui siti che si sono consultati prima di arrivare su quello di Amazon. I collegamenti all’interno di questo esteso database sul comportamento d’acquisto degli utenti, attivati secondo l’algoritmo brevettato di Amazon, sono alla base dei suggerimenti ormai familiari a tutti coloro che utilizzano Amazon, come “Consigliato poiché hai comprato...” o “Consigliato perché aggiungi X alla tua lista di desideri...”. Questi suggerimenti influenzano ogni giorno migliaia di acquisti su Amazon.

* 1. Reti sociali e moderazione automatizzata dei contenuti

Un altro campo rilevante e controverso della governance algoritmica “all’opera” è la moderazione (parzialmente) automatizzata e la regolamentazione dei contenuti sulle piattaforme di social media. In questo caso l’obiettivo è di controllare che i contenuti siano conformi alla legge e alle specifiche regole della piattaforma. La grandissima e crescente quantità di comunicazioni che circolano su queste piattaforme avviene ormai su una scala così ampia che è difficile immaginare che i moderatori umani possano affrontare il loro compito interamente con metodi manuali – anche se, come dimostra Antonio Casilli, il “fattore umano” nella moderazione di contenuti resta importante, anche se poco riconosciuto e quindi poco tutelato (Casilli, 2020).

Nei loro sforzi per trovare soluzioni soddisfacenti a livello globale, le grandi piattaforme dei media sociali si indirizzano verso soluzioni tecniche che hanno un vantaggio economico rilevante. Inoltre, negli ultimi anni si è accresciuta la pressione politica affinché le piattaforme affrontino le questioni dell’incitamento all’odio, della disinformazione e della violazione del copyright sui loro siti. I tentativi di legiferare in questo settore hanno in parte sancito la responsabilità diretta delle piattaforme per questi tipi di contenuti. Le piattaforme di media sociali dunque sviluppano, testano e mettono in funzione un numero crescente di sistemi automatizzati che mirano a identificare e segnalare tali contenuti (Gillespie, 2018). Peraltro, nonostante le piattaforme abbiano a lungo decantato le lodi dell’efficacia ed efficienza degli algoritmi-moderatori, la maggior parte dei sistemi tecnici di moderazione dei contenuti non funziona in modo completamente automatizzato, ma il più delle volte pre-seleziona contenuti potenzialmente problematici che vengono poi esaminati da moderatori umani: secondo stime recenti, 15000 moderatori lavorano per Facebook e 10000 per YouTube (Badouard, 2020).

Piattaforme come Facebook, Twitter e YouTube hanno a lungo mantenuto la segretezza sui criteri decisionali, le tecnologie e i dati alla base delle loro pratiche di moderazione. Il crescente peso politico della questione ha tuttavia spinto le aziende ad aumentare la loro trasparenza al riguardo, anche se per il momento con risultati solo parziali. Recentemente, Facebook, Google, Twitter e Microsoft si sono unite nel *Global Internet Forum to Counter Terrorism* (GIFCT), un’iniziativa per combattere la diffusione di contenuti terroristici online, incentrata su un database condiviso tra queste aziende, ma altrimenti inaccessibile di immagini, video, audio e testi che possono essere ricondotti ad attività terroristiche. Facebook rende ormai pubblica la struttura generale del processo con cui implementa l’attività di moderazione e i relativi criteri decisionali, ma mantiene il segreto sui dettagli del processo e non fornisce dati precisi sui contenuti rimossi. YouTube propone un sistema, chiamato *ContentID*, preposto al blocco o alla monetizzazione dei contenuti protetti da copyright. Anche in questo caso i criteri utilizzati sono conosciuti solo ad alto livello; esistono analisi che dimostrano che il sistema blocca molti contenuti legittimi e segnalano come YouTube manchi ancora di un livello di trasparenza accettabile per quanto riguarda le eventuali procedure di appello.

* 1. Algoritmi predittivi come strumenti di polizia e di inchiesta

Le autorità utilizzano sempre più (per esempio a scopi di sicurezza o di *intelligence*) strumenti di governance algoritmica, che combinano e analizzano diversi insiemi di dati al fine di valutare il rischio di azioni criminose – ad esempio, furti ed aggressioni – e cercare di prevenirle. Le “previsioni di rischio” risultanti dall’analisi algoritmica vengono usate per indirizzare le azioni delle autorità verso particolari aree geografiche o persone specifiche. Queste analisi possono influenzare la decisione da parte delle autorità di assegnare una maggiore presenza della polizia a specifiche aree considerate “a rischio”, di sorvegliare individui identificati come potenziali colpevoli, o persino di avvertire potenziali vittime.

In molti casi, la raccolta e l’analisi dei dati e le misure di governance che ne derivano sono di competenza sia delle agenzie di polizia che di società private, e spesso le relazioni tra questi attori sono opache. Alcuni dei programmi utilizzati restringono le fonti su cui si basano a dati unicamente legati alla criminalità, mentre altri si avvalgono di dati aggiuntivi, come quelli relativi alla meteorologia, al traffico, alle reti, ai consumi e ai comportamenti online. Nella maggior parte dei casi, i software e i principi soggiacenti, così come i dati che usano, non sono noti al pubblico, e lo stesso vale per i risultati dell’analisi e la loro interpretazione. Questo grado relativamente alto di opacità è giustificato dalle autorità con l’argomento che la trasparenza permetterebbe ai criminali di comprendere meglio gli eventuali punti deboli del programma utilizzato e lo renderebbe inefficace. Tuttavia, l’opacità complica enormemente le valutazioni degli effetti sociali di queste pratiche da parte della polizia e dell’*intelligence*; sembra legittimo domandarsi se queste azioni di polizia predittiva rafforzino forme illegittime di discriminazione. Ci si può domandare, d’altra parte, se l’uso di tali programmi sia efficace ed efficiente: uno dei problemi principali di questi software è il fatto di mettere in secondo piano i contesti storici e sociali in cui avvengono determinati crimini. Il software PREDPOL utilizzato negli Stati Uniti, per esempio, combina una serie di variabili spazio-temporali per elaborare la probabilità che un crimine si ripeta presto in un luogo in cui è già avvenuto; per far ciò il software si basa su un modello tratto dalla sismologia, capace di prevedere “l’evento iniziale” di un terremoto molto meno accuratamente rispetto alle scosse di assestamento. Concentrandosi esclusivamente sulla ripetizione del crimine, il software arriverebbe automaticamente alla conclusione che il luogo in questione necessita di maggiore sorveglianza da parte delle forze dell’ordine, senza considerare meccanismi più raffinati che estendano la protezione ad aree e popolazioni particolarmente vulnerabili (Benbouzid, 2017).

È interessante notare che – forse a causa di questi problemi – la maggior parte dei sistemi di polizia predittiva sono ancora concepiti come sistemi di raccomandazione per agenti umani, ai quali rimane la responsabilità di decidere in via definitiva se e come agire. Tuttavia, un numero consistente di ricerche e progetti pilota svolti in questo ambito indica che nel prossimo futuro il grado di automazione di tali sistemi aumenterà considerevolmente. È anche plausibile supporre che essi saranno applicati in vari ambiti, compreso il controllo delle frontiere. Per queste ragioni dovrà essere prestata particolare attenzione al margine di scelta che resterà possibile per gli operatori umani e alla loro capacità di respingere la raccomandazione del programma e prendere una decisione autonoma.

1. Prospettive multi- e inter-disciplinari sulla governance algoritmica

Come mostrano gli esempi discussi nel paragrafo precedente, esiste una molteplicità di modi in cui gli algoritmi contribuiscono a creare l’ordine sociale: la governance algoritmica è una forma di ordinamento sociale che si basa sul coordinamento tra gli attori e regole, e che incorpora procedure epistemiche particolarmente complesse basate su algoritmi per la categorizzazione e il controllo dei comportamenti di individui e gruppi. Nell’ultimo decennio, autori e ricercatori con differenti orizzonti disciplinari hanno analizzato, definito e discusso la nozione di governance algoritmica nei modi che questo paragrafo tenta di riassumere.

* 1. Pensare la governance algoritmica

Le varie discipline che affrontano la questione della governance algoritmica condividono concezioni simili rispetto all’importanza degli algoritmi per l’ordinamento sociale, anche se si concentrano su diversi oggetti di ricerca. Tali discipline includono gli studi sulla scienza e la tecnologia o *science and technology studies* (STS), la sociologia, le scienze politiche, le scienze della comunicazione, l’informatica, gli studi giuridici, l’economia e la filosofia. Se Antoinette Rouvroy e Thomas Berns avevano usato il termine “*gouvernance algorithmique*” già nel 2010, contributi importanti per la concettualizzazione del termine governance algoritmica sono venuti da Barocas *et al.* (2013) e da Müller-Birn *et al.* (2013), che hanno messo in rilievo il legame tra governance algoritmica e altri tipi di meccanismi di governance, in primo luogo istituzionale e sociale. Introdotto da Tim O'Reilly (2013), il concetto di *algorithmic regulation* (regolazione algoritmica) ha dal canto suo evidenziato la possibile maggiore efficienza degli spazi governati automaticamente, trascurando però in larga parte le controversie che possono derivare dalla delega di potere a soluzioni tecnologiche (si veda per esempio Morozov, 2014). Studiosi di scienze giuridiche, scienze politiche, economia e sociologia hanno arricchito il dibattito collegando questa nozione al campo di studi sulla regolamentazione e la governance in generale (Just, Latzer, 2016; Yeung, 2018).

Le basi del concetto di governance algoritmica possono peraltro essere ricondotte all’ipotesi di fondo che ispira gli STS, ossia che la tecnologia ha sempre rispecchiato e riorganizzato il sociale (Bijker, Law, 1992). Dalla “politica degli artefatti” di Langdon Winner (1980) al famoso slogan di Lawrence Lessig, *code is law*, l’idea che gli artefatti tecnologici contribuiscano in diversi modi a governare la società e le interazioni sociali è un tema ricorrente. Inoltre, questioni come: il ruolo dell’invisibilità nei processi di classificazione che ordinano l’interazione umana; le procedure attraverso cui le categorie sono rese invisibili e mantenute tali; i modi in cui gli individui possono modificare questa invisibilità quando necessario; e la misura in cui i sistemi di classificazione sono cruciali per la costruzione delle infrastrutture informazionali sono state per diversi anni tra le preoccupazioni principali degli studiosi di STS (Bowker, Star, 1999). Eppure, la questione della classificazione e dell’organizzazione dell’informazione non è forse mai stata così rilevante come nella nostra attuale epoca di “sovraccarico di informazioni” e di accesso mediato da Internet alla gran parte delle informazioni che ci circondano (Cardon, 2013). In effetti, i dati digitali sembrano proliferare nel complesso mondo di oggi, basandosi sulla varietà di piattaforme e supporti che consentono la loro de-materializzazione, rapida circolazione e distribuzione. I dati servono oggi a scopi diversi, dal *trading* alla sorveglianza, dalla valutazione alla raccomandazione; inoltre sono raggruppati e organizzati attraverso supporti e dispositivi tecnologici che vanno dai motori di ricerca ai siti di e-commerce. Se dal canto loro, una pluralità di attori del settore privato mette a profitto le tracce lasciate dai consumatori sul web per meglio indirizzare, personalizzare (e sfruttare) i loro prossimi acquisti e interazioni, numerosi utenti si preoccupano dei ritratti che tali tracce dipingono di loro, nonché dell’impossibilità di modificarle o cancellarle. Già agli albori dell’era dei big data e degli algoritmi, diversi autori hanno sostenuto che queste dinamiche danno «un’importanza decisiva non solo ai proprietari dei dati, ma anche e soprattutto a coloro che possono renderli intelligibili» (Cardon, 2013, mia trad., p. 10). Gli algoritmi che sottendono le tecnologie dell'informazione e della comunicazione che utilizziamo quotidianamente, Internet in primo luogo, sono (anche) artefatti di governance, accordi di potere e “politica (condotta) con altri mezzi” (Latour, 1988). Basandosi su questa tradizione STS, il campo interdisciplinare degli studi critici sul software e sul codice informatico (*software studies* e *critical code studies*) ha messo in luce le intricate dipendenze reciproche di software e algoritmi da un lato, e le interazioni e le strutture sociali dall’altro (Fuller, 2008; Berry, 2011).

Diversi lavori di ricerca, tra loro strettamente correlati, si interessano oggi di come gli algoritmi contribuiscono a riorganizzare e a riconfigurare le interazioni e le strutture sociali, rifiutando tuttavia la nozione tecno-determinista che vede gli algoritmi come forze esterne indipendenti che governano il nostro mondo[[33]](#footnote-33). Tali ricerche affermano la rilevanza degli algoritmi in questi contesti, ma evidenziano al contempo i contesti economici, culturali e politici che danno forma alla progettazione degli algoritmi e ne permettono il funzionamento (Christin, 2020); in alcuni casi, tentano di stabilire una tipologia di fattori che permettono di identificare e distinguere tra loro diversi modelli di governance algoritmica (Katzenbach, Ulbricht, 2019). Gli studi empirici in questo campo si concentrano sul ruolo degli algoritmi e sul loro effetto di “ordinamento” in specifici contesti sociali (Kitchin, 2016; Ziewitz, 2016), analizzando i modi in cui le basi di dati, i modelli matematici e le procedure di calcolo facilitano nuove modalità di classificazione sociale basata su processi algoritmici. Questo insieme di studi ha contribuito a organizzare, stimolare e sistematizzare la ricerca sulla governance algoritmica e ha reso più chiara la definizione di governance come complesso di «tentativi intenzionali di gestire il rischio o di alterare i comportamenti al fine di raggiungere un obiettivo prestabilito» (Yeung, 2018, p. 507, trad. mia). L’attenzione dedicata agli interventi intenzionali, diretti a raggiungere un particolare obiettivo, costituisce la base di ricerche esplicitamente interessate agli algoritmi come forma di governo appositamente impiegata per regolare i contesti sociali e alterare il comportamento degli individui – in quanto cittadini, lavoratori, consumatori.

La ricerca che analizza le implicazioni sociali e politiche degli algoritmi si sviluppa in parallelo agli studi, principalmente nel campo dell’informatica (si veda ad esempio Gürses *et al*., 2018), che mirano a costruire e ottimizzare sistemi algoritmici per risolvere specifici problemi sociali, individuando contenuti controversi, comportamenti devianti, preferenze o opinioni. La maggior parte di questi studi ha due obiettivi principali. Il primo è di ottimizzare l’individuazione, a partire da dati esistenti, di modelli efficienti, robusti ed equi per classificare soggetti e oggetti in categorie generali o in dimensioni specifiche, come gli stati emotivi o le preferenze politiche. Il secondo obiettivo è di ottimizzare particolari servizi basati su algoritmi per stimolare nel modo più efficace il comportamento degli utenti, cercando di massimizzare i benefici, organizzativi o o di altro tipo, che ne derivano (si veda ad esempio Gürses *et al*., 2018). In alcuni casi, questo tipo di studi si sviluppa non solo in parallelo, ma in dialogo con gli studi che si concentrano sugli effetti sociali e di governance degli algoritmi. Alcune ricerche applicate e critiche al tempo stesso si sforzano di raggiungere entrambi gli obiettivi, come alcuni studi sull’equità o la parzialità degli algoritmi, che cercano anche di concettualizzare specifiche misure e di testarle (si veda ad esempio Waseem, Hovy, 2016), o gli studi sul “*bias* dell’automazione” che analizzano le condizioni in cui gli esseri umani possono prendere una decisione *veramente* autonoma in un contesto algoritmico (si veda ad esempio Parasuraman, Manzey, 2010).

* 1. Studiare le applicazioni della governance algoritmica

Le ricerche sulla governance algoritmica si sono concentrate su diversi settori di applicazione degli algoritmi, di cui il paragrafo 2 ha dato un’anteprima orientata ad alcuni contesti pratici. Un settore d’indagine particolarmente rilevante è senz’altro quello della comunicazione digitale e dei social media. Ormai da quasi due decenni, un insieme di lavori multidisciplinari esplora come i motori di ricerca e le piattaforme dei social media organizzano e strutturano le informazioni disponibili online e come questo influisce sulle percezioni e motivazioni degli utenti. I servizi basati sul Web danno tipicamente priorità ad alcuni tipi di contenuti, spesso basati su misure dell’interesse dimostrato dall’utente per contenuti simili in precedenza, creando così una nuova modalità dominante per attribuire rilevanza sociale a specifiche informazioni. Come già discusso, tali servizi utilizzano algoritmi anche a fini di regolazione di contenuti, bloccando o filtrando discorsi, video o foto ritenuti inaccettabili o illegali (Gillespie, 2018). Con l’aumento della popolarità di tali servizi, ed una crescente pressione politica, i loro creatori e gestori non esitano a ricorrere a soluzioni tecniche e automatizzate per risolvere difficili problemi di governance, come l’incitamento all’odio, la disinformazione e la protezione del diritto d’autore (Gorwa *et al*., 2019; Badouard, 2020). Altre aree che si avvalgono di pratiche di individuazione automatica dei contenuti sono per esempio l’analisi delle percezioni e motivazioni (*sentiment analysis*) per il marketing commerciale e politico e l’analisi del plagio nell’insegnamento e nella ricerca accademica.

Un’altra area di interesse per gli studiosi di governance algoritmica è quello dei servizi, della gestione e della sorveglianza dei cittadini da parte delle istituzioni pubbliche e del settore privato. Gli studiosi di scienze politiche e gli studiosi di giurisprudenza, in particolare, hanno intrapreso ricerche sull’automatizzazione delle procedure per la fornitura di servizi pubblici e il supporto ai processi decisionali amministrativi. L’ambizione di questi attori è che gli algoritmi possano aumentare l’efficienza e l’efficacia dei servizi pubblici, ad esempio razionalizzando la burocrazia, migliorando la ricerca di informazioni o la scelta delle migliori opzioni disponibili; tuttavia, i risultati di queste promesse restano fortemente controversi. I progetti di servizi algoritmici si basano su di una varietà di sistemi di classificazione, punteggio e previsione del comportamento, delle preferenze e delle opinioni dei cittadini, che sono utilizzati per l’assegnazione di prestazioni sociali, per combattere l’evasione e la frode fiscale, per informare la polizia e le agenzie di prevenzione del terrorismo, o ancora per esercitare un controllo delle frontiere o una gestione dei flussi migratori più efficace. Le ricerche in questo campo hanno dimostrato che la diffusione di sistemi algoritmici nel settore pubblico ha prodotto molte conseguenze non solo indesiderate, ma che restano largamente opache per il cittadino (Dencik *et al*., 2018), e che l’applicazione di strumenti di tipo algoritmico nell’amministrazione pubblica è sovente strettamente legata a nuove forme di sorveglianza e classificazione della popolazione da parte di attori statali e privati (Bennett, 2017). Questi dispositivi di valutazione e attribuzione di punteggi non si applicano peraltro agli individui solo in quanto cittadini, ma anche in quanto consumatori, come hanno evidenziato gli studi dei sistemi di punteggi di credito (*credit scoring*). Questi studi evidenziano come le pratiche di valutazione basate su algoritmi influenzano i mercati e creano meccanismi di stratificazione che possono riconfigurare e rinforzare i rapporti di potere, spesso a scapito dei settori già più fragili e impoveriti della società.

Alcuni studiosi sottolineano il potenziale della governance algoritmica di affrontare un alto grado di complessità, grazie alla capacità degli algoritmi di elaborare un numero anche molto elevato di input; in questi casi la governance algoritmica è riconosciuta come una modalità di coordinazione che può potenzialmente offrire nuove opportunità di partecipazione, inclusione sociale, diversità e reattività democratica, con l’obiettivo esplicito di “udire un maggior numero di voci” e di migliorare il rapporto tra utenti/cittadini, piattaforme, ed élite politiche. Questo insieme di ricerche esplora le possibilità offerte dai sistemi algoritmici per migliorare la partecipazione politica attraverso strumenti di partecipazione online, come le elezioni e le petizioni elettroniche, la comunicazione sui social media e il *crowdsourcing* legislativo (si veda per esempio König, 2019), mettendo comunque in rilievo che la governance algoritmica attraverso strumenti partecipativi resta spesso gerarchica e riproduce una distribuzione del potere diseguale (Kelty, 2017).

Infine, la governance algoritmica è di notevole interesse per gli studiosi, in particolare sociologi ed economisti, che studiano le trasformazioni digitali del lavoro. Tali studi si concentrano sui modi in cui viene messa in opera l’automazione della gestione del lavoro all’interno delle aziende, ad esempio attraverso sistemi di *performance management* e di *rating* (Rosenblat, 2018), o la gestione e la sorveglianza algoritmiche dei luoghi di lavoro. Legati all’interazione tra algoritmi e lavoro sono gli studi che vertono sui sistemi algoritmici di realtà aumentata, di riconoscimento vocale e di assistenza per l’esecuzione di particolari compiti, o per il controllo di qualità; le ricerche che analizzano la gestione algoritmica dei trasporti, del traffico, dell’energia, dei rifiuti e della fornitura d’acqua (soprattutto nell’ambito di progetti detti di *smart city*); e le ricerche sul “nuovo capitalismo” delle piattaforme del Web, che esaminano il “fattore umano” di un lavoro sempre più frammentato, ripetitivo e automatizzato (Casilli, 2020).

1. Governance algoritmica, valori e diritti

Come già in parte evidenziato nel precedente paragrafo, un numero importante di controversie e di questioni aperte ricorrono regolarmente nelle discussioni – tra specialisti e nel dibattito pubblico – del fenomeno della governance algoritmica – enfatizzando secondo i casi le sue connotazioni positive o negative. La ragione principale è che questo fenomeno può profondamente influenzare, a più livelli, la protezione delle libertà e dei diritti fondamentali su Internet, a causa dei potenziali problemi che sorgono dall’eccessiva sorveglianza, dalla censura, dai pregiudizi e dal rinforzo delle disuguaglianze dei sistemi di governance algoritmica. Nei prossimi paragrafi si fornisce una breve presentazione delle questioni attorno a cui si articolano le controversie sulla governance algoritmica.

* 1. Predizione, sorveglianza e vita privata

I big data, gli algoritmi e l’intelligenza artificiale hanno il potenziale di modificare in profondità il modo in cui le società percepiscono le proprie popolazioni e “classi sociali”, e gli individui che le compongono. Ciò è dovuto a quella che José van Dijck (2014) ha chiamato *datafication*, l’aumento e la varietà dei dati raccolti dai dispositivi digitali, dai *tracker* online e dai sistemi di sorveglianza. Questo fenomeno beneficia di infrastrutture tecniche sempre più potenti che consentono un’analisi dei dati molto rapida e di norme sociali che favoriscono la quantificazione, la classificazione e la sorveglianza (Rieder, Simon, 2016). Le ricerche sulla governance algoritmica si focalizzano, dunque, in modo importante sui numerosi rischi della *datafication* e della sorveglianza. Sorvegliare intere popolazioni e creare profili dettagliati degli individui sulla base dei loro *data doubles* (duplicati digitali ricomposti a partire da aggregati di dati e diffusi attraverso “assemblaggi” di sistemi informativi; si veda Haggerty, Erickson, 2000) crea ampie opportunità per la classificazione sociale, la discriminazione, l’oppressione e la manipolazione dei consumatori e dei cittadini (Lyon, 2014). La sorveglianza illimitata rappresenta un pericolo per le libertà fondamentali e i diritti umani, come la libertà di parola, la libertà di riunione e la protezione della vita privata.

Molti attori del mondodigitale, in particolare le principali piattaforme, ricorrono sempre più ad algoritmi predittivi, con due obiettivi principali: personalizzare il servizio fornito ai propri clienti o fornire elementi utili al processo decisionale. Questi dispositivi possono tendere a privare gli individui di un'alternativa decisionale e a ridurre le loro possibilità di scelta, non essendo più semplici meccanismi di supporto alle decisioni, ma veri e propri sistemi di decisione automatici o semiautomatici. Come abbiamo visto, gli algoritmi alla base di questi sistemi possono avere un’influenza sui diritti dei consumatori e sul rapporto dell'utente/cittadino con le autorità pubbliche. Parallelamente alla diffusione su larga scala di tali algoritmi da parte del settore privato, sta anche emergendo una vera e propria “azione pubblica algoritmica” (Grosdhomme-Lulin, 2015), destinata, ad esempio, ad anticipare alcuni comportamenti in termini di sicurezza e più in generale in campo sociale, come la previsione dei rischi di abuso, l’assistenza alla diagnosi medica, l’anticipazione dei rischi di abbandono scolastico.

Si pone quindi la questione dell’integrazione dei diritti di informazione, accesso e opposizione, già oggi riconosciuti, attraverso una maggiore supervisione degli algoritmi utilizzati nell’elaborazione dei dati; in particolare è necessario che i processi automatizzati non si sostituiscano all’intervento umano, ma costituiscano un’assistenza alla decisione presa, in ultima analisi, da un individuo o da un gruppo di individui. Il regolamento europeo sulla protezione dei dati personali (RGPD) già vieta che una decisione che produce effetti giuridici nei confronti di una persona sia presa esclusivamente sulla base di un trattamento automatizzato dei dati destinato a definire il profilo dell’interessato o a valutare determinati aspetti della sua personalità (articoli 4 e 22). Tuttavia, nell’epoca della governance algoritmica, questo divieto sembra essere *de facto* sempre meno efficace in considerazione dell’importante asimmetria informativa e decisionale tra progettisti, sviluppatori e gestori delle piattaforme, e gli utenti soggetti dell’azione algoritmica (Rouvroy, Berns, 2010). Sono quindi previste nuove regole a livello nazionale e internazionale per disciplinare il funzionamento degli algoritmi predittivi in materia di protezione dei dati e in relazione al diritto di opporsi alla profilazione (l’insieme delle attività di raccolta ed elaborazione dei dati relativi agli utenti di un servizio, per suddividerli in gruppo a seconda del loro comportamento). Tale pratica infatti può portare a misure che producono effetti giuridici per l’individuo o che incidono in modo significativo sulle sue libertà, interessi e diritti. Queste nuove regole comprendono: l’obbligo di completare le decisioni prese per mezzo di un algoritmo con un intervento umano; l’obbligo di trasparenza sui tipi di dati personali utilizzati dall'algoritmo e sulle modalità della sua parametrizzazione, consentendo, se necessario, verifiche che permettano di contestare la logica generale o la veridicità dei dati analizzati; il divieto di utilizzare algoritmi che direttamente o indirettamente portino a discriminazioni basate, tra l’altro, sulla razza, l’origine etnica, le opinioni politiche, la religione, le convinzioni personali, l’orientamento sessuale o l’identità di genere o l’appartenenza a un sindacato.

* 1. Autonomia e trasparenza

L’uso crescente degli algoritmi come strumenti di governance preoccupa anche per quanto riguarda i suoi effetti sull’autonomia dell’individuo e sulla sua *agency* (possibilità/capacità di agire) (Hildebrandt, 2016). Inizialmente, gli algoritmi erano considerati degli strumenti per aumentare o diminuire l’autonomia dell’individuo. Mano a mano che le ricerche interdisciplinari sulla governance algoritmica si sono sviluppate, è diventato però chiaro che l’interazione tra agenti umani e agenti artificiali (come macchine, robot, algoritmi) è più complessa e articolata di quanto questa visione permetta di comprendere, e necessita di una maggiore differenziazione: le realtà dell’interazione umani-macchine sono molteplici (Gray, Suri, 2018), e la capacità di agire dell’essere umano non può essere valutata solo rispetto alle macchine, ma anche rispetto all’insieme di opportunità e di vincoli posti dalle organizzazioni e dalle norme sociali (Caplan, boyd, 2018).

L’opacità degli algoritmi è ormai da tempo una delle questioni centrali nelle controversie relative alla governance algoritmica e si accompagna generalmente a un appello a una maggiore trasparenza (Pasquale, 2015). La trasparenza viene riconosciuta come un fattore importante per l’autodeterminazione dell’individuo, ma la complessità intrinseca dei sistemi algoritmici pone una sfida di grande importanza a tale riguardo. Pertanto, la trasparenza è una delle principali questioni trattate in ambito accademico e nelle arene in cui si delibera riguardo alle modalità della governance algoritmica, ma anche un tema controverso nello spazio pubblico (Hansen, Flyverbom, 2015). Una certa misura di trasparenza sembra necessaria per una migliore comprensione dei sistemi decisionali e per rendere le norme implicite che regolano questi sistemi più aperte al controllo, alla deliberazione e al cambiamento.

Alcuni studi recenti sottolineano, tuttavia, che la trasparenza assoluta spesso non è possibile né auspicabile e può non essere la soluzione per la maggior parte dei problemi legati alla governance algoritmica, come la mancanza di equità o la manipolazione (si veda per esempio Ananny, Crawford, 2017). In alcuni casi, il fatto di incorporare specifiche norme sociali e determinati valori nel codice informatico, invece di creare più opacità, può contribuire a rendere espliciti norme e processi che prima erano nascosti, come hanno rivelato diverse controversie sui sistemi di valutazione del rischio di disoccupazione, del merito creditizio, o dell’orientamento nell’istruzione superiore. Infatti, l’interesse pubblico per la governance algoritmica ha motivato gli attori della società civile e gli studiosi a indagare la composizione e le ragioni sottostanti ai “punteggi” stabiliti tramite algoritmi e a mettere in discussione i valori sociali sottostanti. Alla luce di questo sviluppo, l’attuale evoluzione della governance algoritmica potrebbe addirittura favorire una maggiore trasparenza; le procedure dette di *auditing* (il concetto riprende la terminologia relativa agli obblighi delle aziende, che prima di comunicare i loro risultati finanziari, si sottopongono ad una approfondita revisione da parte di società esterne all’organizzazione) rendendo pubblico il codice di specifici sistemi, possono contribuire in maniera sostanziale a un’esplicitazione delle ipotesi e dei valori sottostanti.

* 1. Parzialità, equità e censura

Già nei primi tempi dell’ascesa della governance algoritmica, uno dei rischi che gli studiosi hanno identificato nell’uso degli algoritmi come strumenti di governance è stato l’aura di obiettività, neutralità e verità che li accompagnava, così come la loro promessa di risolvere problemi di importante complessità sociale attraverso una “semplice” disponibilità di grandi moli di dati e, in parallelo, un aumento esponenziale della potenza di calcolo (si veda per esempio Boyd, Crawford, 2012). Negli anni seguenti, molti studi hanno contestato l’idea che gli algoritmi possano essere oggettivi e neutrali, mostrando come le disuguaglianze sociali, la mancanza di equità e la discriminazione si riflettono in distorsioni presenti negli stessi dati e nelle pratiche di aggregazione e profilazione ad essi relative. La presa di coscienza su larga scala delle implicazioni sociali degli algoritmi ha portato a numerose discussioni sul rapporto di questi meccanismi con i sistemi di valori sottostanti le campagne politiche, con le disuguaglianze sociali nella fornitura di servizi pubblici e con le implicazioni della sorveglianza effettuata dai datori di lavoro per i diritti civili degli impiegati. La governance algoritmica sta, in questo senso, contribuendo a suscitare discussioni e a esplicitare i modi in cui vengono messi in opera i processi decisionali in numerose aree della società. Alcuni studiosi, tuttavia, sottolineano che la governance algoritmica sta diventando un’“infrastruttura sociale” del mondo contemporaneo e, come tale, verrà profondamente integrata nella nostra vita quotidiana, per infine essere data per scontata come quasi tutte le infrastrutture una volta che esse siano stabilizzate (Gorwa *et al*., 2019).

Un’altra preoccupazione fondamentale, strettamente legata alla precedente, è quella del pregiudizio o distorsione degli algoritmi. L’automatizzazione del processo decisionale tramite sistemi algoritmici tende a favorire individui e comunità che si trovano già in posizione di forza e tende contemporaneamente a discriminare persone e gruppi già emarginati (Noble, 2018).

Gli studi sociali della tecnologia analizzano da tempo come i processi tecnologici, sovrapposti alle procedure amministrative e burocratiche, “materializzano” in artefatti e sistemi specifici fenomeni come il genere, la classe sociale, lo spazio geografico, la razza. Sebbene tali sistemi non diano origine ai pregiudizi, alle disuguaglianze e alle le strutture di potere dominanti, una volta in funzione possono amplificarli.

Al di là delle implicazioni per la privacy, viste all’inizio di questo paragrafo, la generalizzazione della previsione algoritmica e della personalizzazione solleva importanti questioni sull’impatto sociale a medio termine degli algoritmi su una varietà di questioni che includono la censura e la manipolazione – quando la personalizzazione è intesa a influenzare il comportamento – ma anche il rafforzamento di alcune forme di discriminazione generalizzata, la standardizzazione della società o la riduzione della libertà di scelta degli individui. Molti di questi aspetti sono particolarmente rilevanti per l’uso di algoritmi predittivi nel campo dell’*intelligence* e delle indagini di polizia. L’attuale dibattito sulle forme di discriminazione relative alla governance algoritmica fornisce un’opportunità di riflessione su come migliorare l’equità sociale nei sistemi automatizzati.

1. Conclusioni

La pervasività degli algoritmi nella società umana e la diversità del loro impiego hanno implicazioni politiche che appaiono come una questione centrale del nostro tempo: sono una caratteristica chiave sia dell’ecosistema odierno della conoscenza e dell’accesso all’informazione che delle norme culturali sottostanti, in quanto contribuiscono alla formazione delle informazioni a cui accediamo e alla loro organizzazione.

Il concetto di governance algoritmica racchiude una vasta gamma di pratiche sociotecniche che ordinano e regolano il sociale, dalla gestione del lavoro alla moderazione dei contenuti passando per la polizia predittiva e l’accesso all’informazione sul Web. Facendo dialogare questi fenomeni e diversi campi di ricerca, il concetto mette in luce alcune principali controversie comuni a molti contesti di applicazione della governance algoritmica: la parzialità, l’equità, la capacità di agire, l’autonomia, la trasparenza sono questioni importanti e ricorrenti nell’integrazione degli algoritmi all’organizzazione della nostra società.

I casi di studio esaminati brevemente in questo capitolo mostrano che la governance algoritmica dipende dal suo contesto sociale, economico e politico, e può assumere forme molto diverse – a volte ottimizzata per servire gli interessi delle imprese, a volte piegata alla regolamentazione istituzionale o plasmata dalle controversie pubbliche. La governance algoritmica è sia una forma di regolamentazione e di modifica del comportamento sia una forma di gestione e di ottimizzazione dello stesso. Discutere di governance algoritmica oggigiorno significa affrontare questioni come l’opacità, la misura in cui norme, valori e interessi possono essere inscritti negli algoritmi, l’efficienza e l’efficacia, le asimmetrie di potere, la creazione di nuove inclusioni ed esclusioni (e il rafforzamento di quelle che sono meno nuove). Se la governance algoritmica sembra diventare più potente, intrusiva e pervasiva, ha anche il potenziale di diventare più inclusiva, permettere una maggiore diversità sociale e proteggere *by design* alcuni diritti fondamentali.

È probabile che la questione del rapporto tra algoritmi e loro regolamentazione occupi un ruolo sempre più centrale nello studio e nella pratica della governance di internet e, più in generale, della governance dei sistemi complessi e automatizzati che permeano il mondo di oggi, rendendo sempre più evidente la sua centralità e il suo carattere politico. Due facce del rapporto algoritmi/regole sono attualmente all’esame degli studiosi e del pubblico, e probabilmente lo saranno ancora di più nel prossimo futuro. Da un lato, vi è la questione della governance degli algoritmi da parte delle istituzioni. La riflessione giuridica e la normazione dovrebbero estendersi allo sviluppo tecnico degli algoritmi? In quali contesti e in quali forme la regolamentazione degli algoritmi dovrebbe essere promossa o sostenuta? Come valutarne o anticiparne gli effetti? Da un altro lato, sono necessarie ulteriori analisi di quanto il mondo in cui viviamo sia veramente governato da algoritmi – da chi li sviluppa, produce, mantiene. E occorre infine dedicare ulteriori ricerche non solo ai modi in cui gli algoritmi regolano le nostre vite, ma anche – come evidenziato da Solon Barocas e colleghi in un *provocation paper* che ha fatto scuola (2013) – a “cosa significherebbe resistere loro”.

Capitolo 6

Cybersicurezza

Michele Loi

1. Introduzione: perché un’etica per la cybersicurezza?

In questo capitolo presenterò alcuni aspetti essenziali dell’etica della cybersicurezza, che contribuiscono a chiarire l’importanza di questo ambito normativo in generale, e in particolare in relazione ad alcune delle questioni etiche sollevate da robotica e intelligenza artificiale.

L’ambito della cybersicurezza è molto ampio e va dalla protezione dell’informazione (non solo personale) a quella di sistemi e oggetti tecnologici, come pure delle persone che li utilizzano. I tre maggiori obiettivi della cybersicurezza vengono spesso riassunti con l’acronimo *CIA* che sta per *confidentiality* (confidenzialità), *integrity* (integrità), e *availability* (disponibilità)dell’informazione. Ovvero, la cybersicurezza si propone di assicurare che l’informazione venga scambiata in modo confidenziale, cioè che sia accessibile soltanto a coloro che sono autorizzati ad accedervi; integro, ovvero che non vi siano modifiche non previste e non autorizzate nei dati; e disponibile, per cui l’informazione sia sempre accessibile a coloro e ai sistemi che sono autorizzati a usarla, senza interruzioni di servizio.

La prima domanda da farsi è perché ritenere che vi sia una dimensione etica e filosofica nella cybersicurezza. Agli occhi del neofita, la cybersicurezza può apparire una semplice questione tecnologica o, al limite, psicologica. Non vi è necessariamente qualcosa di interessante dal punto di vista etico, o filosofico, in tutto ciò.

Tuttavia, uno sguardo più approfondito mostra che un dominio tecnologico e praticocome la cybersicurezza ha caratteristiche tipiche di quei contesti in cui emergono domande etiche a cui è difficile rispondere. Infatti:

1. L’etica ha a che fare con quelle situazioni in cui un agente umano può rendersi responsabile di benefici o danni per le altre persone. Il compito della cybersicurezza è quello di prevenire i danni alle persone che derivano dalla violazione della confidenzialità, integrità e disponibilità dell’informazione e dei sistemi che dipendono da questa. Inoltre la cybersicurezza garantisce che i benefici della tecnologia siano accessibili a chi ne ha diritto.
2. La cybersicurezza richiede una riflessione filosofica in etica, o quantomeno di una riflessione etica *critica* e più *profonda*, sucome bilanciare una *pluralità* di valori morali che non possono essere tutti promossi, o tutti promossi in massimo grado, o nella stessa misura (Berlin, 1994), per cui si rende necessario stabilire e giustificare determinate priorità tra i valori in questione. In queste situazioni, raramente si riesce a trovare soluzioni che evitino del tutto la necessità di *compromessi* tra valori diversi. Ovvero, è spesso necessario *prima* identificare le soluzioni che implicano un compromesso meno dispendioso dal punto di vista di tutti i valori in questione, o almeno di quelli considerati più rilevanti, e *poi* scegliere in quale misura sacrificare un valore al fine di preservarne un altro. Nella cybersicurezza i conflitti di valore e la necessità di compromessi derivano:
3. dalla natura complessa della sicurezza cyber come obiettivo complessivo, che include obiettivi non sempre perfettamente allineati come la confidenzialità (che include a sua volta la privacy), l’integrità e la disponibilità dei dati (Domingo-Ferrer, Blanco-Justicia, 2020);
4. dalla natura complessa della sicurezza cyber come prodotto emergente di misure, azioni, e dinamiche che coinvolgono reti di artefatti tecnologici e persone estese nello spazio e nel tempo. Questo fa sì che una soluzione di sicurezza cyber abbia molteplici effetti sugli interessi individuali, di aziende e dello stato, spesso in conflitto o tensione fra loro.

Il capitolo è strutturato come segue: nel paragrafo 2 analizzerò tre esempi di bilanciamento di valori nella sicurezza cyber, relativi al tracciamento digitale dei contatti in contesto epidemiologico, allo sviluppo di tecnologie che garantiscono l’anonimato su internet, e all’*hacking* etico. Dedicherò un po’ più di spazio al primo esempio, per due motivi: in primo luogo, perché permette di soffermarsi sul punto (a), ovvero su come le tensioni tra obiettivi contrastanti della sicurezza cyber (in questo caso, confidenzialità-privacy vs. integrità dei dati) diano luogo a tensioni tra obiettivi etici di ordine superiore (libertà dalla sorveglianza vs. protezione dal disordine sociale e dalla manipolazione politica). Gli altri due esempi invece si soffermano sul secondo tipo di *trade-off*, che emerge quando le stesse soluzioni tecnologiche aumentano la sicurezza cyber di un agente a scapito di un altro. Infine, nel terzo e ultimo paragrafo mi occuperò dell’intreccio tra problemi di etica della sicurezza cyber e etica della robotica e dell’intelligenza artificiale, il tema centrale di questo libro. Questa parte del saggio in particolare punta più alla praticità di una eventuale applicazione per analogia degli stessi principi, che alla profondità dell’analisi teorica.

#

# Esempi di scambi di valore (*trade-off*).

Non tutti i problemi di cybersicurezza hanno soluzioni che soddisfano i requisiti di confidenzialità, integrità e disponibilità in modo adeguato, e quando ciò accade ci troviamo davanti a scelte morali. A volte, per esempio, una soluzione più versatile sul piano dell’integrità, come il monitoraggio di tutta l’informazione da parte degli amministratori di sistema, può generare dei rischi per la confidenzialità, poiché informazioni private sono esposte a individui che non dovrebbero accedervi. Oppure, soluzioni capaci in teoria di garantire sia ottimi livelli di confidenzialità sia di integrità possono risultare complicate, riducendo quindi l’usabilità di una tecnologia (per es., occorre autenticarsi per accedere all’informazione), o incoraggiando inadempienze in chi usa il sistema (per es., l’utente scrive la password su un biglietto attaccato al computer dell’ufficio per non doverla memorizzare). Può non essere del tutto evidente che tali tensioni tra obiettivi discordanti, e i compromessi che queste rendono necessari, possono avere una dimensione etica importante. Per mostrare ciò, inizierò l’analisi con la discussione della dimensione della cybersicurezza nella scelta dell’architettura di un sistema di tracciamento contatti attraverso il *bluetooth* dei cellulari, come quello che è stato implementato in diversi paesi in Europa e nel mondo nel 2020 (Morley *et al*., 2020).

## Esempio 1: tracciamento contatti digitale

Consideriamo le applicazioni introdotte allo scopo di tracciare il contatto ravvicinato tra persone potenzialmente contagiate dal Covid-19. In questo capitolo, ci focalizzeremo su una questione specifica: la scelta tra un protocollo *centralizzato* e uno *decentralizzato*. I due protocolli espongono il cittadino a rischi di natura diversa: rispetto al centralizzato, una violazione del server centrale potrebbe rendere possibile risalire ai rapporti sociali di (potenziali) milioni di utilizzatori del protocollo. Nel caso del protocollo *decentralizzato* DP-3T (*DP-3T/Documents*, n.d.), è più difficile impedire attacchi organizzati da parte di organizzazioni criminali volti a creare falsi allarmi (qualora la app risultasse molto popolare e venisse utilizzata dalla maggioranza delle persone, come i suoi sostenitori desiderano). In Italia, la app Immuni, gestita dal governo, utilizza un protocollo del tutto analogo.

Entrambi i protocolli sono stati studiati per minimizzare i rischi per la privacy dovuti al tracciamento dei contatti ravvicinati tra individui. Il tracciamento contatti è un’attività tradizionale di salute pubblica durante le epidemie dovute a malattie infettive. Tuttavia, il tracciamento tradizionale è un’attività che si basa sul lavoro umano. La scarsità delle risorse a disposizione per il trattamento umano ne limita l’efficacia, rendendo il processo impossibile da gestire una volta raggiunto un numero molto alto di contagi. Ciò tuttavia fa anche in modo che il sistema non possa essere convertito in una forma di sorveglianza di massa, quantomeno nascondendolo all’opinione pubblica. Durante le prime fasi dell’epidemia è stata dunque proposta per la prima volta l’idea di ricostruire le catene di contatti tra persone infette e persone a rischio di avere ricevuto l’agente patogeno attraverso i segnali *bluetooth* dei telefoni cellulari di cui la maggioranza delle persone è dotata.

Nel modello decentralizzato l’informazione su chi è stato vicino a chi non può essere ricostruita senza accedere all’informazione contenuta nei telefoni delle persone coinvolte. Non esiste un unico server in cui le catene di contatto siano state ricostruite, neppure in modo anonimo. Il server centrale esiste, ma è solamente un tabellone in cui vengono annunciati dei codici che corrispondono a persone risultate positive al Covid-19. È molto difficile (ma non impossibile, come vedremo) risalire all’identità delle persone a partire da tali codici. Una associazione criminale potrebbe ad esempio creare un virus che accede all’informazione, o nel caso di un obiettivo ad alto valore (ad esempio un politico) potrebbe comperare dei dispositivi da utilizzare esclusivamente quando tale persona, e solo tale persona, è presente. A quel punto è facile inferire che, se il dispositivo riceve una notifica, l’obiettivo ad alto valore ha contratto la malattia. Ciò che si è voluto evitare nel protocollo decentralizzato è il rischio aggiuntivo derivante dall’accumulo di informazione in un server centrale. Vale la pena di citare per esteso la documentazione del protocollo DP-3T:

I progetti con componenti centralizzate, in cui un singolo attore, come un server o uno stato, può apprendere molto sugli individui e sulle comunità, hanno bisogno di un'attenzione specifica perché se vengono attaccati, compromessi o riprogrammati, possono creare un danno maggiore. *Per affrontare questi problemi, noi invece realizziamo lo stesso compito utilizzando un design decentralizzato che non richiede la raccolta e l'elaborazione centralizzata delle informazioni sugli utenti* (Troncoso *et al*., 2020 enfasi nell’originale, trad. dell’autore).

Il valore perseguito in maniera prioritaria è quindi quello di proteggere la confidenzialità dell’informazione medica rispetto a potenziali abusi da parte di agenti che ottengano il controllo del server centrale. Non è possibile eliminare scenari come quelli descritti in cui è possibile identificare persone particolari, ma l’assenza di un singolo punto critico fa in modo che chi ha un accesso privilegiato al server non abbia alcun vantaggio rispetto a un criminale comune. Il rischio di scenari come quello di una schedatura di massa della rete di rapporti sociali dei cittadini risulta particolarmente remoto (Troncoso *et al*., 2020; Vaudenay, 2020).

Possiamo dunque concludere che il modello decentralizzato di tracciamento contatti attraverso il *bluetooth* rappresenti la soluzione ideale, che dovrebbe essere preferita a qualsiasi altra soluzione e la cui adozione dalla maggior parte della popolazione risulterebbe la più sicura possibile? La risposta a tale quesito non appare così ovvia una volta che vengano considerati altri rischi di cybersicurezza oltre alla violazione della confidenzialità. Consideriamo il secondo valore protetto dalla cybersicurezza, quello della *integrità* dei dati. Nel caso del tracciamento contatti, qualsiasi manipolazione dell’informazione che genera un falso positivo può essere considerata una violazione di tale requisito. Ciò risulta, purtroppo, più facile nel caso del modello decentralizzato. Si consideri lo scenario seguente:

Scenario 9 (L’allievo Bart Symptomson). L'allievo Bart Symptomson ha un test la prossima settimana, ma non ha studiato bene. Grazie a un annuncio, trova il signor Starnuto che ha dei sintomi e accetta di prestare a Bart il suo telefono. Bart passa il telefono del signor Starnuto per tutta la classe, poi si aggira per la sala professori. Bart lo restituisce poi al signor Starnuto, che va da un medico. Il medico nota che il signor Starnuto è malato di COVID e lo segnala tramite l'applicazione. Questo fa scattare un allarme per tutta la classe e per tutti gli insegnanti, causando la chiusura della scuola! (Bonnetain *et al*., 2020 trad. dell’autore)

Consideriamo uno scenario (teoricamente ottimale) in cui l’ottanta percento della popolazione sia dotato di un telefono con la app Immuni. Si consideri una potenza straniera ostile, dotata di ampi fondi da dedicare a finalità di terrorismo politico, come per esempio destabilizzare un governo considerato ostile. Potrebbe finanziare una struttura criminale che sfruttando l’informazione resa pubblica dal server centrale dà luogo a un commercio di ID positive per la app Immuni. Si tratta di quelle ID che, pur rendendo molto difficile risalire all’identità della persona da cui provengono, attivano una catena di allarmi che poi permette di generare falsi allarmi, anche di grande entità, dove desiderato, con una logica simile a quella dell’allievo Bart Symptomson dell’esempio. Tale problema risulta di difficile risoluzione nel modello decentralizzato, proprio in virtù del fatto che tutti gli attori in gioco hanno accesso alla stessa informazione: chi controlla il server (realisticamente, lo stato) non può secretare l’informazione sulla positività delle catene di contagio di cui dispone. Non sarebbe troppo difficile per una struttura criminale dotata di ingenti fondi, in questo esempio, creare un sistema simile a una lotteria, dove chiunque abbia acquistato un biglietto riceve un premio se poi risulta contagiato dal Covid, a patto di cedere la possibilità di usare il proprio codice (capace di attivare una catena di allarmi) alla struttura criminale in questione. L’approccio sarebbe efficiente, perché anche qualora aderisse la metà degli utilizzatori della app in tale scenario, l’organizzazione criminale finirebbe per pagare soltanto coloro che risultano effettivamente malati e capaci di attivare una catena di allarmi. Tutto ciò non richiede neppure transazioni di natura legale potenzialmente costose da gestire. Il fatto che l’informazione di contatto sia resa pubblica dal server centrale (anche se in un modo che rende difficile risalire all’identità delle persone) fa sì che sia possibile costruire un contratto *smart*, ovvero un algoritmo che automaticamente garantisce che il pagamento dell’organizzazione criminale arrivi a coloro la cui positività sia dichiarata pubblicamente dal server, e a nessun altro. Il costo per gestire tali pagamenti è molto basso, perché gli algoritmi, non le persone, possono garantire in modo automatico il funzionamento dell’operazione. Nessuna strategia di mitigazione di questo rischio pare compatibile con i principi alla base del protocollo decentralizzato (Vaudenay, 2020).

Abbiamo dunque a che fare con una scelta di questo tipo: da un lato, con il sistema centralizzato, un rischio difficilmente calcolabile di sorveglianza di massa attraverso l’accesso all’informazione in un server, che benché sia priva di identificativi personali palesi come nomi o indirizzi, si presta a inferenze che possono rivelare l’identità delle persone e il loro grafo sociale (ovvero la rete di incontri), per almeno due settimane (il periodo in cui occorre ricostruire il contagio da Covid-19). Dall’altro, un rischio difficilmente calcolabile che organizzazioni terroristiche mettano in atto reti atte a sabotare il vivere civile, attraverso un mercato nero di dati che attivano le allerte per i possibili contagi di app come Immuni.

È plausibile sostenere che stimare quale rischio sia il più grave per una società democratica non sia semplicemente una valutazione tecnica o scientifica, ma una scelta che mette insieme, in modo difficile da districare, considerazioni di natura prettamente tecnica con assunti di natura morale e politica. Una variabile politica importante è, ad esempio, il grado di fiducia nelle istituzioni politiche. Un esperto che abbia un alto grado di fiducia nel governo sarà meno riluttante ad adottare il sistema centralizzato, perché si fida della capacità del governo di proteggere il server con misure di cybersicurezza adeguata e non ritiene altamente probabili abusi con copertura governativa. Un esperto che invece ritenga che l’abuso di potere da parte del governo del proprio paese sia più probabile rispetto al rischio di attacchi da parte di criminali al soldo di potenze straniere potrebbe temere i rischi del sistema centralizzato in misura maggiore rispetto a quelli del sistema decentralizzato. La questione potrebbe essere riformulata come la possibilità di *bias* cognitivi dovuti all’influenza implicita di considerazioni politiche di ampio respiro all’interno delle valutazioni degli esperti riguardo ai rischi di sistemi complessi e capaci di influenzare intere società. Ciò è eticamente rilevante in quanto è molto problematico, anche a livello concettuale, distinguere nettamente il piano valoriale da quello puramente descrittivo nell’ambito di questo genere di valutazione politica. Il primo esperto tenderà a sottostimare i rischi di finire per costruire un sistema di sorveglianza di massa, il secondo tenderà a sottostimare i rischi di rendere il proprio paese vulnerabile ad attacchi e influenze politiche dall’esterno. La riflessione etica può svolgere un ruolo utile anche per riconoscere più agevolmente il *bias* cognitivo in questione in parallelo, e non in opposizione, al tentativo di definire nel modo più chiaro possibile ciò che deriva da fatti osservativi valutati con metodologie ampiamente condivise e quanto sconfina nell’ambito della mera opinione soggettiva, magari non informata dagli esperti dell’ambito. Inoltre, persino esperti che siano in totale accordo riguardo alle probabilità di determinati tipi di attacchi potrebbero essere in disaccordo riguardo a quale dei due rischi sia moralmente prioritario evitare. Tale disaccordo potrebbe essere non risolvibile in presenza di differenze nella valutazione *morale* su quale sia il male peggiore da parte degli esperti. Ad esempio, un esperto che ritenga che il disordine civile dovuto ai falsi allarmi di natura criminale e la manipolazione del processo elettorale sia il male minore a confronto della sorveglianza di massa potrebbe ritenere più accettabile il genere di rischio a cui si espone il modello decentralizzato. Un esperto che abbia una scala di valori opposta riguardo a quale sia il male minore potrebbe invece preferire correre il rischio legato all’adozione del modello centralizzato. Alla fine del secondo paragrafo, tratteremo brevemente le implicazioni normative che derivano da questo modo di concepire il rapporto tra etica e cybersicurezza.

## Esempio 2: tecnologie per l’anonimato su internet

Consideriamo ora quelle soluzioni tecnologiche, ad esempio l’accesso a internet attraverso il software TOR,[[34]](#footnote-34) che rendono molto difficile che la propria attività online venga tracciata. Queste soluzioni sono utili allo scopo di proteggere la confidenzialità dell’attività online. È chiaro come una tale protezione possa aumentare gli spazi di libertà, in particolare la libertà di parola, per categorie oppresse come gli oppositori politici e i giornalisti nei paesi che violano i diritti umani. Inoltre, possono essere viste come strumenti per evitare lo sfruttamento dei propri dati di navigazione da parte delle grandi multinazionali di internet. Dal punto di vista dell’oppositore politico o del giornalista che ne fa uso, tali soluzioni sono mezzi per aumentare la propria sicurezza cyber, poiché rendono più difficile che le loro attività online private vengano spiate da altre persone.

Dal punto di vista dei governi, però, la stessa tecnologia contribuisce al rischio per la sicurezza dei cittadini: possono rendere più facile commettere reati non-cyber, come la diffusione di materiale pornografico o il ricatto online, e possono rendere più facili reati che rientrano nell’ambito della violazione della sicurezza cyber, cioè reati di qualsiasi tipo che comportano la violazione dei requisiti di confidenzialità, integrità, o accessibilità dei dati. In generale, infatti, qualsiasi tecnologia efficace a garantire l'anonimato di un agente rende molto più arduo identificare il responsabile di un crimine, sia esso di natura cyber o meno, perpetrato online. Ad esempio, è più facile diffamare una persona riuscendo a evitare le conseguenze legali di tale crimine. La tecnologia che garantisce l’anonimato può essere usata a tale scopo. Il problema ha una rilevanza non solo civile ma politica, dato che è legato a tentativi di manipolare l’opinione pubblica influenzando i processi democratici, ad esempio quando la persona la cui reputazione è in gioco è un candidato alle elezioni. Ad esempio, l’hacker che ha violato l’account email di una persona nello staff di Hillary Clinton durante la campagna elettorale per le elezioni negli Stati Uniti del 2016 ha utilizzato tecnologie di questo tipo (Groll, Francis, 2016; GOVERNANCE pp….).

Casi simili possono essere analizzati come esempi di *trade-off* tra valori diversi. Da un lato, proteggere le tecnologie di anonimizzazione significa tutelare la sicurezza cyber degli utilizzatori del software stesso, permettendogli di sfuggire a forme di sorveglianza digitale sempre più invasive, tutelando la *confidenzialità* della propria comunicazione. Sull’altro piatto della bilancia troviamo valori come la dignità delle persone (nel caso della pedopornografia o della calunnia) o l’integrità dei processi elettorali (nel caso in cui la reputazione di un candidato alle elezioni venga compromessa per finalità politiche).

Lo stesso esempio, tuttavia, può essere descritto come un *trade-off* tra valori di sicurezza cyber realizzati in attori diversi. Utilizzare un protocollo che garantisce l’anonimità per violare un account email altrui significa infatti dotarsi un alto livello di sicurezza cyber relativamente alla confidenzialità della propria navigazione in rete allo scopo di compromettere la confidenzialità dell’informazione di un’altra persona.

Data l’una o l’altra interpretazione dello scambio di valori in questione, è importante per tutti gli attori implicati nella regolamentazione e nella protezione di reti e software essere consapevoli della natura duplice di tali tecnologie. Ad esempio, un’azienda che fornisca servizi di VPN (un’altra soluzione che rende più difficilmente tracciabile la navigazione) preferirà auto-descriversi come una società che rispetti i valori della cybersicurezza, in particolare quello di proteggere la confidenzialità dell’informazione sulla navigazione online degli utenti. Ma, in virtù della stessa funzione, può essere accusata di essere uno strumento nelle mani di coloro che intendono minacciare la sicurezza cyber di altri individui e organizzazioni. È quindi plausibile che un’azienda di questo tipo dovrebbe dotarsi di una forma di deliberazione etica su come trattare conflitti di valore che emergono, ad esempio, quando viene deciso come registrare le attività dei propri clienti.

## Esempio 3: *hacking* etico

Un insieme di scenari che si prestano a generare importanti dilemmi etici è dato dal contesto del cosiddetto *hacking* “etico”. L’*hacking* etico potrebbe essere definito, in senso molto ampio, la violazione della sicurezza cyber altrui con una finalità pro-sociale. In pratica, la finalità pro-sociale in questione è tipicamente la protezione degli interessi di un committente. In altre parole, nel linguaggio comune si parla di *hacking* etico in senso *non controverso* quando ci si riferisce all’azione di un professionista che effettua un attacco contro la sicurezza cyber di un individuo o di una organizzazione *con il suo consenso* allo scopo di testarne la sicurezza. Il termine viene utilizzato prevalentemente per indicare queste attività, ma ci sono altre attività che potrebbero essere indicate senza grossi problemi sul piano morale come *hacking* etico, in quanto hanno chiare finalità pro-sociali. Si pensi, ad esempio, ad alcuni ricercatori universitari nell’ambito della sicurezza cyber che svolgono analisi su un software per ricercarne le vulnerabilità. Una volta identificata una vulnerabilità, contattano l’organizzazione responsabile, ad esempio il produttore del software nel caso di un software commerciale. Una strategia tipica, che può essere considerata eticamente responsabile in molti casi, è quella di annunciare alla parte coinvolta la decisione di rendere la vulnerabilità pubblica dopo un periodo di embargo, nel caso in cui chi è responsabile della sua esistenza non si preoccupi di risolverla.[[35]](#footnote-35) Una pratica di questo tipo è *hacking* eticonel senso che ha una chiara finalità pro-sociale; ma a differenza di quello che pare affermarsi come l’uso più comune del termine, non si tratta di un’attività di *hacking* perseguita *per conto* di un committente e con il suo *consenso.* Un *hacking* etico di questo tipo può portare a un *conflitto* con gli interessi della parte interessata dall’attacco (Jaquet-Chiffelle, Loi, 2019). Il dilemma etico qui è se rendere pubblica la vulnerabilità quando la parte responsabile dimostri di trascurare il problema. In casi come questo, la conoscenza della teoria etica filosofica può essere preziosa per il ricercatore di sicurezza cyber, al fine di riflettere sul problema da più di una prospettiva. Ad esempio, un approccio consequenzialista implica che la pubblicazione della vulnerabilità sia eticamente obbligatoria quando le conseguenze nel lungo periodo di tale scelta siano tutto considerato positive al netto dei danni prodotti. Un’etica diversa, ad esempio basata sui diritti o sul contrattualismo, potrebbe portare nelle stesse situazioni a privilegiare interessi particolari di breve periodo, in quanto il danno arrecato dalla pubblicazione potrebbe violare un diritto o comunque interessi importanti di specifici individui in modo così serio da non poter essere giustificato dalla somma dei benefici aggregati nel lungo periodo, a prescindere da quanto essi siano grandi (Loi, Christen, 2019).

In tutti questi esempi, l’etica può apparire come una complicazione di cui è inutile occuparsi, poiché, a differenza delle questioni che si prestano ad analisi di natura tecnica, è difficile dotarsi di metodologie etiche che determinino in modo ampiamente condiviso quale sia la scelta migliore (CODICI ETICI pp….). Questa conclusione tuttavia non è sostenibile, in primo luogo perché, come mostra il primo esempio, non esplicitare un disaccordo di valori *implicito* può rendere più difficile dipanare la natura del disaccordo tecnico. L’esempio sul tracciamento contatti digitale mostra, plausibilmente, che quando le valutazioni etiche e politiche sono intrecciate a quelle tecniche, il tentativo di escludere la discussione etica e politica dall’ambito delle scelte sulle tecnologie da adottare rende le motivazioni non strettamente tecniche sullo sfondo poco trasparenti. Il tentativo di escludere il piano della riflessione etica o politica non facilita l’accordo, perché il disaccordo etico o politico continua a influenzare le valutazioni, ma senza essere concettualizzato come tale, in modo palese. I metodi di valutazione morale *strutturati* che sono oggetto di attenzione da parte di filosofi e altri esperti di etica applicata (ad esempio, seguire un approccio deontologico o utilitaristico) probabilmente non riescono a produrre una soluzione ampiamente condivisa nella maggioranza dei casi. Come obiettivo minimo, lo studio dei diversi modelli di valutazione etica ci permette di identificare le premesse nascoste dietro una valutazione pratica. Ad esempio, una premessa utilitaristica come l’idea che il giusto consista nel massimizzare l’utilità generale può essere spesso implicita in una analisi apparentemente di natura tecnica. Ma l’utilitarismo non è l’unica forma di ragionamento etico sistematico e in alcuni casi porta a conclusioni che la gran parte delle persone ritiene del tutto inaccettabili. La capacità di identificare una premessa utilitaristica implicita, quindi, permette di criticare una soluzione che potrebbe essere stata offerta come ovvia e come l’unica possibile, avviando un processo più creativo di esplorazione di alternative diverse. Questo vale, ovviamente, in riferimento a qualsiasi modello etico. Ad esempio, potrebbe essere altrettanto utile dubitare di un ragionamento basato su premesse deontologiche difese in modo dogmatico in contesti nei quali producono effetti non ottimali sul piano del benessere collettivo.

L’utilità pratica dell’etica filosofica applicata alla cybersicurezza non dipende quindi esclusivamente dalla sua abilità di indicare una soluzione preferita da tutti. Consiste anche soltanto nell’aiutare le persone comuni e i politici a riconoscere le premesse normative nascoste nel discorso degli esperti (si pensi all’esempio 2.1). Quando tale riconoscimento ha luogo, seguono alcune conseguenze pratiche importanti, una di ordine epistemico-deliberativo, e l’altra di ordine politico. Dal punto di vista epistemico, una volta riconosciuta la dipendenza della valutazione di cybsersicurezza da premesse morali o politiche in senso più ampio, occorre rendere la natura della deliberazione sulla cybersicurezza più inclusiva, integrando il contributo di esperti in ambiti come la scienza politica, la sociologia, e l’etica. Dal punto di vista politico, in una democrazia il compito di stabilire cosa sia eticamente e politicamente preferibile non è degli esperti di sicurezza cyber, ma in ultima analisi della cittadinanza tutta, la quale esprime la propria volontà attraverso un governo democraticamente eletto.

# Etica della cybersicurezza, robotica e intelligenza artificiale

Se è vero che la cybersicurezza ha una dimensione etica fino ad ora poco esplorata (Christen *et al*., 2020; Taddeo, 2013), l’analisi etica di robot e intelligenza artificiale dovrebbe includere le questioni di sicurezza cyber. In questa ultima parte del capitolo, prenderemo spunto dalle linee guida etiche sull’intelligenza artificiale per determinare in modo pratico e concreto alcuni aspetti di cybersicurezza che dovrebbero essere considerati parte dell’etica dell’intelligenza artificiale. Per semplicità espositiva, seguiremo i princìpi etici illustrati nel documento europeo *Orientamenti etici per un’IA affidabile* (HLEG AI, 2019), per la cui analisi rimandiamo a CODICI ETICI pp…..

## Cybersicurezza e prevenzione del danno

Iniziamo dal principio della *prevenzione del danno*, che è ovviamente il valore più vicino alla finalità intrinseca della sicurezza cyber. Una soluzione informatica che non protegge il benessere e i diritti degli utilizzatori non può essere considerata eticamente appropriata a prescindere dalle qualità ulteriori che potrebbe avere. Sottovalutare i rischi legati alla cybersicurezza può inoltre portare al rifiuto di una tecnologia qualora il danno si materializzasse. Si pensi ad esempio all’utilizzo dei robot per l’assistenza agli anziani. Un hacker – magari un hacker attivista che ritenga moralmente obbligatorio per lui agire in questo modo – che riuscisse a controllare a scopo dimostrativo i robot di una casa di cura per anziani, infliggendo danni fisici o psicologici di entità rilevante, potrebbe realisticamente portare a un rifiuto della tecnologia in questione per un periodo considerevole. In alcuni casi, tuttavia, la protezione del danno deve essere perseguita in modo bilanciato per non compromettere l’usabilità della stessa tecnologia. In ambito medico, ad esempio, alcune procedure di autenticazione che confliggono con l’usabilità delle tecnologie in situazioni di emergenza non sono eticamente giustificabili se portano a un rischio per la salute maggiore rispetto al rischio derivante da un possibile attacco informatico (Loi *et al*., 2019).

## Cybersicurezza e rispetto dell’autonomia umana

Consideriamo ora il principio del *rispetto della autonomia umana*. Uno scenario emblematico di potenziale violazione di questo valore da parte dell’intelligenza artificiale nel contesto della cybersicurezza deriva dalla possibile corsa agli armamenti tra attacchi hacker e sistemi di difesa potenziati dall’intelligenza artificiale. La prima minaccia all’autonomia umana è quella del *deskilling*: se viene consolidata l’abitudine ad avvalersi di sistemi automatici di riconoscimento degli attacchi, vengono a mancare nel lungo periodo le abilità umane che possono risultare essenziali per proteggere i sistemi da attacchi più sofisticati (Taddeo, 2019). Nell’interpretazione data nel documento *Orientamenti etici per un’IA affidabile*, il rispetto di tale principio richiede che venga scongiurata la subordinazione dell’elemento umano a quello dell’IA. Questo principio è violato se alcune attività umane erroneamente classificate come attacchi vengono bloccate e le vittime di tale interferenza non hanno a chi rivolgersi per richiedere la rettifica dell’errore, cosa che potrebbe accadere se non ci fossero esperti umani in grado di valutare la natura dell’attacco con sufficiente sicurezza. Il secondo scenario in cui viene violato il principio di autonomia è legato al principio di spiegabilità degli attacchi e lo discuteremo quando prenderemo in considerazione il quarto principio nel dettaglio.

## Cybersicurezza e imparzialità

Consideriamo ora il valore della *fairness*, intesa come imparzialità o giustizia, in relazione a sistemi automatici di rilevamento degli attacchi hacker. Uno dei problemi più rilevanti di tutti i sistemi basati su meccanismi di apprendimento statistico è quello di essere ingiusti o indirettamente discriminatori. Consideriamo, ad esempio, un sistema per eliminare account fasulli da un social media nel contesto della protezione della sicurezza cyber nazionale contro gli attacchi di manipolazione politica da parte di uno stato estero. È plausibile immaginare che un sistema del genere, sviluppato con l’apprendimento automatico, abbia un’inclinazione statistica a classificare in modo avverso gli account di persone non madrelingua, cioè che tenda più facilmente a generare falsi positivi in questo gruppo demografico e falsi negativi nell’altro. Il problema in questione non è un semplice problema tecnico e non è di facile risoluzione, perché l’eliminazione di una forma di discriminazione indiretta tipicamente porta a generarne una forma diversa. Il riferimento è alle impossibilità matematiche che non ci permettono di ottenere, nei casi più tipici, forme di parità tra gruppi che intuitivamente risultano eticamente desiderabili ma impossibili da garantire simultaneamente (Berk *et al*., 2021). Ad esempio, potrebbe essere impossibile garantire una predizione sia positiva sia negativa che sia egualmente accurata nei due gruppi (per es., italiani vs. stranieri) e che al contempo renda uguale il tasso di falsi positivi e di falsi negativi nei due gruppi (Chouldechova, 2017). Per afferrare la rilevanza del problema, che costituisce uno dei dilemmi etici (e non semplicemente tecnici) irrisolti nell’ambito dell’etica dell’apprendimento automatico, si immagini di dovere scegliere tra due sistemi giuridici con le seguenti proprietà statistiche:

1. Sistema penale dell’eguale opportunità: la probabilità che una persona innocente venga erroneamente condannata e che una persona colpevole venga erroneamente riconosciuta innocente è la stessa per italiani e stranieri
2. Sistema penale dell’eguale accuratezza: la probabilità che una persona condannata sia in realtà innocente e che una persona condannata sia in realtà colpevole è la stessa per italiani e stranieri.

È plausibile che il senso di giustizia di ciascuno suggerisca che, intuitivamente, un sistema penale giusto dovrebbe avere, simultaneamente, le proprietà statistiche di (a) e (b). In assenza di competenze matematiche è persino difficile riuscire a distinguere le due condizioni fra di loro, data la loro somiglianza superficiale. Tuttavia, non è difficile dimostrare che, qualora l’incidenza del crimine sia diversa tra i due gruppi (ad esempio, se in proporzione ci sono più reati tra gli italiani che tra gli stranieri) un qualsiasi sistema di decisione non potrà soddisfare entrambe le proprietà, a meno di non avere una accuratezza perfetta – ovvero, a meno di non essere un sistema che condanna tutte e solo le persone colpevoli (Chouldechova, 2017).

Il problema di identificare un account fasullo tra tanti autentici è strutturalmente identico: basta sostituire nell’esempio il gruppo dei colpevoli a quello degli account fasulli e il gruppo degli innocenti a quello degli account reali, e la condanna penale con la rimozione dell’account. Quindi, se vi fosse, per una qualsiasi ragione sociologica, una diversa distribuzione degli account fasulli tra le due popolazioni, si riproporrebbe il dilemma tra un sistema *equo* in senso statistico secondo la definizione (a) di eguaglianza di opportunità o (b) di eguaglianza di accuratezza. Uno dei temi sui quali la comunità dell’apprendimento automatico si sta interrogando in modo interdisciplinare, includendo studiosi di etica, è quello di giustificare una scelta valoriale tra opzioni di questo tipo, in modo da selezionare l’obiettivo statistico più sensato di equità nel contesto dato.

## Cybersicurezza e spiegabilità

Infine, consideriamo ora il quarto valore etico menzionato in *Orientamenti etici per un’IA affidabile*, quellodella spiegabilità, mettendolo in relazione con il valore del *controllo.* Per controllo si intende la capacità di chi è responsabile della progettazione o dell’uso di un sistema di IA di prevedere il comportamento di tale artefatto e di stabilire quanto esso sia capace di realizzare le finalità a cui è preposto in misura adeguata alle circostanze.

Il valore della spiegabilità sembra avere assunto un’importanza peculiare nel dibattito sull’etica della tecnologia dopo l’avvento di tecniche sempre più complesse di apprendimento automatico (esaminate nel capitolo 2). Nell’apprendimento automatico, il programmatore non definisce i criteri che determinano il comportamento di un sistema. Piuttosto, il programmatore definisce le meta-regole che determinano come il sistema sceglierà da sé tali criteri, ad esempio testandone alcuni, comparandoli, in modo orientato a un obiettivo definito dal programmatore. Dando la possibilità al sistema di modificare gradualmente i propri criteri e di testarli su un numero enorme di casi, il sistema finisce per selezionare (in base a meta-regole date a priori, del tutto trasparenti al programmatore) le regole che sono risultate più idonee a raggiungere il suo scopo. Per farsi un’idea del problema della spiegabilità, partiamo da un esempio che si avvicina al concetto per alcuni aspetti (sebbene, come vedremo, non per altri)*.* Supponiamo che il programmatore debba generare un sistema che classifica le immagini di gatti, massimizzando il numero di immagini classificate in modo corretto. Il sistema è settato sul criterio “se ha quattro gambe, un profilo tondeggiante, è grigio, allora è un gatto”. L’algoritmo di apprendimento automatico definito dal programmatore gli dice di scartare tale criterio, dopo averne generato una variazione (per es.: “ha quattro gambe, un profilo tondeggiante, e una coda”) se il nuovo criterio riduce il numero di errori. Il problema che l’apprendimento automatico pone al principio della spiegabilità è che il genere di criteri che vengono selezionati dai sistemi *non sono per nulla simili a quelli degli esempi utilizzati in precedenza*. Tipicamente, non corrispondono ad alcun concetto umano. Nei casi più fortunati riusciamo a trovare qualche corrispondenza con criteri per noi intellegibili. Ad esempio, riusciamo a capire che determinati gruppi di *pixel* di una determinata forma e colore vengono utilizzati come criterio positivo nella valutazione di un’immagine come gatto. Tuttavia, tale corrispondenza non è mai perfettamente accurata: ci possono essere, ad esempio, delle circostanze in cui lo stesso gruppo di *pixel* porta a una classificazione negativa. Ciò accade, tipicamente, usando algoritmi di apprendimento automatico basati sulle reti neurali. In questo caso è particolarmente improbabile che i criteri che vengono selezionati dal sistema possano essere perfettamente rappresentati attraverso concetti umani. Una rete neurale è generata sulla base di obiettivi chiaramente definiti dal programmatore e sulla base di regole matematiche chiare, ma è come se il sistema scrivesse i propri criteri operativi in un’altra lingua, a noi sconosciuta. Controllare un artefatto di questo genere è possibile, ma non attraverso le stesse strategie di controllo che possono essere usate nei confronti di un sistema il cui codice scritto a mano definisce, in un linguaggio a noi comprensibile, i criteri e i parametri che ne definiscono l’azione. Si parla quindi di scatole nere. La differenza tra il controllo di un sistema “trasparente” rispetto a uno che è una scatola nera può essere paragonata a quella tra leggere il pensiero di una persona e conoscere una persona talmente bene da sapere cosa aspettarsi in quasi tutte le circostanze di cui si ha già avuto esperienza, senza tuttavia poter garantire al 100% che quella persona agirà così, e avendo sempre qualche dubbio su come tale persona potrebbe agire in circostanze insolite.

Consideriamo gli esempi precedenti alla luce del valore della spiegabilità e del controllo. Un algoritmo di difesa dagli attacchi cyber, come abbiamo detto, può minacciare l’autonomia umana qualora portasse al *deskilling*, perdita di abilità e competenze umane in ambito di sicurezza cyber. Questo accadrebbe in misura maggiore qualora gli algoritmi di difesa cyber risultassero governati da ragioni opache, comprensibili soltanto a un livello di astrazione molto elevato (per esempio, nei termini delle funzioni che ottimizzano). Il rischio è quello dell’esperto di cybersicurezza che si limita a programmare gli obiettivi di una macchina che poi apprende e decide in piena autonomia operativa, senza un controllo umano significativo, quali strategie di cybersicurezza siano più adeguate. In tali circostanze, l’esperto diventa lo spettatore passivo di uno spettacolo indecifrabile. Immaginiamo una competizione tra attacchi e difese cyber governate da sistemi di IA nel quale né i progettisti né i gestori di entrambi i processi siano più in grado di stabilire *quali* strategie saranno implementate per arrecare il maggior danno all’avversario, e per quale motivo proprio *quelle* strategie. Il controllo è dunque strumentale rispetto al valore in sé dell’*autonomia*. In secondo luogo, il controllo è strumentale rispetto al principio etico della *prevenzione del danno*, poiché in uno scenario di corsa alle armi tra sistemi di pirateria e difesa informatica entrambi pienamente automatizzati e basati sull’apprendimento automatico la perdita di controllo sulle strategie portate avanti dai sistemi di IA potrebbe portare a rischi maggiori di danni collaterali. Si tratta di uno scenario reso familiare – quantomeno nell’ambito della cultura pop degli anni ’80 – dal film *War Games*. Nella pellicola, tuttavia, i programmi sfuggiti di mano agli umani (tra cui un giovane hacker che ancora frequenta la scuola superiore) rischiavano di innescare una guerra nucleare tra le due superpotenze di allora. Nello scenario reso possibile dall’uso dell’IA per difese e attacchi cyber, invece, il danno che rischia di essere generato è quello della distruzione delle infrastrutture digitali sulle quali la società fa affidamento in misura sempre maggiore (Viganò *et al*., 2019). Occorre poi tenere conto del fatto che il confine tra danno informatico e danno fisico può essere facilmente oltrepassato da attacchi che abbiano come obiettivo sistemi informatici essenziali per le infrastrutture fisiche. Per queste ragioni, alcuni studiosi hanno invocato una qualche forma di regolamentazione internazionale sull’utilizzo di queste tecnologie. Infine, i valori procedurali della spiegabilità e del controllo sono legati al valore della *imparzialità.* Consideriamo di nuovo l’esempio dell’eliminazione di account fasulli dai social network. In assenza di spiegabilità – ossia di un modello mentale del perché un algoritmo classifichi certi profili nel modo in cui lo fa - è più probabile che determinati *bias* a sfavore di specifici gruppi demografici non vengano riconosciuti, e questo rende difficile anche solo iniziare una riflessione normativa sul requisito morale di evitare determinati *bias*.

1. Conclusione

In conclusione, l’etica della cybersicurezza è un aspetto ingiustamente sottovalutato della discussione etica sull’intelligenza artificiale e sulla robotica, e in generale sulle tecnologie digitali. L’esempio della scelta di un protocollo centralizzato o decentralizzato per il tracciamento digitale dei contatti mostra che l’analisi etica della progettazione di un’infrastruttura digitale è incompleta se non vengono considerati i problemi di cybersicurezza. Infatti, le scelte relative alla cybersicurezza investono una gamma di valori morali molto ampia, i quali possono essere sia allineati sia in conflitto con altri valori morali sui quali la riflessione etica pone l’accento, come la privacy, l’imparzialità o la giustizia, la prevenzione del danno e la costruzione di sistemi inclusivi e ampiamente utilizzabili.

Capitolo 7

Armi autonome e controllo umano significativo

Viola Schiaffonati, Guglielmo Tamburrini

# Introduzione

Questo capitolo offre una panoramica delle principali questioni etiche relative ai sistemi d’arma autonomi, alla loro progettazione e produzione, al dispiegamento e all’eventuale impiego in azioni belliche. Dopo aver discusso la questione preliminare di che cosa sia un sistema d’arma autonomo (o, come diremo più semplicemente, un’arma autonoma), il capitolo si concentra sulle problematiche etiche che questi sistemi pongono nello svolgimento di un conflitto armato (*jus in bello*): rispetto del dettato della teoria della guerra giusta (Walzer, 1977) e del diritto internazionale umanitario (in breve, DIU), attribuzione di responsabilità nelle catene del comando militare, rispetto della dignità delle vittime potenziali di un’arma autonoma. Trasversale e decisiva per l’analisi di tali questioni etiche è l’incertezza epistemica che limita la capacità di prevedere il comportamento delle armi autonome. Si introduce poi la nozione di controllo umano significativo su *tutti* i sistemi d’arma, intesa come strumento adatto ad aggirare proficuamente il problema di fornire una definizione adeguata - che non sia troppo restrittiva o inclusiva - delle armi autonome ai fini di un approfondimento del dibattito sulle questioni etiche e di eventuali conseguenti richieste di regolamentazione internazionale delle armi autonome. A tale nozione è dedicata l’ultima parte del capitolo, dove si introduce la distinzione tra politiche uniformi di controllo umano significativo e politiche differenziate in base a varie tipologie di armi autonome. Ci soffermiamo, infine, su una proposta di politica differenziata e prudenziale, che delinea una soluzione alle questioni etiche presentate in questo capitolo dalle diverse prospettive dell’etica dei doveri e dell’etica delle conseguenze.

# Cosa sono le armi autonome

La robotica ha tradizionalmente contribuito a sviluppare armamenti e sistemi di difesa. Fra questi si possono annoverare sia i droni militari sia i robot telecomandati per la ricerca e la neutralizzazione di mine ed esplosivi improvvisati. Lo sviluppo e l’impiego di questi sistemi in ambito bellico è stato raramente oggetto di critiche dal punto di vista etico e legale. La ragione principale sta nel fatto che il loro impiego è stato sempre sotto il controllo degli esseri umani. Le eventuali controversie sono state confinate a una valutazione della loro accettabilità in specifiche condizioni, come per esempio l’uso dei droni per le cosiddette esecuzioni mirate. Molto invece è cambiato negli ultimi anni con l’entrata in scena dei sistemi d’arma autonomi. Per essere considerato autonomo un sistema d’arma deve avere la capacità di selezionare e attaccare un obiettivo senza ricorrere a nessun intervento da parte di un operatore umano dopo la sua attivazione (Dipartimento della Difesa, 2012; ICRC, 2016). A partire da questo requisito, necessario per considerare un’arma come genuinamente autonoma, sono state introdotte ulteriori distinzioni e classificazioni. In (Boulanin, Verbruggen, 2017), per esempio, si distingue tra le seguenti categorie di armi autonome: a) sistemi di difesa antiaerea, b) sistemi di protezione attiva dei veicoli blindati da missili e razzi anticarro, c) sentinelle robotiche per il pattugliamento di determinate zone, d) munizioni guidate di tipo *fire and forget*, lanciate da un velivolo per individuare e colpire autonomamente obiettivi di determinate tipologie all’interno di un’area preassegnata, e) le munizioni *loitering* (in italiano potrebbero dirsi armi capaci di vagabondare o girovagare) che pattugliano, anche per molte ore consecutive, una certa area in cerca di obiettivi da attaccare e distruggere.

Per tenere dietro alla rapida evoluzione delle attività di ricerca e sviluppo in questo settore, la classificazione delle armi autonome è in costante espansione e revisione. Negli ultimi anni, per esempio, è stata sfruttata la cosiddetta “intelligenza dello sciame” (*swarm intelligence*) per costruire sciami di armi autonome di piccola taglia e a basso costo. L’intelligenza dellosciame si manifesta negli schemi di comportamento coordinato di uno sciame di insetti, uno stormo di uccelli, o un branco di pesci. Alla base delle rapide e spesso sorprendenti variazioni di forma e di traiettoria di uno stormo di uccelli, non c’è un sistema centralizzato di controllo che governi e coordini i movimenti di tutti i membri dello stormo. Ciascuno decide come muoversi in base a una regola che tiene conto solo dei movimenti degli uccelli vicini. Il comportamento collettivo dello stormo emerge dalla composizione e dall’applicazione concorrente, da parte di ciascuno, di poche regole locali di controllo. Ci si aspetta che le armi autonome che sfruttano questa idea, come nel progetto pilota Perdix del Dipartimento della Difesa statunitense (<https://www.popularmechanics.com/military/aviation/a24675/pentagon-autonomous-swarming-drones/>) riescano a sopraffare le difese nemiche grazie al loro ingente numero (Verbruggen, 2019).

La transizione dai droni militari che degli operatori umani controllano in remoto ai velivoli senza pilota a bordo che hanno capacità autonome di navigazione, perlustrazione e attacco è un altro sviluppo incoraggiato da spinte operative, economiche e strategiche. Poiché la cabina di pilotaggio di un drone può essere dislocata a migliaia di chilometri di distanza dal teatro delle operazioni, il tempo di latenza dei segnali di controllo inviati dal pilota a terra è generalmente troppo lungo perché il drone possa prevalere in un duello aereo con un caccia pilotato da un essere umano a bordo. Un velivolo autonomo senza pilota ha maggiori possibilità di successo in un combattimento aria-aria, poiché è svincolato dai comandi impartiti dall’operatore umano per le manovre e le procedure d’ingaggio. A queste motivazioni per l’autonomia, che riguardano l’operatività militare del sistema, si aggiungono motivazioni di carattere economico, che riguardano la riduzione del numero e del costo degli addetti che operano nella cabina di pilotaggio di un drone, oltre a quelle strategiche e politiche. Sono numerosi gli Stati che hanno assegnato un ruolo centrale all’autonomia dei sistemi d’arma nei loro piani strategici di sviluppo dei sistemi di difesa. Queste scelte aprono una competizione sulle armi autonome tra Stati Uniti, Russia, Cina e altre nazioni impegnate a incrementare la propria potenza militare (Boulanin, Verbruggen, 2017).

# Il problema della definizione

Cerchiamo ora di capire meglio che cosa effettivamente sia l’autonomia dei sistemi d’arma. Sicuramente il punto di partenza è il requisito che un’arma autonoma debba essere in grado di selezionare e attaccare un obiettivo militare senza l’intervento dell’operatore umano. Nella direttiva del dipartimento di Difesa degli Stati Uniti sull’autonomia dei sistemi d’arma vengono indicate come fondamentali per l’autonomia di un’arma proprio queste caratteristiche (Dipartimento della Difesa, 2012), considerate fondamentali anche in un documento successivo stilato dal Comitato internazionale della Croce Rossa (ICRC, 2016). In altre parole, selezionare e attaccare gli obiettivi in modo indipendente, significa godere di autonomia nelle cosiddette funzioni *critiche* per l’autonomia di un’arma, ossia acquisire, tracciare, selezionare e ingaggiare un obiettivo senza intervento umano. L’identificazione di tali funzioni ha costituito una base di partenza largamente condivisa per l’analisi delle questioni etiche e giuridiche sollevate dalle armi autonome. Bisogna però osservare che si tratta una base poco stringente per il dibattito etico, giuridico, politico e diplomatico, soprattutto quando l’obiettivo è di valutare se bisogna introdurre eventuali restrizioni o proibizioni d’uso su *tutte* le armi autonome. Vediamo perché.

Innanzitutto, esistono sistemi d’arma che rispondono al requisito sopra evidenziato e che sono attualmente usati dalle forze armate di diversi stati, senza suscitare obiezioni di carattere etico o giuridico. Limitandosi a considerare l’autonomia come la capacità operativa di selezionare e attaccare un obiettivo, perfino una mina antiuomo potrebbe essere considerata un’arma autonoma dato che è in grado di discriminare fra la pressione esercitata dal passaggio di un piccolo animale, insufficiente a farla esplodere, e la pressione esercitata dal corpo di un essere umano o da altro oggetto di peso analogo o superiore che, invece, provoca l’esplosione senza l’intervento di un operatore umano. L’identificazione delle necessarie funzioni critiche di un’arma autonoma non basta a tirare una linea chiara per distinguere fra i sistemi d’arma progettati in base alle più avanzate tecnologie di IA e robotica - i quali, come vedremo, sollevano nuovi problemi etici e giuridici - e l’autonomia operativa di sistemi molto più rudimentali, come le mine antiuomo, i cui problemi etici sono già stati riconosciuti e affrontati nelle sedi preposte.

Se i tentativi analizzati finora possono essere visti come troppo permissivi, se lo scopo è quello di arrivare a una vera e propria definizione di autonomia, nella direzione opposta di una concezione troppo restrittiva si colloca la proposta dal ministero della Difesa Britannico (UK MOD, 2017). In base a questa proposta, un sistema d’arma autonomo deve essere in grado di comprendere intenzioni e direttive di alto livello, e di agire in base a esse per raggiungere lo stato desiderato. La criticità di tale definizione risiede nel fatto che nessuna macchina esistente, o concepibile secondo le conoscenze attuali, ha la capacità di comprendere le intenzioni e le direttive di alto livello di un essere umano, anche quando sia in grado di selezionare e attaccare un obiettivo. In questo modo ogni arma esistente o progettabile in un futuro prossimo sarebbe esclusa dal novero delle armi autonome. In definitiva, il tentativo del ministero della Difesa Britannico proietta l’esistenza delle armi autonome in un futuro indeterminato. Questo tentativo è dunque dannoso per promuovere un dibattito costruttivo, nel quale è importante riconoscere l’urgenza delle questioni etiche e giuridiche poste. I protagonisti del dibattito internazionale sulle armi autonome non hanno accolto favorevolmente questa definizione troppo restrittiva e si sono inizialmente basati, in assenza di alternative migliori, sull’identificazione delle funzioni critiche enucleate congiuntamente dal Dipartimento della Difesa statunitense e dal Comitato Internazionale della Croce Rossa.

# Il dibattito sulle armi autonome fra etica e diritto

Il dibattito etico e legale sulle armi autonome ha preso avvio dalle riflessioni di alcuni dei membri della comunità robotica. In particolare, Ronald Arkin e Noel Sharkey hanno discusso rispettivamente i pro e i contro della loro adozione. Arkin ha sottolineato come i robot siano immuni dalle necessità di autoconservazione e dalle reazioni emotive che regolano il comportamento umano, le quali possono indurre un combattente a violare norme morali e giuridiche che regolano la condotta delle parti belligeranti. La sua ricerca in questo ambito è guidata, infatti, dall’idea che i robot debbano essere progettati in modo tale da rispettare *by design* le norme del diritto internazionale umanitario (Arkin, 2009). In contrasto con questa visione, Sharkey ha enfatizzato gli aspetti critici dell’autonomia nei sistemi d’arma, sostenendo che lo sviluppo tecnologico della robotica e dell’IA non può offrire al momento alcun supporto all’idea che le armi autonome possano meglio conformarsi alle norme internazionali. Al contrario, le interazioni fra diverse armi autonome sono difficilmente prevedibili in contesti di guerra non strutturati e così veloci da rendere il loro ritmo fuori dal controllo umano. Sharkey ha contribuito a fondare il Comitato Internazionale per il Controllo delle Armi Robotiche (*International Committee for Robot Arms Control* (ICRAC)) e ha svolto un ruolo fondamentale per l’avvio della campagna internazionale *Stop Killer Robots*. Quest’ultima è un’iniziativa guidata da una coalizione di organizzazioni non governative, fondata nel 2012 con l’obiettivo di vietare l’impiego di sistemi d’arma autonomi capaci di decidere ed eseguire azioni letali contro gli esseri umani. Il dibattito internazionale sulla regolamentazione delle armi autonome si è poi sviluppato soprattutto presso la sede delle Nazioni Unite a Ginevra, a partire dal 2014 e nella cornice della *Convention on Conventional Weapons* (CCW), il cui scopo principale è di restringere, e possibilmente vietare, l’uso di quelle armi che possono causare sofferenze non necessarie e non giustificabili ai combattenti oppure avere un impatto indiscriminato sulla popolazione civile.

# Etica dei doveri e armi autonome

Fin dagli inizi, il dibattito sull’ammissibilità morale e giuridica delle armi autonome ha invocato principi connessi all’etica dei doveri o all’etica delle conseguenze. In questo paragrafo ci concentriamo ad analizzare in modo sistematico le riflessioni etiche sulle armi autonome che prendono come riferimento la cornice dell’etica dei doveri.

# Principi di distinzione e proporzionalità in guerra

Sono riconducibili all’etica dei doveri i principali argomenti contro l’uso delle armi autonome, in quanto possibili fonti di violazione di obblighi fondamentali che, invece, devono essere rispettati da tutti gli attori coinvolti in un’azione bellica. Questi obblighi sono fondamentalmente tre:

1. Il rispetto del dettato della teoria della guerra giusta e delle norme del DIU;
2. Il mantenimento della catena delle responsabilità nelle azioni belliche;
3. Il rispetto della dignità degli esseri umani coinvolti nelle azioni belliche.

Per quanto riguarda la teoria della guerra giusta e il DIU, due principi giocano un ruolo centrale nella discussione etica e giuridica sull’ammissibilità delle armi autonome: *il principio di distinzione* e il *principio di proporzionalità*. Ricordiamo che il DIU limita l’impiego di armi e metodi di guerra e impone varie forme di tutela delle persone e delle cose durante i conflitti armati.

Il principio di distinzione fa riferimento al fatto che nei conflitti gli attacchi militari devono essere diretti solo a obiettivi militari che forniscono un contributo effettivo all’azione bellica e la cui distruzione offre un vantaggio militare. In particolare, i belligeranti non devono indirizzare un attacco contro persone che non siano coinvolte direttamente nell’azione bellica. Per applicare correttamente il principio di distinzione bisogna poter distinguere fra persone attivamente coinvolte nell’azione bellica da un lato, e civili o i nemici fuori combattimento. Inoltre, occorre salvaguardare il personale militare sanitario e religioso, le unità sanitarie e i relativi mezzi di trasporto civili e militari. Queste distinzioni non sono semplici da attuare: si pensi per esempio a come distinguere un nemico fuori combattimento, valutando se non sia nelle condizioni di nuocere perché ferito o in stato di incoscienza oppure se esprima chiaramente l’intenzione di arrendersi.

Proprio il compito di distinguere correttamente fra le varie categorie rilevanti si rivela un problema particolarmente ostico nel caso di sistemi d’arma che sono il frutto della ricerca in IA e robotica. Come è possibile per un sistema artificiale riconoscere, con prestazioni all’altezza di un soldato competente, un nemico che esprime la sua intenzione di arrendersi sulla base di comportamenti molto variegati e poco convenzionali? Come è possibile garantire che il sistema non compia errori percettivi grossolani che, invece, nel caso sono evitati dagli esseri umani in base alla loro condivisione di una conoscenza di senso comune? Si tratta di problemi per i quali non vi è ancora una soluzione tecnologica generale e soddisfacente, tanto più che essi nascono in un contesto, come quello bellico, per sua natura soggetto a cambiamenti dinamici difficilmente prevedibili. Il contesto bellico è ricco di sorprese, alcune delle quali intenzionalmente ricercate dagli avversari, altre che possono nascere anche non intenzionalmente a causa di interazioni non strutturate fra esseri umani e sistemi d’arma. Questi elementi rendono difficile la realizzazione di test empirici per valutare le prestazioni dei sistemi d’arma nelle specifiche situazioni sui campi di battaglia (si veda più oltre il paragrafo 4.2).

Non meno serio è il problema del riconoscimento percettivo dei sistemi d’arma che sono sviluppati mediante tecniche di apprendimento automatico (esaminate nel capitolo 2) e che sono sensibili al cosiddetto *adversarial testing*: piccole modifiche mirate nelle immagini portano a errori significativi nella classificazione delle immagini da parte dei sistemi artificiali, mentre sono del tutto irrilevanti per il sistema visivo umano. Un esempio è rappresentato dai fallimenti di un sistema di IA che utilizza tecniche di apprendimento automatico basate su reti neurali profonde (*deep learning*) per riconoscere un pulmino scolastico (che dovrebbe essere sempre escluso dalla lista degli obiettivi militari). Da alcuni test eseguiti sul sistema nel classificare immagini in ingresso per distinguere un pulmino scolastico, un cane, uno struzzo e altre categorie di oggetti presentati visivamente è emerso che, dopo piccole modifiche introdotte nell’immagine utilizzando il cosiddetto *adversarial testing*, il sistema ha classificato come struzzo la nuova immagine solo leggermente modificata (Szegedy et al, 2013). Ma un pulmino scolastico è protetto dal principio di distinzione mentre uno struzzo non lo è. I casi di *adversarial testing* rilevanti per il dibattito sulle armi autonome sono numerosi: piccole modifiche nell’immagine, irrilevanti per il sistema visivo umano, portano a classificazioni erronee da parte del sistema autonomo e significative sotto il profilo del principio di distinzione.

Il principio di proporzionalità prescrive che non si debbano sferrare attacchi che abbiano costi attesi eccessivi (sia in termini di morti e feriti fra i civili sia in termini di danni a strutture civili) rispetto al vantaggio militare concreto che ne può derivare. Tuttavia, non è un compito facile pesare i vantaggi militari rispetto ai costi per la popolazione civile. Un bilancio di questo tipo può richiedere l’uso di capacità cognitive ed emotive, di competenze sociali ed esperienziali che non sono attualmente alla portata neppure dei sistemi artificiali più sofisticati. Proviamo a ipotizzare la situazione di un comandante che, a una prima valutazione ritiene accettabile la perdita stimata di civili rispetto al presunto vantaggio militare, ma dopo più attenta riflessione basata su esperienze passate condivise con i propri soldati, arriva a concludere che la perdita di civili avrebbe un effetto negativo e duraturo sul loro morale. Il comandante potrebbe decidere così di non lanciare l’attacco, valutando come determinante l’effetto negativo morale dei propri soldati. Come si vede una soluzione tecnologica a questo problema non appare possibile nell’immediato, dato che il comandante ha utilizzato per la sua valutazione competenze sociali ed emotive e delle valutazioni di buon senso difficilmente trasferibili a un sistema d’arma autonomo (Amoroso, Tamburrini, 2017).

Secondo alcuni ricercatori un sistema d’arma autonomo non ha, né ora né in un futuro prevedibile, le capacità necessarie per discriminare fra soldati e civili e per decidere attacchi sulla base del principio di proporzionalità con prestazioni all’altezza di comandante militare competente (Sharkey, 2010).

# Responsabilità e dignità dell’uomo

Supponiamo ora, ai fini dell’argomentazione, che sufficienti prove sperimentali arrivino in futuro a mostrare che un sistema d’arma autonoma è in grado di rispettare le norme della teoria della guerra giusta e del DIU almeno quanto un soldato competente. In questo caso, sarebbe ancora possibile opporsi allo sviluppo e all’uso di armi autonome facendo appello all’obbligo di mantenere la catena delle responsabilità nelle decisioni di attacco e di rispettare la dignità degli esseri umani coinvolti in un conflitto.

Consideriamo, ancora una volta, l’*adversarial testing* prima menzionato, e supponiamo che un’arma autonoma colpisca, insieme a una installazione militare del nemico, un pulmino scolastico che sta passando nelle vicinanze, perché l’ha scambiato per uno struzzo. L’attacco non è giustificabile in base al principio di proporzionalità, perché la distruzione dell’installazione militare non procura un vantaggio militare sufficiente a controbilanciare la morte delle persone trasportate dal pulmino. Chi dovrebbe essere ritenuto responsabile? L’arma autonoma non può esserlo, poiché non è un agente morale. La responsabilità va cercata fra le persone che hanno preso parte alle decisioni per attivare l’arma e sono sfociate nell’attacco al pulmino: ingegneri informatici e robotici e di altri ambiti che hanno contribuito allo sviluppo dell’arma; i responsabili dell’azienda produttrice; il personale politico che ha dato l’impulso all’acquisto dell’arma autonoma da parte delle forze armate; i consulenti e dipendenti del Ministero della Difesa che hanno proceduto all’acquisizione; i militari che hanno redatto il manuale d’uso, il comandante in capo all’operazione militare, il personale militare incaricato di supervisionare il funzionamento dell’arma dopo la sua attivazione. Il problema di determinare la responsabilità quando la catena di comando è così lunga e ramificata è noto anche come “problema delle molte mani” ed è stato già ampiamente analizzato in relazione al malfunzionamento dei sistemi informatici: quando sono numerose le persone che hanno contribuito a vario titolo allo sviluppo del sistema, si può arrivare a concludere che nessuna di esse ha dato un contributo davvero significativo al verificarsi dell’incidente (Nissenbaum, 1996). Anche nel caso ipotetico appena considerato, ogni attore coinvolto concorre al verificarsi dell’evento. Ma è problematico ritenere qualcuno di essi individualmente responsabile dell’attacco al pulmino scolastico, un’azione indotta da un errore madornale di classificazione per il sistema percettivo umano e che si potrebbe classificare come crimine di guerra nel caso in cui l’attacco fosse stato portato avanti da esseri umani.

Anche quando non insorge il problema delle molte mani, la complessità delle tecnologie utilizzate nelle armi autonome, la difficoltà di prevedere il loro comportamento sul campo di battaglia e la loro interazione con gli agenti umani o artificiali con cui hanno a che fare rende difficile accertare la presenza dell’elemento mentale (intenzione, conoscenza o imprudenza) che il DIU per attribuire delle responsabilità criminali. Questo vuoto di responsabilità non è colmabile neppure dalla dottrina della *Responsabilità del comando*. Secondo tale dottrina i superiori, in ottemperanza ai loro doveri di comando, hanno l’obbligo di vigilare sulle azioni dei loro sottoposti, per evitare che essi compiano azioni deprecabili e contrarie al DIU. Tuttavia, non è ovvio che un’assunzione del genere sia valida nel caso di un’arma autonoma, dato che il comandante che ha ordinato l’attivazione dell’arma potrebbe non essere in grado di esercitare in seguito un controllo efficace sul suo comportamento, dovuto anche ai tempi lunghi di reazione degli esseri umani che potrebbero rivelarsi inefficaci per un intervento in tempo utile. È evidente che le armi autonome rappresentano un problema inedito e di difficile soluzione nel processo di attribuzione delle responsabilità.

Nella cornice dell’etica dei doveri c’è un ulteriore argomento che viene utilizzato contro l’impiego di armi autonome. L’argomento fa riferimento al rispetto della dignità umana e alle violazioni che potrebbero derivare affidando a una macchina il potere di vita o di morte su un essere umano. Nel caso di un’arma autonoma, infatti, una vittima potenziale non ha la possibilità di fare appello all’umanità condivisa di qualcuno che si trovi dall’altra parte, il quale possa apprezzare il valore della vita umana in gioco, provare empatia o compassione. Il riconoscimento del valore intrinseco di un essere umano sarebbe perciò negato a priori se si affidasse il suo destino alla discrezionalità di una macchina (Heyns, 2016).

# Etica delle conseguenze e armi autonome

Se l’etica dei doveri fornisce principalmente argomenti rivolti contro le armi autonome, l’etica delle conseguenze offre un panorama di argomenti sia in favore sia contro le armi autonome. Le principali motivazioni a sostegno del futuro impiego delle armi autonome offerte dalla prospettiva dell’etica delle conseguenze sono le seguenti:

1. Le armi autonome potranno garantire maggiore precisione nell’attacco di obiettivi militari, in uno scenario futuro in cui avranno dimostrato capacità di giudizio e di classificazione percettiva almeno uguali a quelle di soldati ben addestrati;
2. Le armi autonome potranno sostituire i soldati diminuendo così le vittime umane nei conflitti ed evitando violazioni del DIU dovute allo stress cui sono sottoposti i soldati nelle azioni belliche;
3. Le armi autonome potranno essere programmate con regole di ingaggio più restrittive di quelle pensate per un comune soldato, dato che il valore della vita di quest’ultimo non può essere paragonato a quello di una macchina; inoltre, un’arma autonoma potrà essere programmata per sferrare attacchi solo quando non vi siano dubbi sulle intenzioni ostili del nemico.

È sulla base di queste considerazioni che alcuni sostengono che non vi siano ragioni sufficienti per passare da una eventuale moratoria temporanea, attualmente giustificabile sulla base della distanza fra le condizioni a), b) e c) e l’attuale sviluppo scientifico e tecnologico, a una messa al bando una volta per tutte delle armi autonome.

L’eventuale riduzione delle vittime nelle azioni belliche in cui siano impiegati armi autonome è un fattore da prendere in considerazione anche dalla prospettiva dell’etica dei doveri, poiché il rispetto degli obblighi deontologici esaminati nel paragrafo 2 deve essere contemperato con il rispetto di altri obblighi morali, come la riduzione delle vittime attese. Da un altro lato, non si può ritenere che le considerazioni in a), b) e c) offrano un sostegno incondizionato all’impiego di armi autonome dal punto di vista dell’etica delle conseguenze, poiché esse riguardano i singoli campi di battaglia e l’azione di ciascuna arma autonoma. È interessante notare che, quando si amplia la visione degli ipotetici benefici delle armi autonome al di là dei singoli campi di battaglia, occorre prendere in considerazione anche altre possibilità nel bilancio dei costi e dei benefici: la possibilità di un ricorso più facile alle armi, rischi di malfunzionamento e hackeraggio malevolo, nuovi incentivi per una corsa alle armi, rischi di destabilizzazione regionale e globale. Un calo di perdite umane attese nel proprio esercito potrebbe indebolire l’ostilità dell’opinione pubblica ai conflitti armati, incentivando così i decisori politici a iniziare nuove guerre. Allo stesso tempo, perturbazioni al corretto funzionamento delle armi autonome derivanti da interazioni impreviste di un’arma autonoma con altri agenti umani o artificiali, possono vanificare le aspettative di diminuzione delle vittime, per esempio provocando episodi di fuoco amico. Analogamente, uno sguardo più ampio sulle conseguenze attese deve tenere in conto di come le armi autonome possono generare una nuova corsa agli armamenti. Le armi autonome sono anche sistemi collegati in rete, e per questo motivo esposti a cyberattacchi e tentativi di hackeraggio (si veda il capitolo 6). Non da ultimo l’introduzione delle armi autonome può comportare il rischio di accelerare il ritmo dei conflitti, fino al punto diventare incompatibile con i tempi di reazione e di controllo da parte degli operatori umani (Altmann, Sauer, 2017).

# Controllo umano significativo e armi autonome

Alla luce delle questioni etiche fin qui sviscerate, passiamo a considerare più approfonditamente la centralità dell’elemento umano, e più precisamente il tipo di interazione tra esseri umani e macchine autonome ammissibile sul piano etico nelle operazioni belliche, e di come l’idea di responsabilità in questo contesto debba essere riconsiderata alla luce delle incertezze epistemiche tipiche dei contesti bellici.

# Controllo umano significativo

A partire dalla preoccupazione di un’interruzione della catena di responsabilità nelle azioni belliche, già da qualche anno si è posta l’attenzione sulla centralità dell’elemento umano nella riflessione sulle armi autonome. È in questo contesto che è stata proposta l’idea del *controllo umano significativo* (CUS) come condizione necessaria negli attacchi sferrati da *tutti* i sistemi d’arma (Article 36, 2013). L’uso dell’aggettivo “significativo” è volto a escludere condizioni e forme di controllo puramente nominali di un sistema d’arma, il che potrebbe accadere quando il controllore umano non ha a disposizione tempo o informazioni sufficienti per intervenire con cognizione di causa sul sistema d’arma. La formula del controllo umano significativo è stata accolta con favore dai molti attori del dibattito sulle armi autonome in quanto può essere facilmente compresa anche in assenza di conoscenze specifiche sulle tecnologie militari più avanzate. Sebbene tale formula non specifichi in che cosa consista un controllo veramente significativo, essa ha il merito di spostare il fulcro della discussione da una questione definitoria (“*Che cos’è un’arma autonoma?*”) a una normativa (“*Quale controllo umano si deve esercitare su ogni sistema d’arma?*”).

Ma cosa rende il controllo umano su un’arma veramente significativo? Per cercare di rispondere a questa domanda riconsideriamo i principi dell’etica dei doveri che abbiamo presentato nel paragrafo 2.1: il rispetto delle norme della teoria della guerra giusta e del DIU; il mantenimento della catena delle responsabilità nelle azioni belliche; il rispetto della dignità degli esseri umani coinvolti nelle azioni belliche. Per garantire tutto ciò è necessario assegnare all’essere umano alcuni ruoli fondamentali nel controllo dei sistemi d’arma (Amoroso, Tamburrini, 2019, 2021; Amoroso 2020):

* Il controllo umano deve intervenire come *sistema ausiliario di salvaguardia* per cercare di impedire che il malfunzionamento o un comportamento imprevisto del sistema d’arma sfoci in un attacco contro la popolazione civile e i suoi beni, in un danno collaterale eccessivo o in altre violazioni del DIU;
* Il controllo umano deve funzionare come *catalizzatore di responsabilità* assicurando la sussistenza delle condizioni per attribuire delle responsabilità personali nel caso di violazioni del DIU;
* Il controllo umano deve essere *garante del rispetto della dignità degli esseri umani* sottoposti a decisioni di vita o di morte in un conflitto armato, consentendo di ricondurre a un agente morale umano tutte le decisioni che hanno un impatto sulla loro vita, sulla loro integrità fisica e sui loro beni.

Per svolgere la funzione di sistema ausiliario di salvaguardia, un controllore umano deve essere in grado di esprimere un giudizio ponderato sulle operazioni che il sistema d’arma sta compiendo, o si accinge a compiere, e di poter intervenire in tempo utile in caso di imprevisti. Ciò significa che l’operatore deve acquisire familiarità con i limiti del sistema d’arma, le cause note di malfunzionamento e le possibili fonti di interferenza; deve essere addestrato a non sovrastimare le capacità dei sistemi informatici, senza sottostare al cosiddetto pregiudizio positivo da automazione che induce a non contraddire le conclusioni alle quali è giunto un sistema artificiale ritenendolo non soggetto a errori. Nel caso dei più recenti sistemi d’arma che utilizzano tecnologie di IA si aggiungono altri problemi, quali la difficoltà di *interpretare* le informazioni che inducono il sistema artificiale a prendere una certa decisione o a compiere una determinata azione; la difficoltà di ottenere dalla macchina delle *spiegazioni*, in modo che l’operatore umano possa formarsi un giudizio ponderato sulla bontà delle decisioni della macchina stessa. Le richieste di interpretabilità e di spiegabilità nascono dall’esigenza di garantire un controllo umano veramente significativo sui sistemi d’arma. Tuttavia, queste richieste devono fare i conti con sfide tecnologiche molto impegnative, soprattutto alla luce dei più recenti sviluppi dell’IA nel settore dell’apprendimento automatico. Questi sviluppi esasperano ulteriormente il *fattore di invisibilità* da cui nascono questioni etiche rilevanti (Moor, 1985): le operazioni interne dei computer sono invisibili, solo input e output sono visibili. In particolare, l’*invisibilità dei valori di programmazione* e l’*invisibilità dei processi di calcolo complessi* sembranoessere elementi decisivi nelle tecnologie dell’apprendimento automatico che giocano un ruolo essenziale nello sviluppo dei sistemi ad autonomia crescente, come le armi autonome.

# Incertezze epistemiche e responsabilità

Le armi autonome dipendono sempre più da architetture basate su apprendimento automatico. Le tecnologie dell’apprendimento automatico hanno allargato notevolmente le frontiere dei compiti che le macchine possono imparare a svolgere, invece di essere programmate esplicitamente a farlo (si veda il capitolo 2). Inoltre, le armi autonome operano in ambienti caratterizzati da un alto grado di incertezza sia per la natura stessa delle operazioni belliche sia per la complessità delle interazioni fra agenti artificiali e agenti umani. Tutto ciò rende difficile valutare le capacità effettive dei sistemi d’arma autonomi e i risultati delle loro azioni nelle singole azioni belliche. Tale incertezza epistemica ha serie ricadute sulla possibilità di ascrivere in modo chiaro la responsabilità umana. Per cercare di chiarire i doveri di controllo del comandante umano tenendo conto dell’incertezza epistemica è stata proposta la nozione di *tracciamento robusto del controllo* (Himmelreich, 2019): un agente *a* ha controllo sul fatto che un risultato *x* si verifichi se 1) c’è un ordine che *a* può dare e che 2) se *a* desse questo ordine, allora *x* si verificherebbe (in tutte le situazioni simili), e 3) se *a* non dovesse dare questo ordine, allora *x* non si verificherebbe (in tutte le situazioni simili). È importante notare che, per poter inferire che un risultato *x* di un sistema d’arma autonoma è sotto il tracciamento robusto del controllo, il comandante militare deve poter asserire con un alto grado di fiducia che l’operazione militare in atto è *simile* in tutti gli aspetti rilevanti al suo modello cognitivo di un’operazione militare e dei risultati attesi. Naturalmente il grado di fiducia, in questo giudizio di somiglianza, è destinato a diminuire al crescere dell’incertezza che caratterizza il modello cognitivo del comandante e i contesti altamente destrutturati delle operazioni militari. In questo scenario, l’autonomia dei sistemi d’arma diventa incompatibile con la possibilità di preservare il tracciamento robusto del controllo.

Quanto discusso finora relativamente all’incertezza epistemica che caratterizza i sistemi d’arma autonoma e i contesti in cui operano si può considerare anche come uno specifico problema di *fragilità* delle applicazioni di IA in contesti critici dal punto di vista della sicurezza (*safety-critical*), come i trasporti, la medicina e le azioni belliche. In tali contesti, i sistemi di IA che utilizzano le tecnologie dell’apprendimento automatico non hanno quelle capacità di generalizzazione che sono necessarie per gestire l’incertezza (Cummings, 2020). Un sistema che utilizza tecniche di apprendimento automatico per percepire l’ambiente è spesso incapace di adattarsi a condizioni che si collocano al di fuori di un insieme ristretto di assunzioni e, quindi, di poter generalizzare di fronte all’incertezza che caratterizza tale ambiente. Il caso del sistema di visione di un’arma autonoma che, a fronte di una minima perturbazione dell’immagine, compie errori percettivi madornali (uno struzzo al posto di un pulmino) rientra in questa nozione di fragilità. Il tentativo di raccogliere e utilizzare più dati per far fronte a quello che è ritenuto un problema di tipo esclusivamente percettivo non è del tutto risolutivo. In altre parole, non si tratta solo di raccogliere più dati per rendere più robusto il processo cosiddetto *bottom-up*, ma di includere anche il ragionamento *top-down* o basato sulla conoscenza in questi sistemi. Tale forma di ragionamento, propria dell’IA classica basata sui modelli, gioca un ruolo fondamentale per realizzare le competenze causali che sono essenziali perché un sistema artificiale riesca ad agire appropriatamente in situazioni nuove. Perché i sistemi critici dal punto di vista della sicurezza possano operare efficacemente in ambienti realistici, caratterizzati da un alto grado di incertezza, occorre che siano dotati di capacità di ragionamento basate sia sui modelli (*top-down*) sia sui dati (*bottom-up*). Tuttavia, ciò non significa che tali sistemi di IA siano in grado di sostituire gli esseri umani nel loro operare, quanto piuttosto che gli esseri umani devono poter sempre esercitare un controllo significativo sulle operazioni di tali sistemi, in particolare in quegli ambienti caratterizzati da un alto grado di incertezza.

Se da un lato lo sviluppo di tecnologie appropriate costituisce un elemento importante per la gestione dell’incertezza, dall’altro lato bisogna riconoscere la necessità di un cambio di paradigma a livello sia epistemico sia etico. In questa direzione sono state avanzate alcune proposte, che enfatizzano la centralità di approcci incrementali per valutare l’impatto di tecnologie di questo tipo: dal punto di vista epistemico è stata proposta la nozione di *esperimento esplorativo* per rendere conto dell’incertezza epistemica che caratterizza la sperimentazione su sistemi di IA e robotica e il superamento del paradigma dell’esperimento controllato (Amigoni, Schiaffonati, 2021); dal punto di vista etico è stata avanzata l’idea di una nuova cornice etica per le cosiddette tecnologie sperimentali, per le quali l’esperienza relativa al loro effettivo comportamento è limitata così da rendere incerti i tentativi di valutare i loro rischi e benefici sociali (van de Poel, 2016).

# Politiche etiche per le armi autonome

L’incertezza epistemica che caratterizza sia i sistemi d’arma autonoma sia i contesti in cui essi operano ha un impatto decisivo anche sulle politiche etiche per le armi autonome ispirate dall’idea del CUS.

# Politiche etiche uniformi per il controllo umano

Secondo alcuni è possibile adottare una ricetta uniforme per il CUS adeguata a gestire tutti i sistemi d’arma in tutte le situazioni d’uso. Una prima proposta, che possiamo chiamare “soluzione olandese” perché avanzata in un documento del governo olandese, ritiene che il CUS possa essere esercitato esaurientemente nella fase di pianificazione dell’azione militare, prima di attivare l’arma autonoma (Ekelhof, 2019). Una diversa proposta è quella che si potrebbe denominare ad “autonomia zero”, poiché impedisce l’attivazione delle armi autonome in qualsiasi situazione e ne richiede la totale messa al bando.

Consideriamo, in relazione alla soluzione olandese, alcune delle armi autonome già operative, in particolare le armi offensive che utilizzando munizioni *loitering*, come il missile israeliano Harpy: velivoli senza equipaggio e armati di una testata esplosiva in grado di vagabondare (*to loiter*) per un certo periodo di tempo alla ricerca di un obiettivo da attaccare all’interno di un certo perimetro. L’operatore definisce non solo l’area di azione, ma anche le categorie di obiettivi che l’arma può attaccare e la durata massima del vagabondaggio. Questi tipi già esistenti di armi sono sufficienti per mostrare che la soluzione uniforme olandese non può assicurare il mantenimento del CUS. Tale soluzione può funzionare solo nel caso di obiettivi militari individuati al momento di pianificare l’azione militare e in situazioni che non evolvono in maniera imprevista successivamente. Per esempio, mentre Harpy vagabonda sull’area prestabilita, potrebbe avvicinarsi a una postazione radar nemica un veicolo per il trasporto di civili. Un attacco da parte dell’arma alla postazione radar potrebbe avere conseguenze incompatibili con i principi di distinzione e di proporzionalità e i danni provocati sarebbero difficilmente riconducibili alle responsabilità degli operatori umani, dato che il contesto operativo è cambiato in modo imprevisto dopo l’attivazione del sistema.

Consideriamo ora la soluzione uniforme ad autonomia zero. La sua inadeguatezza emerge considerando alcuni sistemi puramente difensivi, come il sistema israeliano Iron Dome e il sistema tedesco MANTIS, che funzionano come scudi protettivi, i quali sono dotati della capacità di reagire a missili o proiettili balistici in arrivo e attaccarli autonomamente. La loro utilizzazione in modalità autonoma serve per fornire risposte difensive in tempi assai inferiori a quelli di un operatore umano. Sebbene tali sistemi difensivi non abbiano mai suscitato obiezioni di carattere etico o giuridico, il loro utilizzo è incompatibile con la soluzione ad autonomia zero. Inoltre, sembra possibile mantenere il CUS su questi sistemi senza privarli di autonomia, affidando a un operatore umano di decidere periodicamente se vi siano oggetti o persone da proteggere nel campo di azione dell’arma.

# Politiche etiche differenziate e prudenziali

I problemi che abbiamo evidenziato nel caso delle politiche uniformi suggeriscono di esplorare ora politiche differenziate al problema del CUS, che tengano in conto diverse tipologie di armi autonome e situazioni d’uso. Per introdurre l’idea stessa di un CUS differenziato, è utile richiamare una lista di opzioni per il controllo condiviso dell’azione fra uomo e macchina, ordinata in base ai privilegi di controllo che progressivamente si spostano dall’essere umano (che ha il controllo totale sulle funzioni critiche in L1) alla macchina (che ha il controllo totale in L5) (Sharkey, 2016).

* L1: la selezione dell’obiettivo da attaccare è integralmente effettuata dall’operatore umano;
* L2: la selezione dell’obiettivo da attaccare è effettuata dall’operatore umano in base a un ventaglio di opzioni suggerite dal sistema d’arma:
* L3: l’operatore umano si limita ad approvare o rifiutare la scelta del sistema d’arma in merito all’obiettivo da attaccare;
* L4: l’operatore umano supervisiona la selezione dell’obiettivo effettuata dal sistema, mantenendo la possibilità di riprendere il controllo e interrompere l’attacco;
* L5: l’operatore umano si limita ad attivare il sistema d’arma, definendone la missione nella fase preliminare di pianificazione, senza avere la possibilità di intervenire in fase operativa.

Per ovviare ai difetti evidenziati a proposito di entrambe le soluzioni uniformi precedentemente discusse, proviamo a considerare una particolare soluzione che è *differenziata*, perché modula il controllo umano lungo i cinque livelli (L1-L5), e insieme *prudenziale*, poiché privilegia tra questi i livelli più bassi allo scopo di garantire il mantenimento del CUS (Amoroso, Tamburrini 2021). Questa soluzione impone per default si livelli L1 e L2 di controllo umano, in assenza di informazioni specifiche che indichino il contrario (*aspetto prudenziale della soluzione*). Ma questa soluzione ammette anche eccezioni, che consentono di andare oltre L2 nel caso di specifici sistemi d’arma e modalità d’uso, a patto che il CUS non sia minacciato e le eccezioni siano di volta in volta approvate dalla comunità internazionale degli stati nella sede della Convention on Conventional Weapons dell’ONU o in altre sedi opportune. Affinché il CUS non sia minacciato, le sue funzioni eticamente fondate di sistema ausiliario di salvaguardia, catalizzatore di responsabilità e garante del rispetto della dignità degli esseri umani coinvolti, discusse nel paragrafo 4.1, devono essere sempre adeguatamente preservate. Le eccezioni concordate a queste condizioni comportano i privilegi di controllo vengano opportunamente spostati verso la macchina (*aspetto differenziato della soluzione*).

Le motivazioni principali per sostenere la necessità di dare priorità a una severa regola di default (livelli L1-L2) derivano dalla limitata capacità di prevedere e controllare il comportamento dei sistemi d’arma sviluppati nell’ambito dell’IA e della robotica, i quali danno luogo all’incertezza epistemica discussa nel paragrafo 4.2. Questi limiti dipendono dalle capacità stesse implementate nei sistemi d’arma e dalle loro interazioni con l’ambiente operativo. Abbiamo brevemente considerato, a questo proposito, la capacità di apprendere e le sue limitazioni, la capacità di vagare (*loitering*) per un periodo esteso alla ricerca di un obiettivo in aree soggette a cambiamenti difficilmente prevedibili, la capacità di coordinarsi e comunicare con altri sistemi in assenza di controllo umano, le interazioni con altri sistemi artificiali su un campo di battaglia destinato a divenire sempre più informatizzato e robotizzato. È in questo contesto che nel paragrafo 4 abbiamo menzionato la necessità di un cambio di paradigma che tenga conto dell’incertezza epistemica e che sia in grado di offrire soluzioni incrementali.

Le eccezioni rese possibili da questo esempio prudenziale di politica differenziata dovranno essere subordinate all’elaborazione di opportuni argomenti e prove sperimentali convincenti che mostrino la possibilità di mantenere il CUS a un livello superiore a L2. Nel caso dei sistemi difensivi come Iron Dome o MANTIS (si veda paragrafo 5.1), è ragionevole affermare che il CUS possa essere adeguatamente mantenuto a livello L3, o perfino L4, così come di fatto avviene anche oggi. Ciò significa che un operatore umano deve sempre supervisionare l’attività del sistema e ha diritto di interromperla in fase operativa, per quanto l’efficacia del diritto di veto possa essere limitata dai tempi rapidi di risposta del sistema.

La rinuncia alla ricerca di una soluzione uniforme al problema del CUS sull’autonomia dei sistemi d’arma può contribuire a superare le difficoltà e le obiezioni che, nel dibattito politico e diplomatico sul controllo le armi autonome, sono emerse a proposito della soluzione ad autonomia zero (Amoroso 2020), senza però rinunciare a forme adeguate di CUS che siano eticamente fondate. Tali obiezioni possono infatti essere mitigate riconoscendo la possibilità di negoziare eccezioni alla regola prudenziale che impone livelli di controllo umano L1-L2. Ma solo per quei sistemi d’arma e per quei contesti d’uso che consentono al controllo umano di preservare, anche oltre il livello L2, le funzioni eticamente giustificate e richieste di sistema ausiliario di salvaguardia, catalizzatore di responsabilità e garante del rispetto della dignità degli esseri umani coinvolti.

Capitolo 8

Sostenibilità ambientale della società dell’informazione

Federica Lucivero

1. Introduzione

Questo capitolo affronta un problema finora poco trattato nel dibattito sull'etica del digitale e in particolare dell’intelligenza artificiale (IA): la sostenibilità ambientale dei sistemi hardware e software utilizzati in questo ambito. Le tecnologie digitali possono rappresentare un’opportunità di sviluppo sostenibile: si pensi all’uso di dati sul traffico urbano per ottimizzare le infrastrutture e ridurre l’inquinamento, all’uso di *smart grid* per ottimizzare i consumi energetici domestici, oppure ai modelli per stabilire in anticipo la quantità di prodotti di cui hanno bisogno ogni giorno negozi di alimentari e consumatori e così ridurre al minimo gli sprechi (Herweijer, Waughray, 2018). Allo stesso tempo, elaborare e immagazzinare grandi quantità di dati, e smaltire i dispositivi digitali sono tutte azioni che presentano impatti ambientali e sociali non trascurabili. In tal senso, dunque, si tratteranno alcune delle questioni etiche e sociali collegate alla sostenibilità ambientale della società dell’informazione.

Dopo un breve quadro sul rapporto tra società dell’informazione e sostenibilità, sarà introdotto il concetto di inquinamento digitale, attraverso esempi che richiamano l'attenzione sulla materialità dei dati, la configurazione fisica del servizio digitale e lo spazio occupato dai dati. Ognuno di questi esempi sarà poi utilizzato per discutere varie questioni di carattere etico e sociale: i limiti delle analisi costi/benefici in questo ambito, l’iniqua distribuzione di vantaggi e svantaggi nella popolazione, l’attribuzione di responsabilità per gli effetti nocivi dell’inquinamento digitale. Infine, si rifletterà sulle implicazioni di queste analisi per la governance tecnologica. La tesi centrale di questo capitolo è che la ricerca di soluzioni all’inquinamento digitale non è solo una questione tecnologica, ma anche politica e valoriale, che richiede di articolare la pluralità di valori coinvolti nelle decisioni, di valutare chi ne trae vantaggi e chi ne è leso, e di ragionare su chi ha la responsabilità di intervenire.

1. Sostenibilità ambientale, sviluppo sostenibile e big data

Nel linguaggio corrente, quando si parla di pratiche “sostenibili” ci si riferisce alla loro capacità di essere sostenute, ovvero mantenute nel tempo. Di solito il termine ha una connotazione positiva, ma nel suo significato letterale potrebbe riferirsi anche a pratiche, come la schiavitù, che sono sostenibili nel tempo ma moralmente problematiche. Nel contesto dei movimenti e dell’etica ambientalista, il concetto di sostenibilità si riferisce a pratiche e comportamenti che durano nel tempo perché rispettano limiti ecologici tali da portare benefici alla generazione presente e a quelle future (Attfield, 2013): nella pesca e nell’agricoltura sostenibile, per esempio, non si trattano le risorse ittiche o territoriali come semplici prodotti da sfruttare, ma si fa in modo che gli ecosistemi possano prosperare. Sebbene le pratiche sostenibili siano considerate come esempi virtuosi da un punto di vista ambientale, va sottolineato che il termine è entrato nell’uso comune per indicare la virtuosità di politiche e pratiche commerciali, finendo con l’essere usato spesso anche in modo improprio e privo di significato.

Il concetto di sostenibilità è spesso associato a quello di “sviluppo”. Nella definizione che si legge sul Rapporto Brundtland, pubblicato nel 1987 dalla Commissione Mondiale per l’Ambiente e lo Sviluppo (*World Commission on Environment and Development*, WCED), lo sviluppo sostenibile è «uno sviluppo che soddisfi i bisogni del presente senza compromettere la possibilità che le generazioni future soddisfino i propri» (WCED, 1987, trad. dell’autrice). Correlata a questa definizione di sviluppo sostenibile è quella di “bisogni”, rappresentati nel Rapporto su uno spazio tridimensionale che contempla bisogni di natura economica, sociale e ambientale. Le politiche volte a soddisfare questi bisogni devono quindi affrontare non solo problemi di povertà e malnutrizione – per esempio, attraverso pratiche di agricoltura e pesca sostenibili o attraverso la promozione di energie rinnovabili –, ma anche questioni di parità di genere, accesso a istruzione primaria e assistenza sanitaria di base, fondamentali per garantire il benessere degli individui nel presente e nel futuro. I *Millenium Sustainable Goals* (MSG, 2000-2015) e i *Sustainable Development Goals* (SDG, 2015-2030) sono obiettivi fissati dalle Nazioni Unite perché le politiche internazionali si muovano verso la realizzazione di un ideale sviluppo sostenibile.

È importante notare che la definizione di sviluppo sostenibile è fondata su un concetto di giustizia molto ampio, che promuove un’equa distribuzione di risorse e benessere non solo all’interno della stessa società ma anche tra diverse popolazioni, specie ed ecosistemi a livello globale. Una definizione che, inoltre, non si sofferma ad assegnare un certo status morale a individui, animali e sistemi ecologici nel presente, ma considera anche le responsabilità della generazione presente di garantire condizioni di vivibilità e prosperità ai propri discendenti. Questo ampliamento del raggio di azione di pratiche sostenibili (nello spazio, nel tempo e nelle dimensioni sopra menzionate) crea inevitabili conflitti morali: per esempio, garantire accesso energetico anche ai gruppi più vulnerabili in paesi più poveri potrebbe implicare l’aumento delle emissioni di gas serra a livello globale e di conseguenza danneggiare le generazioni future. D’altra parte, potrebbero esservi economisti o politici secondo i quali la responsabilità di garantire il benessere dei paesi che rappresentano (e dei gruppi vulnerabili al loro interno) è più importante che preoccuparsi del benessere di altre popolazioni o addirittura di generazioni future. Nonostante queste difficoltà pratiche, il concetto di sviluppo sostenibile mira a creare un quadro di riferimento per armonizzare interessi solo apparentemente in conflitto: in sostanza, suggerisce che il benessere delle generazioni future e presenti, dell’ecosistema e di popolazioni su scala globale sono interconnessi.

All’alba della ratificazione degli SDG, un rapporto delle Nazioni Unite ha assegnato alla rivoluzione digitale e ai big data un ruolo chiave per la realizzazione degli obiettivi per uno sviluppo sostenibile (UN IEAG, 2014). Il documento mostra come il crescente volume di dati, generati passivamente e raccolti automaticamente, permetta ai responsabili politici di monitorare e sostenere il raggiungimento di alcuni obiettivi: per esempio il target “energia pulita e accessibile” può essere raggiunto riducendo gli sprechi grazie ai dati sul consumo di elettricità, gas e acqua raccolti da sensori intelligenti; oppure “l'azione per il clima” è favorita dall’uso di dati satellitari volti a tracciare la deforestazione in atto. Nel valutare gli inevitabili punti di conflitto tra l’uso dei dati e il fine ultimo dello sviluppo sostenibile, il rapporto evidenzia il rischio che i dati disponibili non siano rappresentativi di tutti i gruppi della popolazione, soprattutto quelle più vulnerabili e con meno accesso a tecnologie digitali, le quali rischiano quindi di rimanere “invisibili”. Altre questioni sono dovute all’uso improprio dei dati e alle possibilità di discriminare o danneggiare proprio i gruppi più deboli. Pur mantenendo un atteggiamento equilibrato nel discutere le opportunità e i rischi dei big data, il documento delle Nazioni Unite trascura un rischio cruciale per lo sviluppo sostenibile: il fatto che la rivoluzione digitale e i big data abbiano un elevato impatto ambientale. Questo rapporto ambivalente tra sostenibilità e società dell’informazione è stato discusso a lungo da ricercatori in discipline informatiche e ingegneristiche. Come afferma il manifesto redatto nel 2013 dai partecipanti alla prima conferenza ICT for Sustainability (ICT4S), nonostante le tecnologie informatiche abbiano la capacità di rendere i modelli di produzione e consumo più sostenibili, «la storia della tecnologia ha mostrato che una maggiore efficienza energetica non contribuisce automaticamente allo sviluppo sostenibile» (Hilty *et al*., 2013, p. 289, trad. dell’autrice). Vediamo più da vicino che cosa significa ciò.

1. Inquinamento digitale: cos’è?

La terminologia utilizzata per riferirsi alla rivoluzione digitale spesso fa riferimento alla velocità, alla fluidità e alla presunta immaterialità della rete e dei sistemi informatici. La metafora del “*cloud*” – la nuvola eterea e senza limiti, che può essere estesa secondo le esigenze del cliente per pochi euro al mese e dove i dati possono essere immagazzinati e resi accessibili da qualunque postazione o dispositivo collegato in rete – è indicativa di questa fuorviante prospettiva sul digitale come un qualcosa di immateriale. In realtà, questi sistemi necessitano di vere e proprie infrastrutture industriali che occupano terreni, alterano paesaggi urbani e extraurbani, consumano energia, rilasciano gas serra nell’atmosfera e infine terminano generalmente il loro ciclo di vita in discariche.

Con l’espressione “inquinamento digitale” ci si riferisce al fatto che queste infrastrutture hanno un impatto ambientale rilevante dovuto alla produzione di gas serra con conseguenze negative sulla crisi climatica, esauriscono risorse naturali utilizzate per la loro produzione e rilasciano sostanze tossiche quando sono dismesse. Prendiamo l’esempio dei centri di elaborazione dati (CED). Essi sono «il sistema nervoso centrale del XXI secolo» (Whitehead *et al*., 2014, p. 152, trad. dell’autrice): ospitano server e apparecchiature di rete e di archiviazione vitali per servizi come il *cloud computing* e, più in generale, supportano i processi di elaborazione ad alta intensità energetica necessari a (1) memorizzare, gestire ed elaborare dati e (2) fornire applicazioni e servizi per l'elaborazione dei dati. Essi sono anche fondamentali per la raccolta dei big data che si utilizzano per addestrare i sistemi dell’IA (si veda il capitolo 2). I CED consumano una quantità crescente di energia per eseguire le loro operazioni e raffreddare i server (Avgerinou *et al.*, 2017). Hanno anche bisogno di generatori diesel per mantenere i server in funzione in caso di interruzione della fornitura di energia elettrica, producendo così ulteriori emissioni di gas serra con un conseguente impatto sulla crisi climatica. Secondo le stime, a causa della diffusione crescente del *cloud computing* e della crescita dell'uso dei servizi Internet, i CED hanno la più rapida crescita della *carbon footprint* (letteralmente “impronta di carbonio”[[36]](#footnote-36)) di tutto il settore delle tecnologie dell’informazione e della comunicazione (Avgerinou *et al*., 2017).

Allo scopo di fornire una valutazione complessiva dell’inquinamento digitale, si devono considerare ulteriori aspetti che caratterizzano il ciclo di vita delle infrastrutture della società dell’informazione. Lo smaltimento dell'hardware, per esempio, è particolarmente dannoso per l'ambiente. Questi processi di smaltimento spesso avvengono in paesi a basso e medio reddito e con controlli ambientali meno restrittivi, dove è possibile recuperare materiali pregiati dall’hardware (ad esempio, rame e oro), attraverso pratiche come l'incenerimento, che emette sostanze tossiche, aggrava l’inquinamento ambientale (Williams, 2011) e rappresenta anche un rischio per la salute dei lavoratori coinvolti.

Un’altra fonte di inquinamento digitale è direttamente collegata allo sviluppo di sistemi dell’IA. Un gruppo di ricercatori dell’University of Massachusetts ha calcolato il consumo di energia legato all’addestramento di alcune reti neurali profonde (APPRENDIMENTO AUTOMATICO pp….) utilizzate per l’elaborazione del linguaggio naturale (*Natural Language Processing*, NLP). Dallo studio è emerso che l’addestramento del sistema GPT-2 ha comportato l’emissione di gas serra pari a quasi cinque volte le emissioni di un’auto di media cilindrata in tutto il suo ciclo di vita (che comprende la produzione dell'auto stessa) ([Strubell](https://arxiv.org/pdf/1906.02243.pdf) *et al*., 2019). Questi dati diventano particolarmente rilevanti quando si considera che l’addestramento di modelli linguistici con grandi quantità di dati testuali è un ambito sempre più diffuso di ricerca, sperimentazione e applicazione per l’IA.

In sintesi, le tecnologie digitali in generale, e l’IA più nello specifico, comportano un elevato consumo di energia e hanno un notevole impatto ambientale. Bisogna sottolineare che tali stime di impatto ambientale presentano ancora elementi di incertezza e incompletezza, ma danno nondimeno un’idea della portata del problema in un contesto di crescita costante della richiesta e dell’offerta del digitale (Freitag *et al*., 2020). Questi elementi di incertezza non sono motivo sufficiente per trascurare il problema. Infatti, secondo gli accordi di Parigi, per contenere l’innalzamento della temperatura globale al di sotto di 1,5° C, ogni settore produttivo deve adattarsi e operare cambiamenti significativi. Anche prendendo a riferimento le previsioni più ottimistiche, è stato dimostrato che il mondo del digitale e dell’IA non può continuare a produrre la stessa quantità di emissioni, ma deve impegnarsi a una riduzione per evitare un aumento eccessivo del riscaldamento globale (Blair, 2020). Quindi, nonostante la mancanza di un consenso scientifico rispetto all’entità dell’inquinamento digitale, ci sono motivi legittimi per prevedere danni irreversibili in caso di mancata azione, soprattutto considerando il tasso di crescita del settore[[37]](#footnote-37). Come vedremo in seguito, questo è solo in parte un problema tecnico-scientifico (legato, per esempio al miglioramento dell’efficienza dei sistemi o a un calcolo più preciso del loro impatto ambientale), ma anche (e forse soprattutto) un problema etico-politico.

1. Inquinamento digitale e etica

Nell’affrontare il problema dell’inquinamento digitale non possiamo affidarci ad un arido calcolo di dati, ma dobbiamo 1) prendere in considerazione questioni relative alla sfera valoriale, 2) chiederci quali disuguaglianze e ingiustizie vengono generate in questo contesto e 3) riflettere su come attribuire responsabilità ai vari gruppi di interesse. Esaminiamo nell’ordine questi tre aspetti.

* 1. Fatti e valori

Costruire case e riscaldarle, mangiare, spostarsi: è ragionevole pensare che tutte le attività umane hanno un impatto sull’ambiente e sulla crisi climatica[[38]](#footnote-38), ma questo non ci fa rinunciare a queste attività. Solo alcuni tra gli attivisti più radicali della causa ambientalista raccomandano di andare a vivere in capanne senza riscaldamento, raccogliendo frutti spontaneamente offerti dalla natura e spostandosi solo a piedi[[39]](#footnote-39). Un approccio ambientalista più moderato (e più diffuso) consiste nel *valutare costi e benefici* di ognuna di queste attività, promuovendo la costruzione di case a basso consumo energico (mai nullo!), una dieta a base di prodotti con una impronta di carbonio ridotta (per esempio, consumando meno prodotti di origine animale e più prodotti locali) e spostamenti con mezzi pubblici o trasporti meno inquinanti, come auto ibride o treni, e così via, con l'obiettivo di fondo di contenere il riscaldamento climatico entro limiti accettabili. Un approccio simile per quanto riguarda l’inquinamento digitale comporterebbe la ricerca di una riduzione dell’impatto ambientale delle tecnologie digitali e dell’IA, ma non una rinuncia al loro uso. Questo approccio sarebbe in linea con l’etica delle conseguenze e in particolare con una valutazione di ampio respiro delle conseguenze attese (VEICOLI AUTONOMI pp…), in quanto teso a valutare le tecnologie digitali prendendo in considerazione non solo il loro impatto ambientale, ma anche conseguenze in altri ambiti (come quello dell’istruzione, della sanità o anche del risparmio energetico): la valutazione finale riguardo all’uso etico di queste tecnologie scaturirebbe dal rapporto tra i costi e benefici e mirerebbe a massimizzare i secondi. Adottare la prospettiva dell’analisi di costi e benefici, per quanto presenti più vantaggi di una completa rinuncia alla tecnologia e di un ritorno a un presunto stato di natura, non è una soluzione priva di problemi. Si tratta di un tipo di analisi che nasce in un contesto economico in cui il rapporto tra benefici e costi è calcolato attraverso l’assegnazione di un valore monetario. È proprio questa la principale critica mossa a tale prospettiva: decisioni politiche o valutazioni etiche derivano dalla necessità di monetizzare o comunque assegnare un valore quantificabile a conseguenze positive o negative (Turner, 1979). La quantificazione dei costi o dei benefici può essere difficile quando l’impatto di una certa politica o tecnologia non è determinabile con certezza. Inoltre, presuppone un consenso su come definire un beneficio (una conseguenza con una connotazione positiva) o un costo (una conseguenza con una connotazione negativa). Tale consenso è relativamente facile da raggiungere se si considerano valori monetari, facilmente quantificabili, ma diventa complesso quando entrano in gioco altri criteri e prospettive di valutazione. Vediamo più da vicino i limiti delle analisi costi benefici nel caso concreto dell’inquinamento digitale.

Per gestire e ridurre l’inquinamento digitale è importante conoscerne l’entità, sia in termini assoluti (quanto inquinano le tecnologie dell’informazione e l’IA) sia in termini relativi (quanto inquinano rispetto ai vantaggi ambientali e sociali che ne derivano). Queste due domande apparentemente semplici richiedono risposte complesse.

Prima di tutto, non è facile stabilire in termini assoluti l’entità dell’inquinamento digitale. Prendiamo di nuovo l’esempio dei CED. Le analisi dell’attuale consumo energetico e le previsioni di sviluppo futuro variano molto tra loro. Esse dipendono in primo luogo da dati che non sono sempre facili da acquisire perché spesso non sono disponibili pubblicamente e perché le metriche utilizzate per acquisirli variano da contesto a contesto (Whitehead *et al*., 2014). In secondo luogo, per quanto le tecnologie usate nei CED siano in costante evoluzione e tendano a diventare più efficienti, aumenta allo stesso tempo la domanda di accesso, condivisione, immagazzinamento ed elaborazione di dati. Il quadro diventa ancora più vario e complicato considerando altri fattori. Il consumo di energia e le emissioni di gas serra dei CED costituiscono un impatto ambientale *diretto* e relativamente facile da misurare, una volta che vengano decisi dei parametri standard e che i dati vengano resi disponibili. Ma una valutazione più realistica degli impatti ambientali del digitale deve prendere in considerazione anche gli effetti *indiretti* dell'ICT relativi ai comportamenti e alle pratiche degli utenti (Berkhout, Hertin, 2004). Per esempio, una pratica tecnologica come il telelavoro, che comporta un consumo minore di energia per quanto riguarda la guida di un’automobile o l’uso di altri mezzi di trasporto per recarsi al lavoro, può indurre alcuni comportamenti di consumo energetico che hanno un impatto negativo sull'ambiente: i telelavoratori potrebbero infatti utilizzare il tempo utilizzato in precedenza per gli spostamenti tra casa e lavoro per recarsi invece in auto altrove o spendere il denaro risparmiato per acquistare prodotti ad alto impatto ambientale (Williams, 2011). Quindi, anche se le prestazioni più elevate di una tecnologia la rendono più efficiente e consentono una diminuzione unitaria di consumo energetico e di prezzo, i consumatori potrebbero incrementarne l’uso o utilizzare i risparmi conseguiti per utilizzare altre tecnologie e sviluppare nuovi comportamenti ad alto costo ambientale. Effetti indiretti di questo tipo sono anche chiamati “effetti di rimbalzo” (*rebound effects*).

Gli effetti indiretti sono piuttosto difficili da valutare perché richiedono di prendere in considerazione comportamenti e norme sociali e culturali. Le pratiche quotidiane nell'uso di Internet, per esempio le ore della giornata passate a guardare video in streaming, sono una variabile importante per valutare la domanda di dati e quindi per valutare il corrispondente consumo di energia (Morley *et al*., 2018). Troppo spesso, tuttavia, gli effetti legati alle caratteristiche sociali e culturali dell'utente e gli effetti di rimbalzo sono esclusi dalle valutazioni degli impatti ambientali delle tecnologie digitali (Pohl *et al*., 2019). Questa analisi diventa ancora più complessa quando cerchiamo di prendere in considerazione tutto il ciclo di vita dei sistemi hardware e software della IA, dall’estrazione dei minerali che sono utilizzati per costruire i computer, allo sviluppo e all’uso dei sistemi software, fino allo smaltimento dell’hardware nelle discariche[[40]](#footnote-40). I limiti delle metodologie esistenti per valutare tutte le variabili rilevanti e gestire i corrispondenti fattori di incertezza richiedono un monitoraggio costante degli sviluppi della società dell’informazione e complicano le analisi costi/benefici.

È importante anche considerare che un’analisi costi/benefici non solo richiede di avere a disposizione dei valori numerici da poter confrontare, ma anche di avere un sistema di riferimento per valutare il significato sociale di questi valori. Come abbiamo visto nel paragrafo 1, il concetto di sviluppo sostenibile si riferisce non solo alla dimensione ambientale, ma anche a quella economica e sociale. Quindi, per stabilire la sostenibilità di servizi informatici non dobbiamo solo riuscire a misurare il valore numerico del loro impatto ambientale[[41]](#footnote-41), ma dobbiamo anche interrogarci sul loro “valore sociale”, per esempio il modo in cui i servizi informatici promuovono il soddisfacimento di bisogni come l’accesso all’educazione e a cure sanitarie di base, oppure favoriscono una società con meno disuguaglianze e discriminazioni. Sulla base di un’analisi costi/benefici, se i servizi informatici in questione hanno un valore sociale più alto, potremmo essere più disposti a giustificare un impatto ambientale maggiore. Ma come si misura il valore sociale di un servizio digitale? Come qualificare il valore di servizi per la comunicazione o per l’informazione rispetto a servizi per il commercio online o per l’immagazzinamento di fascicoli sanitari elettronici? E come si valuta il valore sociale di tipi diversi di ricerche o applicazioni in IA e in altri settori dell’informatica? Il valore sociale non solo non è facilmente misurabile, ma non è neanche determinabile in modo uniforme. Cambia, per esempio, a seconda della popolazione in riferimento alla quale si valutano i benefici. I servizi di raccolta ed elaborazione di dati sanitari potrebbero essere più importanti in un paese altamente informatizzato di quanto lo siano in un paese meno informatizzato, dove la documentazione cartacea è più utilizzata. Inoltre, come vedremo nel prossimo paragrafo, le popolazioni più vulnerabili alla crisi climatica sono anche quelle attualmente meno avvantaggiate dalla rivoluzione digitale. La valutazione si complica ulteriormente se si prendono in considerazione i bisogni e gli interessi delle generazioni future o se si assume una prospettiva meno antropocentrica, che tiene in considerazione anche le condizioni dei diversi ecosistemi. Una valutazione ambientale basata su un approccio consequenzialista, che valuta le conseguenze di una pratica tecnologica sulla base del benessere collettivo che ne deriva, non è facile da attuare in questo campo, poiché i valori dell’impatto sia ambientale che sociale sono difficili da misurare e variano a seconda dei contesti.

Questi elementi di incertezza e complessità suggeriscono che le attuali misure di valutazione dell'impatto ambientale potrebbero non essere sufficienti per fornire risposte definitive o una chiara ponderazione dei benefici e dei costi. Siffatte difficoltà concettuali ed empiriche non forniscono però una giustificazione per trascurare il problema, poiché la crisi climatica richiede un intervento urgente. Esse mostrano piuttosto che bisogna prendere decisioni al riguardo in un contesto di incertezza e incompletezza informativa. Inoltre, una riflessione etica e filosofica emerge come necessaria per considerare non solo aspetti quantitativi, ma anche dimensioni qualitative, sociali e morali del problema. Capire come le iniziative sui dati mobilitano diversi valori e riflettere su come bilanciarli tra loro è quindi un compito fondamentale per la riflessione etica in questo contesto. Tale riflessione non è un esercizio da svolgere *una tantum*, in via preliminare, quantificando e ponderando l’impatto prima che specifici progetti siano approvati o specifiche tecnologie si diffondano, ma richiede piuttosto una riflessione continua, da affiancare a progetti di ricerca e di innovazione tecnologica, che si concentri anche sugli aspetti più qualitativi, comportamentali e sociali, del loro impatto ambientale.

* 1. Ingiustizia sociale e ambientale

Un secondo aspetto eticamente rilevante è legato a questioni di giustizia rispetto all’equa distribuzione di danni e benefici all’interno della società. Nell’affrontare tali questioni è bene ricordare che IA e tecnologie digitali richiedono risorse naturali che sono limitate, come l’uso di terreni per la costruzione di CED o altre infrastrutture, risorse come l'acqua o, in molti casi ancora, fonti energetiche non rinnovabili per la produzione di elettricità. Queste caratteristiche delle infrastrutture della società digitale hanno significative implicazioni sociali ed economiche. Per esempio, la collocazione di CED nelle periferie di centri abitati dei paesi scandinavi, con climi più freddi che richiedono meno energia per raffreddare i server, è un motivo di forti cambiamenti politici e sociali in quelle zone: grandi investimenti sono stati fatti anche nella speranza che queste infrastrutture portassero localmente lavoro, ma le aspirazioni di cambiamento in tal senso spesso si sono scontrate con una realtà diversa: i posti di lavoro offerti alle comunità locali sono limitati, a causa dell’alta e crescente automazione dei CED. Allo stesso tempo, queste infrastrutture modificano le economie locali e creano nuovi equilibri nazionali in cui zone periferiche acquistano un ruolo più centrale, ma le popolazioni locali spesso non traggono benefici immediati dalle decisioni della classe dirigente (Vonderau, 2018). Tali considerazioni ci spingono a riflettere su come queste infrastrutture non solo alterano lo spazio fisico, ma hanno conseguenze sull'organizzazione socio-economica delle zone in cui vengono costruite e rendono necessario valutare la ripartizione di danni e benefici tra i vari gruppi interessati.

Questioni di giustizia emergono anche quando si considera che la localizzazione fisica delle infrastrutture di rete e dati determina le condizioni di accesso ai servizi. Nel 2013, il giornalista Andrew Blum, ha pubblicato il libro *Tubes: Behind the Scenes at the Internet*, in cui descrive una ricerca durata due anni, volta a capire meglio la materialità di Internet visitando i luoghi reali dove questo servizio è implementato, fornito e distribuito (Blum, 2013). Nel descrivere i dettagli fisici, gli edifici, le persone e le infrastrutture che rendono la rete qualcosa di tangibile, capace di modificare le geografie e il paesaggio, Blum non solo mostra che Internet è un luogo, ma anche che i dati si trovano in alcuni luoghi più che in altri. La vicinanza degli utenti ai server permette alle informazioni di raggiungerli più velocemente. Questa caratteristica è abbastanza rilevante per varie tipologie di aziende che beneficiano di un rapido accesso ai servizi. Per questa ragione, le società che offrono sistemi di immagazzinamento di dati tendono ad acquistare spazi in aree urbane, come il centro di Manhattan, al fine di consentire alle aziende di intermediazione finanziaria che operano in borsa di lavorare alle velocità necessarie per essere più competitive (Carp, 2013). Nel contesto di un’economia dei dati (*data economy*) le condizioni di accesso ai dati implicano accesso alla ricchezza e allo sviluppo economico. Se la vicinanza alle infrastrutture ha implicazioni sulla competizione economica per l'accesso rapido o l'elaborazione dei dati, è giusto lasciare che siano le società di borsa a usufruire di connessioni migliori? Ci sono modi più trasparenti e democratici per decidere come ridistribuire lo spazio urbano, tenendo anche conto delle esigenze delle comunità locali? Lasciare queste decisioni esclusivamente nelle mani del settore economico potrebbe essere problematico dalla prospettiva del bene comune e della giustizia distributiva.

Le decisioni relative alla localizzazione delle infrastrutture digitali non sono mai neutrali, poiché hanno forti impatti sociali e dovrebbero perciò essere discusse tenendone debitamente conto. Questo elemento emerge anche nel caso dello smaltimento dei dispositivi digitali di cui si è parlato sopra. Gli impianti di smaltimento per i dispositivi elettronici dismessi sono spesso localizzati in paesi a basso reddito nel sud-est asiatico dove sono presenti centri per il riciclo con strutture spesso fatiscenti e non a norma. Le persone che lavorano in questi centri sono spesso sottopagate ed esposte a gravi rischi per la salute, dovuti all’esposizione a prodotti chimici utilizzati nella lavorazione o presenti insieme a metalli tossici negli stessi dispositivi elettronici. Le reali condizioni di smaltimento sollevano ulteriori questioni relative all'equa distribuzione dei costi e dei benefici della rivoluzione digitale: le popolazioni più vulnerabili subiscono in questo caso le conseguenze ambientali della crescente domanda del digitale nei paesi ricchi (Maxwell, Miller, 2012; Gabrys, 2013). Simili considerazioni emergono anche riguardo all’estrazione di minerali per la produzione di dispositivi digitali in paesi ricchi di risorse naturali, dove la manodopera locale viene spesso assunta in condizioni di semi-sfruttamento.

La rivoluzione digitale distribuisce benefici e danni ambientali tra gruppi e comunità, ma quali criteri dovranno essere considerati oltre a quelli economici e chi dovrebbe partecipare a queste decisioni? La “giustizia ambientale” non è un concetto nuovo, ma trova le sue radici nei movimenti politici statunitensi già dalla fine degli anni ‘80. Come è ben spiegato sul sito della *US Environmental Protection Agency*[[42]](#footnote-42), la ricerca di giustizia ambientale è motivata dal fatto che non tutti «godono dello stesso grado di protezione dai rischi ambientali e sanitari» e che i criteri con cui benefici e svantaggi ambientali sono distribuiti devono essere ispirati a principi di giustizia e devono affrontare le ragioni strutturali di tali squilibri (Schlosberg, 2009). Studiosi del concetto di *data justice* hanno ampiamente dimostrato che la rivoluzione digitale crea nuove disuguaglianze e ne consolida altre già presenti nella struttura sociale (Dencik *et al*., 2019; Taylor, 2017). È importante sottolineare che alcune delle disuguaglianze prese in esame in questi studi sono collegate proprio alle questioni ambientali indotte dalla rivoluzione digitale in corso, come per esempio la localizzazione dei CED in determinate aree e il loro impatto sulle popolazioni locali, oppure l’uso massiccio di risorse naturali e l’emissione di sostanze tossiche a danno di popolazioni più vulnerabili, che fanno tra l’altro un uso minore dei servizi digitali. Questioni di disuguaglianza, discriminazione ed esclusione sociale sono trasversali alla riflessione etica sulla rivoluzione digitale e all’etica ambientalista; per affrontare il fenomeno dell’inquinamento, che si trova al bivio tra queste due discipline, bisogna interrogarsi su come ingiustizie vecchie e nuove e i rapporti di potere tra gruppi sociali ostacolino il raggiungimento degli ideali di sviluppo sostenibile.

* 1. Responsabilità individuale o istituzionale?

Al momento, la questione ambientale è un argomento ancora piuttosto periferico nella riflessione etica sull’IA (Jobin *et al*, 2019; Lucivero, 2019). Allo stesso tempo, come abbiamo illustrato con vari esempi, ci sono numerose questioni politiche, sociali e etiche ad essa connesse che devono essere affrontate. Ma da chi? Nell’articolo sui consumi energetici dell’addestramento di modelli per NLP descritto nel secondo paragrafo, i ricercatori della University of Massachusetts ad Amherst si interrogano sulle responsabilità del settore industriale, di quello accademico e pubblico rispetto all’inquinamento ambientale della ricerca sulla IA, suggerendo la necessità di distribuire più equamente le limitate risorse IA (che al momento sono in larga parte nelle mani dei privati) e di sviluppare e promuovere l’uso di algoritmi computazionalmente più efficienti e di dispositivi a minore consumo energetico. Nella stessa direzione sono orientate le raccomandazioni del Gruppo indipendente di esperti ad alto livello sull’IA nominato dalla Commissione Europea per elaborare delle linee guida per una IA affidabile (HLEG AI 2019; CODICI ETICI pp….): tra i criteri identificati c’è anche la sostenibilità ambientale dell’IA. Il monitoraggio dell’impatto ambientale derivante da sviluppo, distribuzione e utilizzo dei sistemi di IA deve essere messo in atto insieme a misure appropriate per ridurre il loro impatto ambientale su tutto il ciclo di vita. Queste richieste sono motivate da questioni etiche e di giustizia ambientale declinate nel contesto specifico dell’IA. Allo stesso tempo, tuttavia, queste richieste rimangono ancora ad un livello tale di astrazione da renderle difficili da implementare. Quali misure dell’impatto ambientale sull’intero ciclo di vita di un sistema di IA sono ritenute affidabili? Chi si assume i costi di tale monitoraggio, della realizzazione di sistemi più efficienti e con un’impronta ambientale complessivamente migliore? Come assicurare che questi costi non penalizzino ulteriormente i piccoli centri di ricerca, le piccole e medie imprese e gli utenti a favore delle già potenti multinazionali nel settore?

Il fatto che queste richieste eticamente motivate siano nella sostanza ancora parte di un sistema di autoregolamentazione, e quindi del tutto volontarie, lascia aperte varie questioni riguardo alla distribuzione delle responsabilità in questo contesto e al rispetto di determinati standard di sostenibilità ambientale. Se le responsabilità morali e legali rispetto all’impatto ambientale del digitale e dell’IA non sono ben identificate, un’azione di mitigazione di tale impatto perde di forza e rimane legata alla volontà del singolo attore. Per esempio, alcune grandi società tecnologiche come Microsoft, Google, Amazon e Apple si sono impegnate spesso pubblicamente a ridurre le emissioni di gas serra. Sebbene questo impegno sia apprezzabile in linea di principio, non è chiaro in pratica che cosa promettano queste aziende. Alcune si propongono di raggiungere un obiettivo *carbon neutral*, che consiste nel compensare le proprie emissioni di gas serra nell’atmosfera con azioni di eliminazione di questi gas (come per esempio attività di riforestazione). Al momento non esiste uno standard comune che determini: 1) quali elementi della filiera vengano considerati nel calcolo delle emissioni prodotte da un’azienda (per esempio se si considerano solo emissioni prodotte dai combustibili fossili *in situ* oppure se le misure comprendono le emissioni prodotte dall’uso di elettricità e gas comprate da terzi, o se si calcolano anche le emissioni indirette prodotte in diversi stadi della filiera, come estrazione, manifattura, trasporto, uso, smaltimento); 2) come queste emissioni debbano essere compensate (per esempio attraverso l’uso di energie rinnovabili o l’acquisto di crediti di carbonio) e 3) quali strategie di compensazione siano efficaci (per esempio acquistare e tutelare una foresta che potrebbe finire distrutta o finanziare progetti di riforestazione). Queste ambiguità portano alcuni attivisti e ricercatori a sostenere che un obiettivo carbon neutral non è abbastanza ambizioso, perché si limita a *compensare*, ma non a *ridurre* la totalità delle emissioni (Freitag *et al*., 2020). Secondo tale prospettiva, mirare a *zero net energy* è un obiettivo più desiderabile, dal momento che implica non solo una compensazione tra gas emessi e rimossi, ma anche una riduzione delle emissioni in tutta la filiera. In teoria, un’azienda che mira alla “neutralità di carbonio” potrebbe aumentare le emissioni totali, senza perciò contribuire allo sforzo di limitare l’incremento del riscaldamento globale entro 1,5° C, obiettivo che richiede la riduzione del 50% delle emissioni annuali fino al 2030 e il raggiungimento di “zero emissioni nette” (*net zero)* entro il 2050 (https://www.ipcc.ch/sr15/). Date queste premesse, le aziende che si impegnano a raggiungere obiettivi *carbon negative* sono ancora più encomiabili, perché intendono eliminare più gas serra di quanti ne emettono attraverso tutta la filiera. Eppure questi obiettivi sono spesso presentati senza informazioni precise sulla loro attuazione e hanno quindi portato alcuni commentatori a parlare di iniziative di *greenwashing* (CODICI ETICI pp….),in cui le aziende adottano cioè un ecologismo di facciata, strumentale per influenzare l’opinione pubblica e restituire un’immagine virtuosa di impegno ambientale. Per esempio, alcune aziende affermano di voler utilizzare solo energia che proviene da fonti rinnovabili (solare, eolica idrica), ma allo stesso tempo non specificano se e come il passaggio all’*100% renewable* porterà a una diminuzione effettiva delle loro emissioni e quali siano queste fonti di energia. Questi esempi mostrano i limiti della scelta di lasciare questioni di riduzione di impatto ambientale alla autoregolamentazione delle singole organizzazioni.

D’altra parte bisogna chiedersi se sia legittimo assegnare la responsabilità per la riduzione e compensazione dell’inquinamento digitale solo ai singoli individui (siano essi ricercatori, dipendenti o cittadini in generale), nonostante ci siano state delle iniziative come *Clicking Clean* di Greenpeace[[43]](#footnote-43), che hanno contribuito a sviluppare la coscienza critica dei consumatori relativamente all’impatto ambientale delle attività delle aziende fornitrici di servizi digitali e quindi a spingere le stesse aziende ad attuare politiche ambientaliste più efficaci. Nel complesso, i consumatori hanno ancora poca consapevolezza dell’impatto ambientale delle loro abitudini digitali (come l’utilizzo di sistemi di *cloud* per l’immagazzinamento di dati o lo streaming di video). In modo analogo, non possiamo assegnare la responsabilità di rendere i sistemi digitali più sostenibili ai singoli ricercatori. Infatti i singoli individui sono parte di una rete o ecosistema di cui fanno parte istituti di ricerca pubblici e privati, case editrici, associazioni di categoria, consigli di ricerca e finanziatori della ricerca. Se all’interno di un tale ecosistema non c’è una condivisione di che cosa significa agire responsabilmente, diventa difficile per i singoli attori riconoscere e attuare linee di condotta responsabili in modo coerente ed efficace (Samuel *et al*., 2019). Per essere riconosciuta e attuata dai singoli individui, la sostenibilità ambientale dei sistemi digitali deve essere riconosciuta a livello sistemico, per esempio da enti finanziatori che devono esplicitare nei loro bandi come affrontare le questioni ambientali in un progetto di ricerca. L’agire responsabile in questo ambito comporta cambiamenti profondi nei modi di condurre la ricerca, a cominciare dai premi e dagli incentivi distribuiti dagli enti finanziatori e dalle associazioni professionali (Samuel, Lucivero, 2020).

1. Conclusioni

I servizi e i dispositivi digitali, la raccolta e l’analisi di grandi quantità di dati, l’addestramento di sistemi di IA - in breve la società dell’informazione - contribuiscono in molti modi allo sviluppo sociale ed economico e alla ricerca di soluzioni ai problemi ambientali. In questo senso, si tratta di tecnologie che possono contribuire a realizzare uno sviluppo sostenibile e a soddisfare i bisogni della popolazione globale, delle specie animali e degli ecosistemi. Allo stesso tempo, la sostenibilità del digitale presenta ombre e ambiguità che ci costringono a interrogarci sul rapporto tra sviluppo sostenibile e inquinamento digitale.

Con inquinamento digitale si intendono diversi tipi di effetti negativi che la rivoluzione digitale ha sull’ambiente. Abbiamo illustrato le conseguenze del consumo di energia, dell’uso di risorse naturali e delle emissioni di gas serra da parte dei CED, dell’emissione di sostanze tossiche durante lo smaltimento di dispositivi elettronici e delle emissioni prodotte dall’addestramento di reti neurali profonde dell’IA. Nonostante i dati sul ciclo di vita di questi processi e prodotti siano difficili da raccogliere e analizzare e le previsioni sulle emissioni future siano discordanti, la crescita costante della domanda in questo ambito e gli accordi internazionali per prevenire danni irreversibili da riscaldamento globale rendono evidente che la questione della sostenibilità ambientale di queste tecnologie deve essere affrontata con decisione.

Come è stato più volte ribadito in questo capitolo, cercare soluzioni all’inquinamento digitale non è una questione esclusivamente tecnologica, ma anche politica e valoriale. In questo senso, le analisi costi/benefici che mirano a quantificare vantaggi e svantaggi presentano un’utilità limitata in un contesto in cui i dati (su costi e benefici) sono incerti perché fortemente dipendenti da aspetti comportamentali e sociali. Anziché tentare solo di quantificare queste implicazioni, è importante articolare la pluralità di valori coinvolti e di prospettive rilevanti in simili valutazioni. L’importanza della dimensione valoriale emerge anche quando si considera che la configurazione fisica dei servizi digitali e la loro localizzazione spaziale presentano distribuzioni non uniformi dei danni ambientali tra gruppi e popolazioni. Il rischio di creare nuove disuguaglianze e consolidare quelle esistenti deve essere affrontato e discusso secondo modalità inclusive, che comprendano i diversi gruppi di interesse e non solo i gruppi di potere. Infine, la questione dell’inquinamento digitale non può prescindere dalla discussione dell’attribuzione e della distribuzione di responsabilità tra le parti in causa. Al momento, le soluzioni a tale problema sono spesso lasciate al mercato e all’adozione volontaria da parte del settore privato di obiettivi di riduzione delle proprie emissioni. Come abbiamo visto, questa impostazione soffre di una mancanza di standard chiari e e di garanzie sull’efficacia delle soluzioni adottate. Si devono prendere in considerazione anche le responsabilità dei singoli, anche se, come abbiamo visto, esse sono spesso limitate dal contesto sociale e organizzativo. In definitiva, ciò significa che le organizzazioni pubbliche, nazionali e sovranazionali hanno un ruolo decisivo da svolgere nel creare standard, meccanismi di controllo e di verifica.

Al momento esistono diversi codici di autoregolamentazione nel settore privato e alcune normative che specificano degli standard (per esempio, per quanto riguarda il miglioramento dell’efficienza energetica dei CED o l’inserimento dei rifiuti elettronici in speciali piani di riciclo e smaltimento)[[44]](#footnote-44), ma l’approccio delle istituzioni pubbliche è frammentario e carente. In particolare, mancano politiche e normative che affrontino la questione della sostenibilità digitale in modo efficace attraverso incentivi o sanzioni. Alla base della risposta istituzionale c’è una fiducia di base nel fatto che l’efficienza energetica di questi sistemi aumenti con l’avanzamento tecnologico e che azioni volte a finanziare il settore e sostenere l’incremento della domanda siano più efficaci di azioni volte a ridurre le emissioni (Freitag, 2020). In conclusione è necessario e urgente sviluppare una visione di insieme che guidi un progetto di governance tale da affrontare efficacemente le questioni relative alla sostenibilità della società dell’informazione e definire i criteri generali da porre alla base dei singoli strumenti normativi. La riflessione etica gioca un ruolo significativo in questo contesto, perché offre gli strumenti per chiarire i concetti e i valori in gioco, diagnosticare i problemi moralmente rilevanti da affrontare, e identificare e mitigare i conflitti tra valori e bisogni diversi, fino a sviluppare una base morale condivisa per giustificare un progetto di governance.

*Parte terza*

Persone e interazioni

Capitolo 9

Robotica sociale: persuasione, inganno e etica del design

Fabio Fossa

1. Introduzione

Più le tecnologie sono pensate per interagire con noi, più diventa importante interrogarsi sui modi in cui facciamo esperienza e formiamo relazioni con gli oggetti tecnologici. Sviluppare tecnologie che rendano l’interazione facile, coinvolgente ed efficace partendo dallo studio delle relazioni tra esseri umani e sistemi tecnologici è lo scopo principale della robotica sociale. Da ciò dipende l’adozione di sistemi intelligenti in molti contesti sociali che ne potrebbero beneficiare sostanzialmente, dall’abitazione (per coadiuvare la cura degli anziani o fornire assistenza nelle faccende domestiche) all’ospedale (in corsia, per il monitoraggio dei pazienti), dal posto di lavoro (per l’assistenza ai lavoratori o l’accoglienza dei clienti) alla camera da letto (si veda anche il capitolo 10), fino ai supermercati e alle scuole (si veda il capitolo 11). In tutti questi casi la dimensione sociale è inestricabilmente intrecciata alla dimensione più propriamente tecnologica, il che genera diversi problemi di carattere morale – primi fra tutti le questioni della manipolazione e dell’inganno: fino a che punto è lecito ingannare l’utente, generando ad arte la sensazione che stia interagendo con un essere senziente? Fino a che punto è lecito influenzarne implicitamente il comportamento, fosse anche per spingerlo ad agire in modi ritenuti socialmente desiderabili? Non si tratterebbe di una lesione della dignità dell’utente, trattato con condiscendenza e abilmente manipolato? Ma è davvero importante tutto ciò, se l’interazione va a buon fine e l’utente, per quanto forse ingannato, ne ricavasse dell’utile o fosse convinto di beneficiarne?

Il capitolo si apre con un’introduzione sulla robotica sociale, i suoi scopi e le sue metodologie, per poi soffermarsi su captologia e tecnologie persuasive come metodi per migliorare la qualità delle interazioni tra esseri umani e sistemi tecnologici. Oltre a chiarirne l’approccio, se ne esplora il significato morale, percorrendo la linea sottile che separa spinta gentile [*nudge*] e manipolazione. Si considera poi la questione dell’inganno, analizzandola dalle prospettive dell’etica dei doveri e dell’etica delle conseguenze. Infine, si presenta il *Value Sensitive Design* come strategia per integrare i valori morali in gioco nelle tecnologie stesse in sede di progettazione, così da limitare potenziali conseguenze eticamente problematiche e promuovere l’adozione di valori tramite l’interazione con le tecnologie.

1. Robotica sociale

In molti casi, i sistemi tecnologici raggiungono livelli più elevati di efficienza se si riduce il coinvolgimento di esseri umani nel loro funzionamento. Si pensi, ad esempio, alla piena automazione di alcuni sistemi di produzione, dove l’intervento umano è un fattore da minimizzare data la sua imprecisione, lentezza, imprevedibilità e fragilità. In altri casi, invece, i sistemi tecnologici sono progettati proprio per interagire con gli esseri umani. Ciò vale per tutte le situazioni in cui il sistema tecnologico è pensato per funzionare in ambienti popolati non esclusivamente da tecnologie, ma anche da esseri umani. Se si pensa a quanti e quali contesti potrebbero beneficiare di un supporto tecnologico capace di interagire efficacemente con utenti umani – dall’ospedale all’ospizio, dall’azienda all’officina, dall’abitazione privata all’ufficio e alla strada – se ne colgono facilmente le potenzialità.

Date queste premesse, è altrettanto facile capire come il compito dei progettisti si complichi. Non si tratta più solamente di organizzare diversi prodotti tecnologici in modo che la loro azione combinata conduca allo scopo prefissato, come nel caso della piena automazione di un processo industriale. Alla logica delle interazioni tra macchine si intreccia la logica, spesso oscura e contorta, delle interazioni sociali. Per funzionare efficacemente in un contesto sociale, il sistema tecnologico deve saper stare al gioco delle relazioni interpersonali. Se si vuole che la tecnologia si integri senza intoppi nel tessuto sociale nel quale deve svolgere le sue funzioni, è necessario che sia percepita comeun elemento che appartiene, almeno a prima vista, a questo tessuto. È necessario, cioè, che sia in grado di esibire un comportamento *sociale* *quanto basta* perché l’interazione sia tanto efficace quanto credibile. Ottenere questo risultato è lo scopo della robotica sociale(Breazeal, 2003)*.*

L’etichetta “robotica sociale” può sembrare limitante se la si applica solo a tecnologie robotiche. Ne rimarrebbero esclusi gli agenti conversazionali come gli assistenti vocali, una delle più diffuse applicazioni dell’IA in grado di interagire con noi. È meglio quindi adottare un’interpretazione più flessibile. Fa al caso nostro quella proposta da Wallach and Allen (2009, p. 3), i quali parlano di «(ro)bot» per includere anche programmi dotati o meno di un corpo digitale (avatar) o materiale. Se decliniamo la proposta nel contesto di cui ci stiamo occupando, le condizioni minime perché una tecnologia rientri nel campo di interesse della robotica sociale sono dunque due: 1) che sia dotata di autonomia operativa; 2) che sia pensata soprattutto per interagire con esseri umani.

Ciò detto, ogni robot sociale deve essere attentamente valutato a partire dalle sue caratteristiche software e hardware, in quanto diverse scelte progettuali pongono problemi di tipo o grado differente. Ad esempio, un agente conversazionale per l’accoglienza dei clienti in un centro commerciale pone sfide diverse a seconda che sia dotato di un corpo umanoide libero di muoversi in un dato spazio o, più semplicemente, di una colonnina provvista di schermo tattile e altoparlante. Il problema etico di garantire l’incolumità degli utenti si pone solo nel primo caso, in cui i movimenti del corpo robotico possono servire a migliorare la qualità dell’interazione ma pongono anche il rischio di collisioni pericolose. Altri problemi etici, per esempio la manipolazione del comportamento degli utenti attraverso specifiche scelte di progettazione, accomunano invece le due tecnologie ma, come vedremo più avanti, possono assumere forme e gradi assai diversi.

Per capire meglio in cosa consiste il compito della robotica sociale, prendiamo brevemente in esame un carattere tipico dell’interazione tra umani: la dimensione emotiva (Adams *et al.*, 2000). Tramite l’espressione delle emozioni comunichiamo una grande varietà di informazioni sui nostri stati interiori: preferenze, stati d’animo, disposizioni, intenzioni e così via. Saper leggere le emozioni e modulare il comportamento in modo consono è un’abilità sociale primaria, di cui facciamo largo uso nel quotidiano. Ne deriva che, tra esseri umani, lo scambio di informazioni più rapido e preciso potrebbe non essere sempre il più funzionale. Alcune volte è opportuno tralasciare delle informazioni se il nostro interlocutore ci fa intendere di esserne annoiato; oppure, se mostra segni di stress emotivo, è meglio addolcire messaggi critici, spendendo più tempo ed energie per coltivare il rapporto umano che ci lega. Cosa succede quando un sistema tecnologico prende il posto di un interlocutore umano? Perché l’interazione sia efficace, lo scambio di informazioni implicito veicolato dall’espressione delle emozioni deve continuare a giocare il suo ruolo, in entrambe le direzioni. Di conseguenza, il sistema dovrebbe essere in grado non solo di leggerele emozioni dell’utente umano, ma anche di esprimere emozioni adeguate alla natura dell’interazione. Se ciò non avviene, l’interazione non sarà abbastanza efficace e l’estraneità del prodotto tecnologico – un alieno tra esseri umani – ne metterà a rischio l’adozione. La componente sociale e quella tecnica diventano inscindibili.

La rilevanza assunta dalla dimensione sociale è sicuramente fonte di ulteriori difficoltà per i progettisti, che si trovano a dover trattare una materia quanto mai difficile da tradurre in buoni modelli di interazione umani-robot. Allo stesso tempo, si apre l’opportunità di esplorare soluzioni forse meno convenzionali, ma non per questo meno funzionali. Dal momento che l’accettazione sociale del sistema tecnologico è basata innanzitutto sulla percezione che se ne ha, i progettisti hanno ora a loro disposizione l’intera dimensione dell’apparenza. Far sì che il prodotto appaiain grado di comportarsi in modo socialmente impeccabile diventa tanto importante, ma spesso meno dispendioso, che renderlo effettivamente in grado di fare ciò. Ad esempio: invece che predisporre un intero modello computazionale della vita emotiva umana, far sì che qualche variazione nel tono di voce o nell’espressione facciale del robot accompagni determinati input o output può essere sufficiente per suggerire agli utenti che anche il sistema abbia emozioni e, così, attivare reazioni che migliorano la percezione della qualità dell’interazione.

Indurre gli utenti a proiettare sul sistema qualità significative da un punto di vista sociale – affidabilità, competenza, autorità, affabilità, consapevolezza, solidarietà, impegno, attaccamento – diventa una strategia praticabile e parsimoniosa in termini di sforzo progettuale. Influenzare il modello mentale che gli utenti formano della tecnologia tramite accorgimenti *ad hoc* può risultare ugualmente funzionale e più fattibile che modificare le funzioni svolte dal sistema. Sorgono però delle domande etiche: quanto è lecito far credere all’utente che il sistema possa contare su capacità di cui non dispone? Quali sono i rischi e, se ci sono, vale la pena di correrli? La liceità o meno di ingannare l’utente tramite precise scelte di progettazione allo scopo di migliorare la qualità delle interazioni conduce al filo rosso che lega a sé molti dei problemi morali connessi alla robotica sociale. Prima di occuparcene direttamente, chiariamo in che senso alcune tecnologie possono *persuadere* l’utente a comportarsi in determinati modi.

1. Captologia e tecnologie persuasive

Il potere persuasivo delle tecnologie digitali capaci di interagire in modo socialmente credibile con gli utenti umani è emerso in tutta la sua rilevanza dapprima nel campo di studi sulle interazioni tra esseri umani e computer (Nass, Steuer, Tauber, 1994), dove è poi divenuto unico oggetto di indagine della *captologia*, la studio delle tecnologie persuasive (Fogg, Cuellar, Danielson, 2008; Siegel, Breazeal, Norton, 2009). La disciplina prende le mosse dal riconoscimento che l’interazione con programmi informatici influisce significativamente sul comportamento degli utenti. Lo scopo è di sfruttare consapevolmente tale influenza, indirizzandola verso obiettivi rilevanti da un punto di vista progettuale. Se la progettazione di programmi informatici negli ultimi decenni può essere descritta come dominata da tre motivi principali – funzionalità, intrattenimento e facilità d’uso – ora ne emerge un quarto: la persuasività. La captologia si occupa proprio dello sviluppo di strategie di progettazione che servano a gestire efficacemente il potenziale persuasivo delle tecnologie, cioè a influenzare il comportamento degli utenti indirizzandolo verso scopi e obiettivi di interesse – primo fra tutti, migliorare la qualità delle interazioni tra esseri umani e sistemi digitali.

La captologia mette in luce quale e quanto sia il potere persuasivo che può essere esercitato sugli utenti attraverso precise scelte di progettazione delle tecnologie. Le decisioni prese dai *designer* di interazioni plasmano l’esperienza d’uso, riuscendo a mobilitare ad arte emozioni, idee, reazioni e comportamenti degli utenti. Ci sono molti modi in cui scelte di progettazione, che si traducono poi in caratteristiche del prodotto, possono persuadere gli utenti ad adottare certi comportamenti o sviluppare determinate abitudini. In generale, le tecnologie persuasive possono essere divise in due classi a seconda dell’importanza che la persuasione gioca nella loro progettazione. Si parla di *macrosuasione* quando persuadere gli utenti a comportarsi in un certo modo o a convincersi di una certa idea è lo scopo principale che motiva la progettazione del prodotto tecnologico. Ne è un esempio un programma di realtà virtuale che simuli realisticamente l’esperienza della guida in stato di ebbrezza con lo scopo di aumentare la consapevolezza dei rischi connessi e, dunque, convincere gli utenti ad evitare di mettersi al volante dopo aver consumato alcolici. Si parla invece di *microsuasione* quando elementi persuasivi sono integrati al progetto di una tecnologia con la quale, però, si vogliono perseguire altri scopi. Un’interfaccia digitale per la compravendita di prodotti online che ospita un conto alla rovescia entro il quale è necessario concludere l’affare agisce sul comportamento dell’utente non per renderlo in generale più disposto agli acquisti, ma perché lo scopo dell’interfaccia – vendere le merci disponibili sul sito – sia raggiunto più efficacemente. Nel primo caso la persuasione degli utenti è l’unico scopo primario della tecnologia, nel secondo è un mezzo al suo servizio.

Il potenziale persuasivo di una tecnologia dipende in buona misura dalle sue funzioni e dal contesto d’uso. Di conseguenza, anche le metodologie volte al suo sfruttamento dipendono dai medesimi fattori. Poniamo che un sistema, dato il suo contesto d’uso e le funzioni che svolge, sia percepito come un attore sociale. In casi simili, la persuasione può essere ottenuta almeno tramite tre strategie. La prima è la gratificazione: ad esempio, facendo sì che il sistema lodi gli utenti quando si comportano in modi desiderabili. Oppure, è possibile modellare direttamente attitudini e comportamenti – ad esempio, nei videogiochi, costruendo personaggi virtuali che incarnano modelli desiderabili di comportamento a cui gli utenti sono incoraggiati ad ispirarsi. Infine, si può far leva su pregiudizi, dinamiche e costumi per venire incontro alle aspettative degli utenti e migliorare l’accettabilità del sistema. Si immagini un assistente virtuale che svolga alcune funzioni tradizionalmente affidate ai medici di base con lo scopo di diminuirne il carico di lavoro. In sede di progettazione, dotarlo di un avatar che riprende gli stereotipi associati, in alcuni settori della nostra cultura, all’idea del medico esperto e affidabile (per esempio: maschio, di stirpe caucasica, sulla sessantina) potrebbe ben disporre l’utente verso la tecnologia e persuaderlo circa l’affidabilità dei risultati dell’interazione.

È a questo livello che emerge la domanda etica circa l’inganno e la sua liceità. Un sistema tecnologico che loda l’utente, ne mima il comportamento e simula altri caratteri dell’interazione sociale è pensato per indurre gli utenti a proiettare su di esso qualità tipicamente umane o animali, cosicché appaia familiare e sia agilmente accettato come “uno di noi”. Il fatto che gli utenti siano stimolati *by design* (cioè dal modo in cui il prodotto è progettato) a costruire modelli mentali illusori della tecnologia allo scopo di migliorare la qualità delle interazioni non può che far nascere dei sospetti.

In verità, l’utilizzo di strategie che rasentano o praticano l’inganno per influenzare il comportamento degli utenti va ben oltre l’ambito delle tecnologie digitali, interessando più o meno tutti i sistemi che prevedono una qualche interazione tra esseri umani e macchine. Come spiegato in (Adar, Tan, Teevan 2013, p. 1863), l’inganno si trova «ovunque nella progettazione di sistemi concreti, anche se raramente lo si presenta in questi termini». La pratica è diffusissima in virtù della sua efficacia, dal momento che

può riempire molti vuoti e risolvere molte tensioni che emergono in sede di progettazione da diverse preoccupazioni (ad esempio, per il bene individuale e il bene del gruppo), da diversi obiettivi (ad esempio, princìpi che confliggono), o da diversi stati del sistema (ad esempio, tra le prestazioni desiderate e effettive del sistema). In ogni situazione in cui si dà uno scarso allineamento tra *desiderata* (il modello mentale o le aspettative sviluppate dagli utenti) e realtà (il sistema effettivo), si dà anche l’opportunità di servirsi dell’inganno. Questo vuoto, estremamente comune, motiva e attiva l’inganno (ibid., 1864; trad. dell’autore).

Quando le aspettative dell’utente e le capacità del sistema non collimano, l’inganno intelligentemente progettato può riempire i vuoti e condurre ad un’interazione percepita come migliore di quanto realmente sia. Nascondendo mancanze sul lato delle prestazioni, mascherando momenti di incertezza, progettando interfacce antropomorfe e mantenendo le funzioni digitali coerenti con quelle di tecnologie antecedenti (skeuomorfismo) – ad esempio, associando il suono dell’otturatore che scatta alla funzione che permette di fare foto con lo smartphone – gli utenti sono continuamente ingannati allo scopo di offrire loro una esperienza migliore, che si sposi bene con le loro aspettative e, per quanto possibile e utile, con la loro esperienza passata. Così, l’innovazione tecnologica è facilmente accettata e prontamente inserita nello schema della vita quotidiana, a costo però di ingenerare possibili fraintendimenti circa le effettive capacità e qualità del sistema. Per riassumere, l’inganno *by design* rende possibile manipolare la comprensione che l’utente si forma di ciò che il sistema sia o meno in grado di fare, così da promuoverne l’accettazione e l’uso. Vertendo sull’inganno – sul dare ad intendere – si tratta di una strategia tanto funzionale quanto delicata da un punto di vista etico. Ciò vale ancor di più per i robot sociali.

1. Spinta gentile e robotica sociale

Per chiarire quali possano essere le problematiche etiche connesse all’atto di persuadere tramite la robotica sociale è necessario introdurre la nozione di *nudge*, o spinta gentile. Resa popolare da Thaler e Sunstein in un libro di grande successo (Thaler, Sunstein, 2008), la nozione si riferisce all’idea di progettare architetture decisionali in modo da orientare il comportamento delle persone coinvolte verso direzioni ritenute auspicabili. Con il termine “architettura decisionale” si intende il modo in cui diverse alternative sono presentate a chi deve scegliere tra di esse. A seconda del contesto e dell’ambiente in cui una scelta avviene, la struttura tramite cui varie alternative vengono presentate può influenzare significativamente, e in modo prevedibile, il comportamento di chi decide ed agisce. Un esempio è la sistemazione delle merci sugli scaffali dei supermercati. Gli scaffali sono un’architettura decisionale, in quanto la posizione dei diversi prodotti – le diverse alternative – influisce sulla propensione dei clienti ad acquistarli. Supponiamo, ad esempio, che prove empiriche suggeriscano una propensione da parte della clientela ad acquistare prodotti sistemati ad altezza occhi. Supponendo che sia nell’interesse dei clienti vivere una vita sana, si potrebbe spingerli gentilmente verso questo obiettivo piazzando ad altezza occhi prodotti sani e sistemando prodotti che lo sono meno in posizioni appena più scomode.

Il punto cruciale dell’intera teoria è che la spinta gentile non deve tramutarsi in vera e propria manipolazione. Perché ciò non avvenga sono necessarie almeno due condizioni. Primo: la spinta deve essere effettivamente gentile, facile da contrastare. Secondo: l’influenza deve essere esercitata esclusivamente nell’interesse di chi decide. Se le condizioni sono soddisfatte, la spinta gentile inclina, senza obbligare, le scelte e le azioni degli utenti verso gli esiti presumibilmente più allineati ai loro stessi interessi. Invece che ridurre l’autonomia e la libertà di scelta, la spinta gentile ha lo scopo di promuoverle rendendo più difficoltose, ma comunque percorribili, le scelte che le minacciano. Il nostro avventore potrà sempre acquistare il prodotto meno sano, ma per farlo dovrà chinarsi leggermente. Se questo piccolo disturbo lo orienterà verso prodotti più sani, grazie a cui vivrà una vita meno spiacevole e dolorosa, il suo interesse e la sua autonomia saranno state non solo rispettate, ma sostenute.

Detto ciò, viene subito da chiedere: perché i prodotti sani dovrebbero avere la precedenza sui prodotti convenienti o su quelli che hanno un minor impatto ambientale? Chi decide che la salute fisica debba avere la precedenza sulla stabilità economica o sullo stato dell’ambiente? Come possiamo essere certi di che cosa sia presumibilmente nel miglior interesse di chi sceglie, e che diritto abbiamo di esercitare tale, seppur minima, influenza? È pur vero che ogni architettura decisionale ha un’influenza su chi decide, e allora è meglio esercitare questa influenza in modo cosciente piuttosto che abbandonarsi al caso o, peggio, a chi ne sa piegare la logica ai propri interessi. Tuttavia, i problemi rimangono: *chi* ha il diritto di decidere *quali* direzioni siano da favorire tramite spinte gentili? E *perché*? E *fino a che punto* una spinta gentile non è una forma di manipolazione?

Ecco che emergono le diverse difficoltà etiche collegate alla spinta gentile. Innanzitutto, si fa avanti la questione del paternalismo. Pur semplificando, per paternalismo si intende per lo più la pratica di decidere per altri cosa è meglio per loro contro la loro volontà, pur godendo costoro in principio del pieno diritto di deciderlo in autonomia. Ciò vale soprattutto per questioni relative alla sfera individuale come, ad esempio, l’intimità, le opinioni etiche e politiche, i gusti e le propensioni individuali. Trattandosi di un’interferenza nell’autonomia dell’individuo – un carattere essenziale della sua dignità morale – il paternalismo è un comportamento eticamente controverso e richiede attente giustificazioni. La spinta gentile, orientando l’esercizio dell’autonomia, può sempre correre il rischio di tramutarsi in paternalismo e minacciare l’autonomia individuale. Ma le difficoltà non finiscono qui. C’è, ad esempio, la questione dei valori - quali valori promuovere tramite spinte gentili e quali marginalizzare? E, ancora, il problema di chi abbia il diritto di prendere una simile decisione: chiunque costruisca un’architettura decisionale? gli esperti? o tutte le parti coinvolte, attraverso processi democratici di discussione partecipata?

Le molte problematiche sollevate danno un’idea di quanto la questione sia delicata. Ancora non è stato chiarito, però, che cosa abbia a che fare con la robotica sociale. Il collegamento è tracciato proprio dalla captologia e dallo studio delle tecnologie persuasive applicato alla robotica sociale. I robot sociali infatti, in un certo senso, possono essere pensati come se fossero architetture decisionali. Il modo in cui sono progettati e funzionano può influenzare in maniera prevedibile il modello mentale che gli utenti costruiscono della tecnologia e, dunque, il loro comportamento. Il progetto, in altre parole, diventa uno strumento tramite cui controllare il modello mentale che gli utenti formano del sistema, dunque le aspettative che ne traggono e i modi in cui si comporteranno in relazione alla tecnologia. Si tratta di un’opportunità impossibile da ignorare. Si pensi, ad esempio, ad un sistema intelligente robotico pensato per la cura e la compagnia degli anziani. Dato il suo scopo, può sembrare ideale far sì che gli utenti sviluppino delle relazioni emotive con il robot: che si affezionino ad esso, ne abbiano a cuore le sorti, vogliano passarci del tempo insieme e magari confessargli le loro ansie o paure. Ora si supponga che alcuni tratti somatici – occhi grandi, viso rotondo, corporatura esile – o funzionali – ad esempio, la capacità di emettere frasi con contenuto emotivo (“sono felice di vederti” o simili) – abbiano l’effetto psicologico di indurre l’utente a proiettare caratteri umani sulla tecnologia, rendendo più facile lo sviluppo di legami emotivi. Perché non servirsi, per progettare robot sociali efficienti, di questo genere di spinte gentili, che non obbligano ma inducono l’utente a godere appieno della tecnologia (il che, si presuppone, sia nel suo interesse)?

La robotica sociale, soprattutto alla luce della captologia, difficilmente può rinunciare all’impiego di *robotic nudges* (Borenstein, Arkin, 2016):elementi che hanno il preciso scopo di suggerire all’utente caratteristiche o capacità da proiettare sul sistema cosicché l’interazione sia percepita come gratificante, ricca, coinvolgente e soddisfacente. Anche le spinte gentili mediate dai robot sociali, però, presentano le stesse sfide etiche che già conosciamo. Che ne è dell’autonomia dell’utente? Che ne è della sua dignità? Con quali mezzi e fino a che punto è permesso ingannarlo benevolmente, cioè allo scopo di aumentarne la soddisfazione? Chi è lecito che decida quali siano i presunti interessi degli utenti, in che modo, e perché? Il problema dell’inganno, a cui dedichiamo la prossima sezione, conduce a intricati conflitti di valore e rappresenta una delle questioni etiche più complesse in relazione alla robotica sociale.

1. Inganni robotici

La questione dell’inganno è endemica in robotica ed intelligenza artificiale. Se anche si tralascia il peso epistemologico dell’apparenza e del dare ad intendere in giunture cruciali per l’intero dominio di indagine – si pensi qui al gioco dell’imitazione di Turing (1950) – quanto detto su tecnologie persuasive e spinte gentili mediate attraverso robot sociali basti come riprova che l’inganno ne sia anche di fatto una componente ineliminabile. Ciò vale in particolare per la robotica sociale, dove alla dose di inganno già propria della tecnologia si somma la dose propria del contesto sociale, che senza piccoli o grandi inganni non può nemmeno essere concepito. In un certo senso, la tollerabilità delle relazioni sociali dipende dalla nostra abilità di celare opinioni, modulare l’onesta e dissimulare intenzioni in modo da non offendere il prossimo ad ogni piè sospinto e trasformare lo spazio sociale in un campo di battaglia. Il gioco dei rapporti sociali è delicato e difficile e i robot devono essere progettati per padroneggiarlo, se vogliamo che stiano efficacemente al nostro fianco come se fossero attori sociali. E siccome senza un pizzico di inganno si è condannati all’espulsione dal convivio sociale, bisogna aggiungerlo anche alle funzioni implementate nei robot sociali.

Un esempio forse innocuo, ma che aiuta ad avvicinarsi al cuore della questione, consiste nel dotare gli agenti conversazionali (provvisti o meno di corpi robotici) di funzioni che riproducono tratti del modo in cui comunichiamo a voce. Intonazioni, pause, ritmo, piccole bugie innocue (*white lies*), esclamazioni tipiche del parlato umano (“ehm”, “uhm”, “hmm”) sono spesso riprodotte in sistemi tecnologici con il solo scopo di mimare l’interazione tra umani, dando ad intendere in modo gentile che l’interazione con la macchina non sia poi così diversa, che ciò che vale nell’un caso vale anche nell’altro, e che quindi non ci sia troppo da preoccuparsi o da pensare: basta interagire come sempre (Isaac, Bridewell, 2017). Il vantaggio da un punto di vista dell’efficienza e della gradevolezza delle interazioni è evidente. Ma cosa succede se l’imitazione è così buona da suggerire in modo non aderente alla realtà che dietro al sistema ci siano desideri, passioni, intenzioni, secondi fini, propensioni, attitudini? E che succederebbe se le aspettative dell’utente circa le capacità e il futuro funzionamento del sistema dovessero venirne influenzate?

Riprendiamo l’esempio, meno innocuo, in cui robot per l’assistenza di soggetti vulnerabili (diciamo, anziani soli) siano progettati per indurre una certa indecisione sull’effettiva natura della tecnologia, lasciando aperto uno spiraglio per la convinzione, superficiale o profonda, che il robot abbia una propria vita emotiva, tenga all’utente, abbia dei bisogni emotivi e richieda anch’esso compagnia ed attenzione – che sia, in fondo, più un amico che uno strumento. Diventa immediatamente evidente la problematica etica: è lecito tutto ciò, o si tratta di un inganno perpetrato a danno di un soggetto già vulnerabile, e quindi esposto a manipolazioni e paternalismi di ogni genere? Quali sono i valori in gioco, che cosa merita rispetto e protezione, chi ha il diritto di pronunciarsi a riguardo e chi di stabilire quale sia il bene da perseguire in un simile contesto?

La questione è oltremodo spinosa e dà una chiara idea di quanto sia complesso il panorama delle problematiche etiche in robotica sociale. Si può portare un po’ di ordine analizzandola secondo le prospettive, già introdotte nel capitolo 4, dell’etica dei doveri e delle conseguenze.

Se si prende in considerazione l’etica dei doveri, la questione dell’inganno assume i connotati di una minaccia all’autonomia e alla dignità dell’utente, in particolar modo se quest’ultimo si trova in una posizione di vulnerabilità che lo espone ancor di più a simili rischi. In particolare, l’attenzione si concentra sulla linea sfumata che separa spinta gentile e manipolazione con l’intento di proteggere il diritto di ognuno di vivere la propria vita liberamente, secondo i propri valori, senza subire interferenze degradanti e condiscendenti. L’inganno robotico deve quindi essere valutato in base all’impatto sulla dignità e l’autonomia degli utenti, a prescindere dai suoi effetti.

La valutazione fornita da molti non è stata delle più accomodanti. Secondo Robert Sparrow (Sparrow, Sparrow, 2006), ad esempio, prodotti robotici progettati ad arte per invogliare la formazione di legami emotivi danno ad intendere che la tecnologia sia in grado di fare ciò che essa non è – stabilire legami emotivi autentici e bidirezionali. Dare ad intendere ciò a soggetti vulnerabili, bisognosi di compagnia e contatto, come ad esempio degli anziani soli, significa ingannarli in un modo che non è allineato al loro interesse ma, invece, ne sfrutta le debolezze, quale che ne sia il risultato percepito. Non importa, cioè, se gli utenti percepiscano o meno come gratificante l’interazione con i robot sociali, e magari sentano di essere più felici grazie alla nuova tecnologia. L’intera relazione, si sostiene, è basata su un’incomprensione orchestrata ad arte che sfrutta la proiezione di qualità emotive su robot che ne sono assolutamente privi. Qualsiasi beneficio si baserebbe comunque su un inganno, sull’indurre una rappresentazione falsata della realtà, il che di per sé vale come un’offesa alla dignità e all’autonomia dell’individuo, il quale si può dire goda in generale di un diritto a non essere preso per il naso – il cui rispetto è ancor più importante in situazioni di vulnerabilità. Un simile inganno robotico ne sminuisce la dignità, ne disconosce i diritti e indica una mancanza nel nostro dovere morale di provvedere al contatto umano di cui gli anziani hanno bisogno. Dalla prospettiva della deontologia, di conseguenza, il giudizio è negativo.

La rigidità della posizione appena esposta non può che attrarre critiche da parti più disposte a riconoscere i molti vantaggi, non solo pratici, associati all’adozione di spinte gentili robotiche. Squalificare senza appello ogni tipo di influenza sulle aspettative e i comportamenti degli utenti non sembra una posizione percorribile, sia per la difficoltà di metterla effettivamente in pratica, sia per il potenziale che andrebbe sprecato. Non si tratta, si badi bene, solo di rendere la vita più facile ai progettisti, che così avranno uno strumento in più per perseguire i propri fini. Si tratta soprattutto degli utenti, dei loro diritti e di ciò che ritengono abbia valore. Come si può pretendere di deprecare le relazioni emotive tra esseri umani e robot sociali, tacciandole di inconsistenza e falsità, se fossero proprio gli utenti ad affermare di trarne vantaggi moralmente significativi, come la sensazione di vivere una vita più ricca, meno sola, più attiva e gratificante? Guardando alle conseguenze di un inganno robotico, si suggerisce, sarà possibile distinguere tra progetti eticamente inaccettabili, in quanto irrispettosi della dignità degli utenti, e inganni accettabili, i cui effetti promuovono proprio ciò che sta loro a cuore. Tanto più che alcune forme di inganno robotico, come quelle che stimolano processi di antropomorfizzazione, fanno leva su tratti comuni della psicologia umana, a cui siamo esposti quotidianamente anche al di fuori dell’ambito della robotica senza che ciò generi eccessive preoccupazioni. Esagerare, o ingannare a scopo maligno, sono certamente pratiche da condannare, ma deve essere anche lasciato lo spazio per un uso intelligente e responsabile dell’inganno robotico, volto a fare gli interessi di chi si aspetta un miglioramento delle proprie condizioni di vita dall’interazione con la tecnologia.

In questo senso, l’idea di un inganno benevolo, finalizzato al bene degli ingannati, sembra avere più possibilità in ottica consequenzialista. L’approccio dell’etica delle conseguenze, grazie alla sua flessibilità, potrebbe offrire la metodologia più adatta a valutare caso per caso, lasciando aperto uno spiraglio per l’impiego etico delle spinte gentili in robotica sociale (Arkin, 2010; Sharkey, Sharkey, 2011). Certamente si tratta di un giudizio delicatissimo, che richiede grande cautela, ma se non altro si guadagna lo spazio per un dibattito più aperto – ma anche, quindi, più esposto ai rischi evidenziati dall’impostazione deontologica. Rimangono comunque, anche da questa prospettiva, molti problemi da discutere e numerosi fonti di possibili disaccordi. Rimangono difficoltà che hanno a che fare con la prevedibilità delle conseguenze o l’orizzonte temporale da prendere in considerazione. Rimangono problemi legati alla gestione delle aspettative degli utenti, che potrebbero essere eccessivamente sviate dall’uso smodato di elementi di progetto ingannevoli. Rimane da stabilire quali siano i valori che convergono a determinare il “bene” dell’utente, e rimane il dubbio su chi si trovi nella migliore posizione per definirli – soprattutto, di nuovo, in casi in cui il giudizio dell’utente stesso potrebbe essere indebolito dalla sua posizione di vulnerabilità.

Attorno alla domanda sui valori e sul bene dell’utente ruota l’ultimo gruppo di considerazioni che vale la pena tenere a mente. Abbiamo visto come atteggiamenti paternalistici costituiscano un rischio intrinseco di ogni tecnologia persuasiva. Per evitare che l’autonomia e la dignità degli utenti vengano violate è necessario fare sì che le decisioni circa i valori morali non siano risolte in modo autoritario da politici, tecnologi o chi per loro, ma sia garantita agli utenti la facoltà di esercitare il loro proprio giudizio nei limiti imposti dalla legge e dalla comune appartenenza al circolo dei soggetti morali. Detto ciò, fino a che punto è lecito dubitare della capacità degli utenti, per quanto vulnerabili siano, di distinguere la realtà dall’inganno? Quasi sempre si è supposto, finora, che l’utente occupi una posizione di svantaggio e sia alla mercé di un inganno che subisce, soprattutto in casi di vulnerabilità. Eppure, il presupposto potrebbe essere troppo forte. Tutti noi siamo soliti lasciarci coinvolgere in attività basate sull’inganno e sul dare ad intendere, in cui sospendiamo – ma solo parzialmente, e per trarne qualche beneficio – la consapevolezza di essere immersi in un inganno, in una dimensione di apparenza. Lo facciamo quando un film o una rappresentazione teatrale ci commuove, lasciandoci portare al pianto o all’euforia da eventi che non sono reali nel senso ordinario in cui usiamo la parola. Lo facciamo anche quando giochiamo ai videogiochi (si veda il capitolo 12) e lasciamo che la loro narrativa ci assorba completamente, tanto da ignorare stimoli corporei ben più concreti. In queste situazioni, che dall’esterno possono sembrare strane se non addirittura preoccupanti, spesso la consapevolezza di cosa è reale e cosa è finzione è solo volontariamente sospesa – mai perduta – allo scopo di godere della propria vita interiore, di fare esperienze emotive e cognitive gratificanti, che impattano poi significativamente sul benessere personale. Può quindi esistere un autoinganno etico, una ricerca soggettiva di esperienze immerse nella finzione che sia allineata a valori morali e resista a giudizi esterni circa la sua effettiva consistenza o significatività. Anche l’interazione con robot sociali potrebbe essere tra queste esperienze, il che richiede ulteriori cautele nella riflessione sul valore morale dell’inganno in tale contesto pratico.

1. Uno strumento metodologico: *Value Sensitive Design*

Alla luce di quanto detto, progettare robot sociali pone numerose sfide allo sforzo di allineare le tecnologie alle aspettative morali degli utenti e, più in generale, del contesto sociale in cui vengono introdotte. Non si tratta solo di fare sì che quanto comunemente riconosciuto come rilevante da un punto di vista morale sia effettivamente integrato al progetto tecnologico, ma anche di prendere coscienza di aspetti tuttora controversi – forse destinati a rimanere tali – e agire almeno in modo da mitigare i rischi connessi. Il primo caso, per fare un esempio, corrisponde all’allineamento dei prodotti tecnologici a valori ampiamente condivisi quali la non discriminazione, l’equità, l’inclusività, la sicurezza, e così via. Qui il problema non è tanto capire se il valore in quanto tale sia effettivamente un valore – cioè, se la non discriminazione o la sicurezza siano effettivamente valori all’interno della nostra cultura. Ciò, infatti, è dato per scontato. È invece assai meno scontato capire che cosa si intenda, quando si chiamano in causa i valori citati: sia in generale (che cos’è la non discriminazione?, che cos’è l’equità?), sia nella misura in cui riguardano la robotica sociale (cosa implica il valore della non discriminazione per la robotica sociale? Cosa implica l’equità nel medesimo contesto tecnologico?). Il secondo caso, invece, riguarda questioni come quella dell’inganno, cioè problematiche che dividono le opinioni e per la gestione delle quali non si può contare su un vasto consenso o su linee guida già istituzionalizzate. Come ci si deve comportare in relazione a problemi controversi, a cui non è possibile dare risposte che incontrino il favore dei più, è una domanda di ardua soluzione. Entrambe le difficoltà sono rilevanti per chi progetta nuove tecnologie come i robot sociali e richiedono di essere prese in considerazione sin dall’inizio, così che il prodotto tecnologico sia sviluppato tenendo conto del suo impatto morale e siano prese le dovute precauzioni perché i rischi siano tendenzialmente limitati e siano invece aumentati i benefici anche di carattere morale.

La sfida, dunque, consiste nel considerare questioni etiche già dalle prime fasi della progettazione, così da integrare al processo di sviluppo della tecnologia i valori morali pertinenti. Per riprendere le parole di Helen Nissenbaum (2001, p. 119; trad. dell’autore),

Gli ingegneri si trovano ora ad affrontare il dovere, per loro così poco familiare, di cogliere non solo quell’insieme usuale di proprietà che i sistemi incorporano, ma anche le proprietà morali di quegli stessi sistemi: bias, anonimità, privacy, sicurezza, eccetera. La sfida di costruire sistemi informatici si trasforma così in un’occasione di attivismo – di attivismo ingegneristico. Un attivismo che rappresenta un appello per il quale non solo molti si possono sentire impreparati, ma anche in sé difficile da soddisfare.

Come Nissenbaum nota, un ostacolo evidente consiste nella estraneità del compito rispetto alle usuali pratiche di progettazione. L’etica è un ingrediente nuovo ed eterogeneo da molti punti di vista (metodologico, epistemologico, contenutistico, formale). Perché l’integrazione abbia successo, c’è bisogno non solo di promuovere l’organizzazione di gruppi e sessioni di lavoro interdisciplinari, ma anche di elaborare procedure e metodologie grazie a cui le prospettive filosofiche, sociali ed ingegneristiche possano superare le diversità che le distinguono, collaborare all’analisi dell’impatto etico del prodotto tecnologico e lavorare insieme a possibili soluzioni.

Una delle metodologie che merita di essere discussa è il *Value Sensitive Design* (VSD). Inizialmente elaborato da Batya Friedman (1996) e successivamente sviluppato anche da altri autori come Peter H. Kahn (Friedman, Kahn, 2008; Friedman, Kahn, Borning, 2008) e Jeroen van den Hoven (2008), negli ultimi due decenni VSD ha ottenuto un ampio riconoscimento internazionale ed è stato applicato in diversi contesti di progettazione sensibili al richiamo della ricerca e dell’innovazione responsabile (Owen *et al.*, 2012). In breve, VSD è una metodologia di indagine che si ripropone di integrare sistematicamente considerazioni etiche nel processo di progettazione dedicando specifiche sessioni di lavoro all’analisi dell’impatto etico dei prodotti in via di sviluppo. La metodologia presenta una struttura tripartita, volta ad approfondire aspetti diversi ma strettamente connessi tra loro. Inoltre, la struttura è pensata non come lineare, ma iterativo: le diverse fasi non sono semplicemente da completarsi una dopo l’altra, ma rappresentano momenti diversi di una indagine riflessiva, pronta a tornare sui propri passi e rivalutare il lavoro svolto alla luce dei risultati conseguiti.

Illustriamo brevemente in che cosa consiste la strategia di indagine VSD. Innanzitutto, si afferma che più il gruppo di lavoro include membri dotati di competenze eterogenee (per quanto rilevanti all’indagine), migliori saranno i risultati ottenuti. Una precondizione, quindi, consiste nel formare il gruppo di lavoro in modo equilibrato, tenendo conto delle diverse competenze necessarie. Il gruppo di lavoro sarà poi impegnato in tre compiti da intendersi, come anticipato, in maniera iterativa. Un primo compito consiste nell’indagine concettuale relativa all’impatto etico del prodotto in questione. Qui si dovrà determinare quali valori etici siano chiamati in causa dal progetto, secondo quale gerarchia sarebbe idealmente opportuno organizzarli per risolvere potenziali conflitti, e secondo quali procedure gestire i *trade-off* associati a quest’ultimi. Accanto all’indagine concettuale deve anche essere svolta una ricerca empirica, volta a comprendere meglio i contesti di uso del prodotto, le aspettative morali che gli utenti si troveranno a nutrire nei suoi confronti, i modi in cui esso influenza le esperienze, le attitudini e le idee degli utenti, e così via. Se l’indagine concettuale usa gli strumenti della riflessione razionale per prendere coscienza del profilo valoriale connesso al prodotto tecnologico, la ricerca empirica è volta a registrare cosa avviene – o potrebbe avvenire – una volta che il prodotto sia nelle mani dell’utente finale. Entrambe le prospettive sono fondamentali per comprendere appieno il possibile impatto etico di un prodotto tecnologico e, allo stesso tempo, realizzare quali azioni intraprendere per allinearlo alle aspettative che ci si propone di soddisfare. L’ultimo momento, infatti, ha proprio lo scopo di esaminare diverse possibilità progettuali alla luce di quanto emerso, così da optare per il progetto meglio in grado di ridurre i rischi e promuovere valori etici condivisi. Non è escluso, poi, che quanto emerso in sede di analisi concettuale e ricerca empirica non stimoli nuove soluzioni progettuali particolarmente allineate ad alcuni valori emersi come rilevanti, offrendo così un ventaglio più ampio di opzioni alla decisione finale. Questo non significa che ogni conflitto e ogni rischio debba essere eliminato, ma che si sviluppi almeno la consapevolezza di entrambi, in modo da esercitare la cautela necessaria in altre sedi o di attivare altre strategie per il monitoraggio degli impatti del prodotto.

VSD permette una gestione consapevole della dimensione morale sollevata dalla progettazione di ogni genere di tecnologia e può risultare uno strumento utile per affrontare le molte sfide etiche della robotica sociale. È bene però sottolinearne i limiti, così da promuoverne un uso consapevole. È importante in particolare capire che VSD è uno strumento pensato per la progettazione di tecnologie allineate a valori etici, ma che la progettazione non è necessariamente la sede più consona per prendere tutte le decisioni circa la dimensione morale implicata nello sviluppo delle nuove tecnologie. Alcune decisioni di ampia portata richiedono il coinvolgimento sia di diverse competenze che di diverse procedure. Si pensi, ad esempio, alle decisioni circa quale valore morale sia degno di considerazione e quale no, a quale valore si debba in generale dare priorità in caso di conflitti o, ancora, circa l’interpretazione da dare a un certo valore così che sia possibile calarlo nelle situazioni concrete. Simili questioni, in un certo senso, rappresentano la condizione perché la metodologia VSD sia utilizzabile, ma possono rischiare di essere fagocitate dalla metodologia stessa. In altre parole, può sembrare alle persone coinvolte nel processo del VSD che sia loro compito non solo comprendere quali problematiche etiche possano emergere dall’uso di un prodotto progettato in un certo modo e come sia possibile affrontarle attraverso scelte mirate, ma anche deliberare su quali valori siano davvero importanti, sulla definizione da dare loro, e sulla loro gerarchia.

Lasciare che questioni tanto cruciali siano decise solo da chi è coinvolto nella progettazione di tecnologie conduce alla cosiddetta *tecnocrazia*, in cui decisioni che riguardano la collettività sono prese attraverso processi e in sedi che la escludono. I rischi posti dalla tecnocrazia sono particolarmente gravi da un punto di vista etico, in quanto implica atteggiamenti paternalistici e manipolatori che toccano sfere cruciali della vita morale individuale quali l’autonomia, l’autodeterminazione e la dignità. Di conseguenza, è fondamentale saper distinguere tra le decisioni che possono essere prese in sede di progettazione e le decisioni che invece richiedono un contesto differente. Se il tentativo di allineare le nostre tecnologie a valori socialmente accettati è una questione tecnica, la determinazione, la definizione e la specificazione dei valori è una questione sociale. Una società democratica non può accettare che questioni di valore siano decise esclusivamente in sede di progettazione, ma esige che la collettività sia adeguatamente rappresentata, che i processi decisionali siano pubblici, trasparenti e scrutinabili, che la responsabilità per le decisioni prese sia chiara e distribuita in modo equo, e che vengano istituite politiche etiche chiare, adottabili e passibili di controllo circa il loro rispetto.

Per concludere, il metodo VSD è pienamente efficace solo in un contesto sociale che assume con serietà la sfida etica posta dall’introduzione di nuove tecnologie. Non è possibile, e nemmeno giusto, addossare la piena responsabilità dell’allineamento etico delle tecnologie ai soli gruppi coinvolti nel loro sviluppo. Senza un adeguato supporto istituzionale, inclusivo e trasparente, l’obiettivo non può essere raggiunto. Assicurarsi che le nostre tecnologie, e in particolare i robot sociali, limitino i rischi e promuovano valori ampiamente condivisi è un compito che non può essere affidato a un solo attore, ma richiede l’impegno di una società nel suo insieme. Solo uno sforzo collettivo volto a sostenere lo sviluppo eticamente consapevole di robot sociali può garantire una reale e motivata fiducia nei vari attori coinvolti – ingrediente indispensabile per coglierne i vantaggi promessi senza venire meno alle aspettative etiche che è lecito nutrire a riguardo.

Capitolo 10

Sex robots

Maurizio Balistreri

1. Introduzione

Una critica che ricorre nel dibattito afferma che i sex robot sarebbero incompatibili con la moralità, perché – dietro il loro acquisto ed uso – ci potrebbero essere soltanto motivazioni poco apprezzabili.

Il capitolo, dopo aver spiegato cosa sono i sex robot, mostra che le ragioni per acquistare un sex robot possono essere diverse, e che dietro questa scelta potrebbe esserci anche una preoccupazione sincera, ed apprezzabile per il piacere di altre persone, per la stabilità di una relazione affettiva e per il proprio benessere. Nella seconda parte del capitolo esamineremo se stuprare un robot è segno di un carattere vizioso. I sex robot non possono essere aggrediti perché, dato che sono macchine, non hanno una volontà e di conseguenza non possiamo costringerli a fare cose che non desiderano o che non gradiscono. Ad ogni modo, le loro espressioni facciali, la voce, la postura e le reazioni possono essere programmate per farci credere che non vogliono fare sesso o che si eccitano quando vengono maltrattati. La maggior parte dei sex robot che oggi troviamo in commercio hanno fattezze femminili. Alla fine del capitolo pertanto discuteremo la questione se i sex robot fanno aumentare le violenze sulle donne e promuovono nel fruitore l’immagine della donna come oggetto sessuale e sempre disponibile.

1. Che cosa sono i sex robot?

I sex robot sono macchine dotate di una qualche forma di intelligenza artificiale progettate per il piacere sessuale dell’utente. I primi sex robot sono stati messi in vendita dalla True Companion nel 2010 per meno di 10.000 dollari: Roxxxy è la versione femminile e sia il suo aspetto che il suo carattere possono essere personalizzati. Farah è riservata e timida, Wendy è avventurosa e spregiudicata, Susan pronta a soddisfare qualsiasi fantasia, Yoko invece è così giovane – appena 18 anni – ed aspetta soltanto di imparare. Ad oggi il modello di sex robot tecnologicamente più avanzato è quello prodotto dalla società Abyss Creations e prevede una testa robotica parlante e smontabile sul corpo di una bambola (RealDollx). La testa robotica non è dotata di videocamera ma può comunque rispondere alla voce come fa qualsiasi assistente digitale ed, anche in questo caso, l’utente potrà scegliere tra un ventaglio di personalità. I sex robot della Abyss Creations non sono ancora in vendita, ma i prototipi, Harmony e Solana, sono già stati presentati ed è stato stimato che il loro costo si aggirerà intorno ai 20.000 dollari (Devlin, 2018). Di questi robot esistono anche delle copie virtuali con le quali gli utenti possono interagire attraverso la Realbotix Harmony App scaricabile a pagamento sul loro smartphone. Bambole di silicone in grado di muovere la testa, la bocca, gli occhi, battere le ciglia e anche rispondere ad alcune semplici domande sono già state pubblicizzate e vendute anche da altre società. Altri modelli – come ad esempio quelli della linea della Android Robot Store – sono dotati di esoscheletri che permettono loro sia di camminare che di muovere le braccia (Forsch-Villaronga, 2020). Roxxxy, infine, avrebbe un’apertura vaginale che massaggia qualsiasi cosa venga inserita. Al momento la maggior parte dei sex robot ha le sembianze di donna (Wosk, 2015; Devlin, 2018): la versione maschile, però, non è soltanto un’invenzione cinematografica o letteraria (McEwan, 2019). Ad esempio, dopo aver messo in commercio Harmony e Solana, l’Abyss Creations ha presentato Henry; la True Companion, invece, ha nel suo catalogo Rocky – che è la versione maschile di Roxxxy (Devlin, 2018). Dato, comunque, che a volte è possibile scegliere le sue principali caratteristiche, chi acquista un sex robot potrebbe anche optare per un’identità sessuale meno definita – *queer*: ad esempio, potrebbe scegliere un robot con il corpo di una donna ma i genitali e la voce di un uomo oppure con il corpo di un uomo ed i genitali, il timbro della voce e anche il seno di una donna. Si può, comunque, intervenire sull’identità sessuale del robot anche dopo averlo acquistato. Un *transgender* *converter* già esiste e permette di sostituire la vagina di qualsiasi Realdoll con un pene. Tuttavia, i sex robot non devono per forza essere umanoidi ed assomigliare ad un essere umano, in quanto potrebbero evocare un personaggio mitologico, di fantasia o un animale (Devlin, 2018). Un grande quantità di giocattoli del sesso che oggi viene prodotta non rappresenta più i genitali: è possibile che se i sex robot non assomigliassero agli esseri umani potrebbero dare più piacere. Se avessero, ad esempio «un corpo di velluto o di seta, sensori e genitali misti e tentacoli al posto delle braccia» (ivi, p. 266), chi interagisce con loro proverebbe piacere con tutti i sensi. (In questo modo, per altro, i sex robot non rappresenterebbero più gli esseri umani e, pertanto, non potrebbero essere accusati di promuovere l’immagine degli esseri umani come meri oggetti sessuali.) Se, poi, collegati a tecnologie che consentono il trasferimento di dati da un computer all’altro, i sex robot potrebbero essere usati anche per fare sesso a distanza o con una comunità di persone, in quanto quello che una persona fa al suo robot potrebbe essere trasmesso ad un’altra persona (attraverso la strumentazione che essa indossa o il sex robot con il quale, a sua volta, interagisce).

Quello che comunque distingue i sex robot dalle bambole (*sex doll*) è il fatto di possedere software di intelligenza artificiale che controllano il loro comportamento (socialità, apprendimento, mimica ecc.). Per la descrizione delle componenti software possiamo seguire Forsch-Villaronga e Poulsen (2020). La maggior parte dei sex robot oggi in commercio ha un programma di riconoscimento vocale che permette loro di partecipare ad una conversazione e di rispondere a domande semplici dell’utente. Alcuni modelli, poi, muovono anche le labbra in sincrono con la parola e prevedono la funzione di “comando vocale”: cioè, l’utente può controllare alcune loro funzioni o movimenti ‘a voce’. Gli aggiornamenti software sono automatici perché i sex robot possono connettersi alla linea wifi: attraverso la rete, poi, essi possono accedere ai motori di ricerca per rispondere alle domande degli utenti. I sex robot che hanno sensori interni e/o esterni sono in grado sia di riconoscere le azioni dell’utente sia di rispondere agli input con la voce (e gemiti) o con i movimenti del corpo o della testa. Se il robot non è autonomo, la risposta può essere attivata dall’utente, a voce oppure manualmente. L’utente può ordinare a Rocky di penetrarlo, a Roxxxy, invece, di muoversi avanti e indietro. I sex robot che possiedono anche programmi basati sul *deep learning* (ad esempio, Emma ed i RealDollx) possono apprendere gli interessi e le preferenze dell’utente attraverso la conversazione: attraverso, invece, l’interazione, ai sex robot della True Companion possono essere insegnate nuove personalità. L’abilità di imparare, però, è limitata soltanto alla conversazione e alle espressioni facciali. È comune, inoltre, che i sex robot abbiano software che regolano le loro espressioni facciali, in quanto questo rende più facile l’interazione e permette all’utente di simpatizzare con loro più facilmente. Ad esempio, i RealDollx possono sorridere, guardare in maniera minacciosa o sembrare scioccati: inoltre, possono anche sbattere le palpebre velocemente, alzare ed abbassare le sopracciglia. Oltre, poi, ad avere software che regolano il movimento della testa e del collo, alcuni sex robot (ad esempio Emma e i RealDollx) hanno un sistema di riscaldamento interno controllabile dall’utente.

Per il momento si può avere un rapporto sessuale con un sex robot soltanto acquistandoli online attraverso siti e-commerce che garantiscono ai clienti interessati una consegna nella massima discrezione. Non è possibile, invece, avere rapporti occasionali a pagamento, ma le cose in futuro potrebbero cambiare, perché case di appuntamento con le bambole (i *sex doll brothel*) sono già state aperte in alcune città (ad esempio, Barcellona, Parigi, Torino) con alterne fortune (Devlin, Lake, 2018). A Houston (Texas) il consiglio comunale si è opposto alla loro apertura accogliendo la preoccupazione che l’offerta delle *sex doll* potrebbe incoraggiare la tratta degli esseri umani (*sex-trafficking*).

1. Chi fa sesso con un robot è una persona viziosa?

Un tempo i giocattoli del sesso erano oggetti che non si potevano menzionare (erano un taboo!), ma oggi essi sono ampiamente usati. Non soltanto il loro uso privato non è più considerato motivo di vergogna, ma vengono sempre più percepiti come uno strumento che contribuisce al benessere personale (Kontula, 2009; Forbes, 2016). Se c’è ancora qualche remora nel loro acquisto è forse soltanto perché gli utenti si divertono a preservare l’aspetto proibito dell’oggetto e trovano la cosa eccitante (Piha *et al.*, 2018). Del resto, che cosa ci sarebbe di male nel ricercare il piacere attraverso oggetti – da quelli più semplici a quelli ad ‘alta definizione’ – che rappresentano alcune parti del corpo? Non si arreca un danno a nessuno e si possono trascorrere piacevoli momenti che possono aiutare a superare lo stress quotidiano o ad affrontare le fatiche e le responsabilità della giornata successiva. L’idea, cioè, di fare sesso con un robot può sembrare anche bizzarra e ad alcuni ripugnante; si tratta, però, soltanto di un atto di autoerotismo, molto elaborato, ma comunque innocente.

Anche se l’uso dei giocattoli per l’autoerotismo viene sempre meno percepito come un problema morale, alcuni ritengono che sia la produzione sia l’uso dei sex robot aprano questioni morali importanti. Una critica che ricorre nel dibattito afferma che i sex robot sarebbero intrinsecamente immorali, perché – dietro il loro acquisto (ed uso) – ci potrebbero essere soltanto motivazioni indecenti. Il fatto che i sex robot abbiano soprattutto le sembianze di donna non sarebbe un caso: l’ipotesi, infatti, è che essi sarebbero acquistati soltanto da uomini che hanno difficoltà ad avere relazioni con le donne o che desiderano un partner sottomesso, sempre disponibile e senza alcuna pretesa. Anche a loro – si dice – piacerebbe avere accanto a sé una donna e alle volte in passato l’hanno avuta ma essi ricorrono ad un sex robot o comunque sognano di farlo (Gutiu, 2012; 2016) perché non sarebbero in grado di costruire o di portare avanti relazioni durature ed importanti, desiderano un partner controllabile e non sono in grado di riconoscere alle donne la loro stessa dignità (Sparrow, 2017). Del resto, – sostiene Gutiu (2016) – la relazione con una macchina è più facile rispetto a quella con una donna: non c’è il rischio di soffrire e la sua presenza non implica alcun vincolo o limitazione; quando, poi, il sex robot viene a noia si può sempre spegnere, sostituire o cambiargli il carattere.

L’ipotesi che i sex robot siano ricercati soltanto come partner sostitutivo da persone sole appare plausibile, in quanto è soprattutto a questo tipo di persone che si rivolgono le società che li producono.

Nel video promozionale della AI-Tech, una azienda cinese leader nel campo della robotica, vediamo un uomo camminare tra le bottiglie di alcol, consumate la sera prima, e vecchie fotografie della ex fidanzata. Non occorre molta fantasia per immaginare che cosa è successo: la relazione, per qualche motivo, è finita e lui, che ripensa a lei con rimpianto, non riesce ancora ad accettarlo e continua a star male. Le cose, però, cambiano: portano una cassa e dentro c’è Emma che promette di essere la soluzione a tutti i suoi problemi, perché è la compagna che ogni uomo aspira ad avere (Devlin, 2018). Secondo Michael Hauskeller soltanto le persone che sono abituate a trattare o considerare gli altri solo come mezzi possono pensare che un robot possa sostituire un essere umano: «Un amante reale può essere rimpiazzato da un robot senza alcuna perdita se, e soltanto se, le altre persone non possono essere per noi mai qualcosa di più di un semplice mezzo, se esse già sono (per noi), per tutti i nostri intenti e scopi, meramente dei robot del sesso mascherati» (Hauskeller, 2014, p. 14, trad. dell’autore).

Non è facile realizzare ricerche scientifiche che permettano di stabilire con precisione quali sono le motivazioni che spingono o potrebbero spingere le persone ad acquistare e fare sesso con un sex robot. Tuttavia, dal fatto che una persona decida di acquistare un sex robot non si può concludere che, allora, vuole rimpiazzare un partner umano con uno robotico o che sia sessista, intemperante e violenta. Le ragioni, infatti, potrebbero essere altre e moralmente giustificabili (come, ad esempio, la ricerca del proprio piacere) o anche apprezzabili (ad esempio, la ricerca del piacere del proprio partner). I sex robot sono un fenomeno abbastanza recente e non è semplice raggiungere i potenziali acquirenti, in quanto la maggior parte dei sex robot in commercio viene acquistata (o pre-ordinata) online. Inoltre, si tratta di giocattoli sessuali che molte persone guardano ancora con sospetto e diffidenza – soltanto un uomo su 4 (25%) e una donna su 10 (10%) farebbe sesso con un robot (YouGov, 2017) – per cui molte persone che acquistano sex robot potrebbero non aver voglia di raccontarlo. Anche se però non abbiamo ancora evidenze empiriche sia per quanto riguarda le motivazioni dietro al loro acquisto (o uso) sia per quanto riguarda il grado di soddisfazione degli acquirenti, conosciamo quali sono le ragioni che spingono le persone ad acquistare un giocattolo sessuale. A partire, pertanto, da quello che sappiamo riguardo al mercato dei giocattoli sessuali, possiamo avanzare delle ipotesi sui sex robot che possono servire a ragionare sulla moralità del loro impiego. C’è bisogno, comunque, di più ricerche, in quanto le ipotesi avanzate sulle implicazioni benefiche o dannose dei sex robot non sono a tutt’oggi corroborate da osservazioni sperimentali adeguate.

Innanzi tutto, qualcuno potrebbe ricorrere ad un sex robot per avere una vita sessuale più soddisfacente. I sex robot, infatti, sono un giocattolo sessuale che può essere acquistato da qualsiasi persona: è vero che al momento la maggior parte dei sex robot presenta l’aspetto di donna (Wosk, 2015), ma essi possono assomigliare anche ad un uomo ed essere adattabili a qualsiasi gusto o preferenza. Secondo Sparrow (2021), i sex robot sono sopravvalutati come strumenti che aiutano la masturbazione: non soltanto la natura fisica di questi artefatti pone limiti importanti alle fantasie sessuali, ma – dato che il piacere è raggiunto con l’immaginazione – «a meno che una persona non sia eccitata al pensiero di fare sesso con un robot, uno farebbe meglio a masturbarsi alla vecchia maniera» (ivi, p. 33, trad. dell’autore). Tuttavia, alcune persone – potendo scegliere – potrebbero preferire masturbarsi o raggiungere il piacere sessuale con un sex robot piuttosto che da sole o con altri tipi di giocattoli sessuali. Ad esempio, una persona potrebbe scegliere di acquistare o fare sesso con un robot perché ha fantasie sessuali particolarmente eccentriche che non riesce a soddisfare (o preferisce non farlo) con il proprio partner, perché non sono corrisposte o semplicemente preferisce non confessarle (Balistreri, 2018). Il sesso con un robot, poi, potrebbe permettere di provare situazione estreme – dove, ad esempio, c’è una persona sottomessa e l’altra dominante o giochi che prevedono anche la violenza, in quanto il sex robot può essere danneggiato o distrutto, ma non è possibile arrecargli sofferenza o ucciderlo. Naturalmente qualcuno potrebbe avere fantasie che nessun essere umano potrebbe soddisfare. Ad esempio, la bestialità non è offensiva o disumanizzante per l’essere umano (Soble, 2007). Il problema riguarda casomai gli animali: è vero, infatti, che il sesso con gli animali non è necessariamente crudele (Dekkers, 1994) ma spesso è violento e può comportare la loro morte (Singer, 2001). C’è, poi, la questione del consenso: un animale può essere consenziente e come facciamo a stabilirlo? La soluzione a questi problemi potrebbero essere sex robot che hanno le fattezze degli animali.

Un sex robot, inoltre, potrebbe aiutare qualsiasi persona a conoscere ed a esplorare la propria sessualità e ad imparare a dare piacere non soltanto a sé stessa ma anche alla persona con cui fa l’amore. Altre persone, poi, potrebbero scegliere di avere accanto a sé un robot non tanto per immaginare di avere un rapporto con una persona reale, ma solamente perché desiderano accanto ‘qualcuno’ che faccia loro compagnia e con cui stare insieme quando la solitudine diventa insopportabile. In questa categoria potrebbero rientrare sia le persone che, ad esempio, hanno perso il proprio partner o che non vogliono impegnarsi in una nuova storia d’amore o avere incontri occasionali sia quelle che non sono in grado di avere rapporti o di stare a contatto fisico con altre persone. A questo si aggiunga che, con l’aiuto dei sex robot, le persone che soffrono di ansia sociale o che, per altri motivi, non hanno relazioni potrebbero (ri)acquistare fiducia in se stesse e nelle altre persone, in quanto – come raccontano le persone che vivono con una bambola (Devil, 2018) – le persone con un partner sintetico costituiscono una comunità che condivide le proprie esperienze.

C’è chi ipotizza che con l’uso dei sex robot le persone con disabilità o malformazioni potrebbero avere una vita sessuale più completa, senza bisogno di ricorrere alla prostituzione o all’assistenza sessuale (Jecker, 2021). La libertà sessuale è un diritto umano fondamentale (Fosch-Villaronga, Poulsen, 2020). La possibilità di scegliere la propria vita sessuale e il proprio partner senza impedimenti da parte dello stato, della famiglia o di terze persone permette di avere un maggior controllo sulle proprie esistenze e di fare esperienze piacevoli con un effetto positivo anche sulla propria salute (McArthur, 2017). Questo è stato ribadito dai programmi per la salute sessuale dell’Organizzazione mondiale della sanità e dalla dichiarazione sui diritti sessuali della *World Association for Sexual Health*. Tuttavia, ci sono persone disabili che non sono in grado di avere una vita sessuale o di raggiungere una vita sessuale veramente soddisfacente (Ulivieri, 2014; Forsch-Villaronga, Poulsen, 2020): ad esempio, non possono masturbarsi oppure, a causa delle loro disabilità, non riescono a trovare una persona con cui avviare una relazione o avere rapporti sessuali (Appel, 2010). Queste persone possono avere una vita sessuale soltanto ricorrendo al sesso a pagamento o – quanto meno nei paesi in cui questo servizio è legalmente riconosciuto – ricorrendo ad un assistente sessuale. I sex robot non sembrano da scartare per principio, in quanto potrebbero essere un’altra opzione. Alcuni, però, temono che con l’impiego dei sex robot come strumenti ‘assistenziali’ le persone con disabilità potrebbero avere ancora più difficoltà ad avere relazioni affettive, in quanto, una volta che avranno accanto a sé un robot, esse smetteranno di cercare un partner umano (Forsch-Villaronga, 2020). Altri, invece, ritengono che sarebbe comunque sbagliato affidare la cura delle persone con disabilità alle macchine, in quanto l’attività di cura, per sua natura, sarebbe compito degli esseri umani (Sparrow, 2017). Tuttavia, i sex robot potrebbero essere un’opzione in più per le persone con disabilità.

I sex robot, comunque, potrebbero essere utilizzati non soltanto dalle persone sole, ma anche da quelle coppie (di qualsiasi orientamento) che desiderano esplorare nuove fantasie o ravvivare una relazione spenta. Quando pensiamo al sesso a tre (*threesome* o *ménage à trois*) immaginiamo, in genere, tre persone, ma “una” di queste – almeno nella fantasia – potrebbe essere anche un sex robot (Liberati, 2021). In caso, invece, di disfunzione erettile, i sex robot potrebbero essere anche una soluzione terapeutica. I sex robot, poi, potrebbero aiutare le persone che non hanno il partner vicino – secondo la pornostar Jessica Ryan una bambola di silicone è lo strumento ideale per le relazioni a distanza (Devlin, 2018) – o che, per qualche ragione, non possono avere – con lui o con lei – rapporti sessuali (ad esempio, il partner potrebbe essere in quarantena o aver subito una operazione ed essere in convalescenza).

Qualcuno ha anche suggerito che i sex robot potrebbero essere impiegati per aiutare gli autori di reati sessuali (o chi sente il desiderio di compierli) a controllare le loro azioni nei confronti di donne e bambini (Prigg, 2014). In base a quest’ipotesi, cioè, i sex robot potrebbero essere per i criminali sessuali quello che il metadone è per i tossicodipendenti: con il loro aiuto essi potrebbero perdere o avere un desiderio meno irresistibile di abusare di donne o bambini o avere una maggiore capacità di controllarlo. L’uso di sex robot su autori di reati sessuali è l’area più controversa del campo della robotica: il loro impiego non è al momento corredato da evidenze empiriche che ne sostengono l’efficacia. Lo scetticismo e le resistenze, poi, non mancano poiché i robot sono percepiti come una minaccia e perché c’è la preoccupazione che questo porti ad un aumento incontrollato della violenza (Zara, 2018). C’è il rischio che tali dispositivi promuovano, invece che scoraggiare, le fantasie dei criminali sessuali, contribuendo a colmare il divario esistente tra fantasia e realtà (Brown, Shelling, 2019). Inoltre, alcuni ricercatori ritengono che il dibattito sul possibile uso dei sex robot per finalità terapeutiche non consideri con sufficiente attenzione il fatto che l’abuso ha a che fare più con un desiderio di dominio e di controllo che con la sessualità (Facchin, Barbara, Cigoli, 2017; Balistreri, 2018). Comunque, anche se è stato affermato che l’idea non avrebbe nemmeno dovuto essere avanzata (Strikwerda, 2017), l’ipotesi di ricerca è interessante e merita di essere studiata (Moen, Sterri, 2018), in quanto potrebbe aiutare a ridurre il numero di violenze e di reati sessuali (Zara, 2018).

1. C’è qualcosa di sbagliato nello ‘stuprare’ un robot?

È stato affermato, comunque, che è più difficile mettere in dubbio il carattere vizioso di una persona che decide di comprare e fare sesso con un sex robot che sembra non gradire o rifiutare eventuali avances sessuali ovverosia che è stato programmato per non assecondare il potenziale partner o resistergli. I sex robot non possono essere aggrediti perché, dato che sono macchine, non hanno una volontà e di conseguenza non possiamo costringerli a fare cose che non desiderano o che non gradiscono. Ad ogni modo, le loro espressioni facciali, la voce, la postura e le reazioni possono essere programmate per farci credere che non vogliono fare sesso o che si eccitano se vengono maltrattati. Robot di questo tipo sono già in commercio: ad esempio, esiste una versione di Roxxxy della True Companion che non a caso viene presentata sulla pagina web della società come la ‘Frigid Farah’. Per fare sesso con Samantha della società robotica Synthea Amatus c’è bisogno, invece, di conquistarla: se non viene accarezzata e baciata, si mostra poco disponibile e seccata e dice di ‘no’ (Miley, 2018). Si possono scegliere comunque, anche altri modelli: i sex robot della Z-Onedoll gemono se vengono presi a schiaffi e i gemiti aumentano con l’aumentare delle violenze (Devlin, 2018). Secondo Sparrow (2017; 2020) non ci vuole molto sforzo per comprendere che chi è interessato a vivere con un robot l’esperienza di una violenza sessuale è una persona violenta, in quanto soltanto una persona violenta potrebbe avere piacere a stuprare un robot. Fa violenza su un robot soltanto perché desidera stuprare una persona reale (Gildea, Richardson, 2017).

I sex robot che vengono oggi prodotti non provano emozioni e sentimenti e non sono nemmeno autocoscienti: siamo ancora lontani dal poter interagire con robot che meritano di essere trattati come persone. Secondo Sparrow, comunque, il fatto che una persona faccia sesso con un sex robot programmato per rifiutare qualsiasi approccio sessuale, e goda anche a stuprarlo, ci direbbe che persona è, in quanto non sarebbe possibile eccitarsi con questi giocattoli, senza immaginare veramente una violenza. Una persona, cioè, può eccitarsi ad “infliggere sofferenze” ad un robot soltanto perché è crudele. Le reazioni di molte persone al video di Youtube in cui gli ingegneri della Boston Dynamics prendono a calci il cane robotico Spot per mostrare la sua stabilità avvalorano la conclusione di Sparrow[[45]](#footnote-45). Nello spazio dei commenti alcune persone affermano di sentirsi molto dispiaciute per Spot. Altre, invece, scrivono che gli ingegneri sono persone malvagie perché si compiacciono di quello che fanno. Questa conclusione – ammette Sparrow – può sembrare avventata in quanto un robot non può soffrire, ma non è sbagliato trarre conclusioni sul carattere di una persona da come tratta una macchina, in quanto il modo in cui trattiamo la rappresentazione delle cose rivela quali pensieri e sentimenti abbiamo.

Da un recente sondaggio sappiamo che il 49% del campione intervistato si aspetta che nei prossimi cinquant’anni avere rapporti sessuali con un robot diventerà qualcosa di normale (YouGov, 2017). Non abbiamo, comunque, informazioni sul tipo di carattere delle persone che acquistano o che acquisterebbero volentieri sex robot programmati ad opporre resistenza a qualsiasi approccio sessuale. La possibilità di raggiungere il piacere simulando una violenza sessuale non nasce con i sex robot. Giochi di questo tipo, infatti, possono essere praticati anche con il proprio partner o con qualsiasi persona: è sufficiente che una persona reciti la parte della vittima e l’altra quella dell’abusante. Non è necessario, poi, che sia sempre la stessa persona a interpretare la parte della vittima: in qualsiasi momento le parti si possono invertire e chi era lo ‘schiavo’ può diventare il ‘padrone’. Questi giochi, poi, si possono praticare non soltanto nel mondo reale con altri adulti consenzienti, ma anche da soli – attraverso un avatar – entrando in una realtà virtuale e giocando – nel ruolo, ad esempio di un soldato o di un cavaliere medievale – a stuprare donne e bambini (Luck, 2009). Secondo Sparrow, questi comportamenti (non importa se praticati su una persona consenziente, un robot o personaggi virtuali) rivelerebbero che una persona è veramente crudele e violenta, in quanto essi esprimerebbero pensieri, emozioni e fantasie del suo carattere (Sparrow, 2020).

Tuttavia, il rapporto tra simulazione di comportamenti e carattere è più complesso di come lo presenta Sparrow. Innanzi tutto, se il modo in cui una persona tratta la rappresentazione delle cose rivelasse veramente che tipo di persona è e fosse il segno del suo atteggiamento nei confronti delle cose rappresentate, si dovrebbe poter affermare che una persona che tratta bene un robot ha un carattere virtuoso. Questa conclusione, però, – come ammette lo stesso Sparrow – appare davvero contro-intuitiva. Non è facile, infatti, accettare che una persona sia virtuosa soltanto perché tratta bene un robot, in quanto una persona virtuosa si prende cura del benessere delle altre persone, non delle macchine: «Immagina qualcuno che è sempre gentile con il proprio cane robot, che si ricorda di accarezzarlo ogni giorno e lo porta a fare ‘passeggiate’, che lo coccola, gli dà i dolcetti e gli parla con voce gentile. (…) non direi che tale comportamento sia moralmente ammirevole o ‘virtuoso’» (Sparrow 2020, p. 4, trad. dell’autore). Secondo Sparrow, comunque, si può ancora sostenere che la gentilezza nei confronti di un robot non è gentilezza ‘autentica’ (e che pertanto non è indicativa di un carattere virtuoso), senza con questo dover rinunciare all’idea che il maltrattamento di un robot sia un vizio reale. È sufficiente riconoscere che quando consideriamo la nostra relazione con i robot c’è un’asimmetria: sulla base di come trattiamo i robot, possiamo infatti rivelare i nostri vizi (o un carattere vizioso), ma non possiamo mostrare le nostre virtù (di avere, cioè, un carattere davvero virtuoso). L’asimmetria – dice Sparrow – si può spiegare con il fatto che, quando non ci sono conseguenze, siamo più veloci a condannare il vizio di una persona che ad apprezzare le sue intenzioni. (L’incapacità di una persona di portare a termine i suoi piani malvagi non la rende meno disprezzabile; non apprezziamo, invece, chi ha buone intenzioni che però non producono mai risultati.) Mentre, poi, per essere viziosi è sufficiente esprimere le proprie emozioni, disposizioni e fantasie; per comportarsi bene, invece, è necessario possedere anche il giudizio o la saggezza pratica (APPRENDIMENTO AUTOMATICO pp…). La virtù, cioè, non richiede soltanto pensieri, fantasie ed emozioni moralmente apprezzabili, ma anche la capacità di impiegare queste disposizioni in un modo appropriato e produttivo. Per questa ragione, – dice Sparrow – una persona che tratta bene un robot non può essere virtuosa: se fosse virtuosa, infatti, saprebbe che non serve a niente prendersi cura di esso (Sparrow 2020).

Sparrow, comunque, non tiene conto che il comportamento di chi stupra un robot può suscitare disgusto, ma non possiamo affermare che stia facendo qualcosa di male perché non danneggia nessuno. Inoltre, non è dal suo comportamento nei confronti degli oggetti inanimati (o delle macchine) che noi conosciamo il suo carattere, ma dalle conseguenze delle sue azioni sulle altre persone. Se è vero, cioè, che chi stupra un robot è una persona crudele, allora dovrebbe esserci una vittima. Qui, invece, c’è soltanto un sex robot e qualsiasi cosa facciamo non possiamo comunque danneggiarlo come vittima: lo possiamo maltrattare e, se vogliamo, anche distruggere, ma non possiamo fargli male (naturalmente le cose sarebbero diverse se un robot fosse cosciente e capace di provare sentimenti: forse un giorno avremo anche macchine di questo tipo, ma per ora si tratta di fantascienza). Inoltre, non è dimostrato che il modo in cui trattiamo la rappresentazione delle cose è segno (e di conseguenza permette di inferire qual è) il nostro comportamento nei confronti delle cose stesse. A livello empirico – come ammette lo stesso Sparrow – manca una prova di questo tipo: non c’è, cioè, alcuna evidenza scientifica che l’esposizione alla violenza sia indicativa di un carattere violento. Il fatto, poi, che una persona provi piacere nel fare sesso con un robot che dice ‘no’ non è necessariamente il segno di assenza di qualsiasi scrupolo morale né dimostra che abbia fantasie di violenza, in quanto non ci dice ancora nulla di cosa per lei sia eccitante e faccia scattare la molla del piacere. Ad esempio, potrebbe identificarsi con la vittima e trarre piacere al pensiero di essere dominato: l’identificazione, infatti, è sempre imprevedibile e non dipende dalla realtà concreta rappresentata. Inoltre, anche se la scena è violenta (il sex robot che viene stuprato), la persona può avere in mente scene di sadomasochismo consensuale e cooperativo con il proprio partner ispirate a valori di cura e amore. Infine, una persona potrebbe trarre piacere nel fare violenza ad un robot proprio perché sa che è soltanto un gioco e che lo spazio dell’immaginazione ammette cose che la realtà, invece, non permette. Del resto, non soltanto parlare di stupro su un robot non ha senso – il sesso con un robot è autoerotismo, in quanto il sesso è qualcosa che avviene tra persone o tra animali o tra animali e persone – ma godere di una violenza rappresentata non è la stessa cosa che godere di una violenza reale (e nemmeno presuppone necessariamente il desiderio di mettere in atto questa violenza su una persona), in quanto gli “oggetti” della violenza, qualsiasi cosa rappresentino, sono fittizi o immaginari.

1. I sex robot promuovono la violenza sulle donne?

Anche quando facciamo sesso possiamo lasciarci andare a fantasie poco convenzionali: non è necessario però che queste fantasie poi si traducano in realtà, in quanto la cosa bella di fantasticare è proprio quella di non doversi confrontare con la realtà e con le sue possibili conseguenze (Soble, 2007). Se è vero che durante un rapporto sessuale possiamo lasciarci andare alla fantasia e immaginare situazioni che restano poi confinate nella nostra ‘camera da letto’, perché dovremmo pensare che acquistare e fare sesso con un sex robot predisponga le persone alla violenza ed allo stupro, ovvero che ciò sia la prova più evidente e incontrovertibile che prima o poi lo faranno (Sparrow, 2017)? Almeno per il momento, i sex robot sono prodotti soprattutto con le fattezze di donna: di conseguenza, essi – è stato affermato – incoraggerebbero l’immagine delle donne come oggetto e alimenterebbero anche l’idea che si può fare sesso con loro anche senza chiedere il consenso. Chi sostiene inoltre che la violenza sui sex robot condurrà alla violenza sulle donne, afferma che c’è da aspettarsi che a lungo andare la violenza sui robot abbasserà le resistenze delle persone a praticare violenza sulle donne, in quanto renderà la violenza sempre più eccitante (Sparrow, 2017). Il fatto, poi, che i sex robot non oppongano resistenza e non reagiscano potrebbe desensibilizzare le persone verso le azioni che producono un danno alle persone reali (Maras, Shapiro, 2017). Si potrebbe perdere la capacità di associare le azioni violente alle sofferenze che esse producono e, progressivamente, perdere il rispetto per le altre persone (Danaher, 2017; Strikwerda, 2017).

Chi ha studiato le conseguenze dell’esposizione prolungata alla violenza e se essa possa, a lungo andare, corrompere il carattere ha rivolto l’attenzione soprattutto ai videogiochi (si veda capitolo 12) e alla pornografia. La cosa non deve sorprendere, in quanto i videogiochi – in particolare quelli violenti – incoraggiano e premiano il giocatore che si mostra abile a compiere azioni che sarebbero considerate immorali o anche dei crimini nel mondo reale come ad esempio rubare, torturare, stuprare ed uccidere (Luck, 2009). La pornografia, poi, indugia spesso su rappresentazioni di violenza su persone non consenzienti. Tuttavia, la convinzione che l’esposizione alla violenza o alla rappresentazione della violenza renda più facile compiere attività violente è fortemente contestata nella letteratura scientifica. Non abbiamo, cioè, alcuna evidenza empirica che le esperienze di violenza che viviamo nello spazio della fantasia (ad esempio, attraverso la pornografia, la letteratura, il cinema, i videogiochi o qualsiasi altro mezza) incidano sul nostro carattere ed influenzino il nostro comportamento: «La relazione tra film e violenza televisiva e violenza reale, e la relazione tra violenza dei videogame e violenza reale, è stata dibattuta sempre da quando certe tecnologie sono diventate di ampio consumo. Allo stesso modo, si è discusso a lungo sulla relazione tra pornografia e violenza sessuale e/o sessismo. Anche se da decenni abbiamo argomenti a favore di una tesi e dell’altra e migliaia di studi accademici, il consenso su queste questioni rimane elusivo» (Sparrow, 2017, p. 470).

Spesso gli atti di delinquenza, risse a scuola e qualsiasi tipo di comportamento criminale violento (ad esempio, aggressioni, rapine ed omicidi) vengono spiegati con l’esposizione ‘frequente’ ai videogiochi. Alcuni studi, poi, sottolineano una relazione stretta tra consumo di materiale pornografico e la violenza e/o una maggiore tolleranza verso le aggressioni sessuali nei confronti delle donne. Tuttavia, non c’è alcuna evidenza scientifica che trarre piacere da rappresentazioni di violenza renda gli individui più predisposti alla violenza nel mondo reale (Sparrow, 2017; Balistreri, 2018). Chi lo sostiene non considera che la violenza non è per forza disadattiva o socialmente scoraggiata (Zagal, 2012) e che essa potrebbe essere il risultato di altre influenze negative (Ferguson, 2012). Secondo alcune ricerche, infine, la pornografia potrebbe avere effetti anche positivi sul carattere e aiutare a scaricare le fantasie più ‘violente’ e renderle, comunque, più controllabili e meno pericolose. Nello studio di Riegel (2004), ad esempio, su persone che avevano posseduto materiale pedopornografico, la maggior parte dei partecipanti ha riferito che l’immagine ha agito come ‘sostituto’ del bambino reale senza aumentare il desiderio di commettere crimini sessuali (Young, 2016). Alcuni autori di crimini sessuali hanno affermato che il consumo di materiale pedopornografico prodotto interamente attraverso il computer (perciò, senza immagini di persone reali) ha effetti positivi e fa sentire loro meno il desiderio di commettere nuove violenze (Strikwerda, 2017). Comunque, anche se non crediamo che la pornografia possa avere un effetto catartico, essa potrebbe avere un effetto liberatorio e contribuire a portare alla luce un’ampia gamma di abitudini sessuali. Attraverso la pornografia, cioè, forme di sessualità in precedenza segregate e represse possono vedersi pienamente riconosciute nel loro diritto al piacere e nella loro identità (Soble, 2007).

Se partiamo da quello che sappiamo riguardo all’esposizione alla violenza o alla rappresentazione della violenza, la paura che i sex robot favoriscano la violenza sulle donne non sembra molto giustificata. Si tratta comunque di una preoccupazione che non dovremmo sottovalutare e per questo, prima di trarre conclusioni definitive sui sex robot, è necessario studiare il fenomeno con più attenzione. Del resto, le fantasie sessuali potrebbero facilitare o agire da stimolo per l’atto abusante o rappresentare un fattore disinibente che, oltre a desensibilizzare «rispetto al comportamento sessualmente deviante», riduce la capacità del soggetto di mettere in atto meccanismi di controllo (Zara, 2018, p. 256, trad. dell’autore). Non soltanto il rapporto con un robot potrebbe essere più intimo e coinvolgente di qualsiasi relazione con personaggi di finzione, immaginabile attraverso la pornografia o i videogiochi. Ma in un rapporto con un sex robot può essere più difficile prendere le distanze da se stessi: chi gioca ad un videogioco, invece, può sempre agire come se fosse un’altra persona, in quanto compie le proprie azioni attraverso un avatar ed in un mondo non reale ma virtuale (Danaher, 2017). Inoltre, dovremmo considerare anche gli effetti a lungo termine della loro commercializzazione, in quanto sex robot sempre più interattivi e somiglianti agli esseri umani ma ‘schiavi’ potrebbero agire da rinforzo e da stimolo per le fantasie più violente nei confronti dei nostri simili. Più, cioè, le rappresentazioni dei sex robot saranno realistiche, maggiori potrebbero essere le probabilità che incoraggino le persone a mettere in atto nella realtà i comportamenti fantasticati (Sparrow, 2017). Tuttavia, come ammette lo stesso Sparrow, le rappresentazioni più realistiche potrebbero incoraggiare alla violenza meno rispetto a rappresentazioni più semplici e schematiche. Inoltre, non abbiamo evidenze che un’interazione o simulazione immersiva e coinvolgente con il sex robot faccia perdere al soggetto coinvolto la capacità di controllare i propri comportamenti o che lo predisponga a trasportare in modo automatico la sua esperienza in qualsiasi relazione. Il fatto che – attraverso la realtà virtuale immersiva – una persona possa praticare (e sviluppare) certe capacità che, in seguito, può impiegare con successo e cognizione di causa nella realtà non la priva della libertà di decidere se ricorrere ad esse e in quale contesto o occasione utilizzarle. Nemmeno, poi, l’ipotesi dell’uso terapeutico dei sex robot per la gestione di patologie sessuali presuppone che la simulazione con il robot possa facilitare il transfer a situazioni non simulate, in quanto l’obiettivo non è quello di incoraggiare la violenza sessuale su donne e bambini, ma di offrire ai criminali sessuali un ‘oggetto sostitutivo’ che possa aiutarli a contenere le loro perversioni. Non è necessario, poi, che l’oggetto sostitutivo assicuri la stessa gratificazione dell’oggetto reale, ma è sufficiente che permetta al soggetto di avere un maggiore controllo sulle sue pulsioni.

1. Conclusioni

Accade spesso che i sex robot vengano presentati come un prodotto il cui uso sarebbe sempre immorale, in quanto – questa è la critica – ad essi potrebbero ricorrere soltanto uomini soli e viziosi. Noi abbiamo mostrato che le ragioni per acquistare o fare sesso con un sex robot possono essere diverse, e che dietro questa scelta potrebbe esserci una preoccupazione sincera, legittima ed apprezzabile per il piacere di altre persone, per la stabilità della propria relazione e per il proprio benessere. Si può, poi, anche ammettere che per alcune persone i sex robot potrebbero essere una scelta obbligata, perché hanno difficoltà a socializzare con altre persone o a portare avanti un rapporto. Tuttavia, una persona che acquista un sex robot per addolcire la sua solitudine non è criticabile, in quanto non ha un carattere spregevole, ma un problema personale che deve ancora risolvere. Anche se, comunque, uomini e donne potrebbero avere beneficio dalla compagnia di un sex robot, il ricorso a questi strumenti potrebbe intensificare la loro solitudine (Forsch-Villaronga, 2020). Avere, inoltre, un rapporto sessuale con un robot non è la stessa cosa che stuprare una donna, in quanto nel primo caso l’oggetto verso il quale l’eccitazione è rivolta è semplicemente una macchina. Un sex robot, poi, non ha una volontà e, di conseguenza, non può essere veramente stuprato: pertanto, l’idea che una persona possa comprare un robot per stuprarlo può suscitare disgusto, ma questo sentimento di ripugnanza non sembra giustificato moralmente, in quanto non fa male a nessuno. È sbagliato, cioè, valutare un contesto di finzione come se appartenesse al mondo della realtà, in quanto lo spazio della finzione (o della fantasia) ammette cose che, invece, la realtà non permette. Infine, non c’è alcuna evidenza scientifica che i sex robot promuovano la violenza sulle donne, in quanto le esperienze di violenza che viviamo nello spazio della fantasia o dell’immaginazione (ad esempio, attraverso la pornografia, la letteratura, il cinema, i videogiochi o qualsiasi altro mezzo) non sembrano incidere sul nostro carattere ed influenzare il nostro comportamento. Tuttavia, è ancora presto per esprimere una valutazione morale conclusiva sui sex robot: le preoccupazioni relative alle possibili conseguenze della loro produzione sono legittime e devono essere affrontate approfondendo di più il fenomeno e raccogliendo evidenze scientifiche adeguate. Avere, poi, un quadro più chiaro delle implicazioni benefiche e dannose in gioco è importante anche per avviare una riflessione sulla responsabilità delle persone che producono sex robot. Noi possiamo avere rapporti sessuali con i robot, e – come abbiamo visto – volendo anche stuprarli, soltanto perché i produttori dei robot fanno certe scelte e programmano i robot un certo modo. Riuscire a dare, pertanto, una risposta alle questioni morali che precedentemente abbiamo discusso potrebbe aiutarci a stabilire se la produzione dei sex robot è moralmente accettabile o da incoraggiare. Serve, comunque, anche una riflessione morale più ampia per considerare se la produzione di sex robot sempre più realistici promuove un’immagine pericolosa del piacere e della sessualità – il corpo, soprattutto quelle donne, come oggetto da stuprare o comunque sempre disponibile – o se invece favoriscono trasformazioni culturali che incoraggiano l’esplorazione del piacere.

Capitolo 11

Le implicazioni etiche degli usi educativi e didattici dei robot

Edoardo Datteri, Luisa Zecca

1. Dalle “macchine per insegnare” agli insegnanti robotici

Nel 1924, lo psicologo statunitense Sidney Pressey presentò al convegno dell’*American Psychological Association* una versione della sua “macchina per insegnare” (*teaching machine*). Consisteva in un rullo che, come nelle macchine da scrivere, faceva avanzare un foglio di carta su cui erano scritte delle domande a risposta multipla. L’utente aveva a disposizione quattro pulsanti per scegliere una delle risposte. Nella modalità “test” il sistema memorizzava il numero di risposte corrette mentre avanzava con le domande; nella modalità “teach” il sistema ripresentava la medesima domanda finché l’utente non sceglieva la risposta corretta. Raggiunto un certo numero di risposte corrette, la macchina erogava una caramella. Per motivi analizzati da Benjamin (1988), Pressey non riuscì a creare un mercato attorno al suo dispositivo il cui ricordo progressivamente svanì fino al suo brusco recupero, agli inizi degli anni ’50, da parte di Burrhus Skinner. La macchina di Skinner presentava caratteristiche diverse da quella di Pressey: l’utente era chiamato a comporre la risposta anziché sceglierla da un insieme predefinito di opzioni, e i problemi venivano presentati secondo una sequenza programmata in modo funzionale al progressivo raggiungimento di certi obiettivi di apprendimento (Skinner, 1958). Skinner salutava con favore l’uso delle macchine per insegnare: esse, a suo avviso, permettevano di personalizzare il processo di apprendimento, di renderlo più rapido (poiché l’utente veniva subito messo a conoscenza dei propri errori) e di raggiungere molti più studenti a parità di numero di insegnanti. Non tutti erano d’accordo: tra i contemporanei di Skinner vi era chi dubitava seriamente della possibilità di meccanizzare l’apprendimento, temendo che la massiccia introduzione delle “macchine per insegnare” nel mercato avrebbe potuto causare effetti de-umanizzanti e perdita di posti di lavoro (Benjamin, 1988).

La ricerca sulle “macchine per insegnare” è proseguita lungo il corso dei decenni successivi, producendo sistemi informatici e robotici per facilitare l’insegnamento e l’apprendimento in contesti scolastici e non scolastici. Questo capitolo si concentrerà sulle applicazioni educative e didattiche, attuali e futuribili, dei robot. Forniremo anzitutto una panoramica dei principali scenari d’uso. Presenteremo poi alcune questioni etiche e sociali sollevate da tali scenari, che riguardano la protezione dei dati personali, l’attribuzione di cognizione ed emozioni ai robot, la possibilità di meccanizzare l’insegnamento. Lungo tutto il testo utilizzeremo, se inevitabile, il genere linguistico maschile interpretandolo come neutro: il termine “bambini”, per esempio, dovrà essere inteso come non necessariamente riferentesi a persone di genere maschile.

1. Scenari di uso dei robot in contesti educativi e didattici
	1. Lo scenario 1: la robotica educativa

La discussione sulle questioni etiche sollevate dagli usi educativi e didattici dei robot sarà articolata attorno a tre scenari che corrispondono rispettivamente alla cosiddetta robotica educativa (*educational robotics*), all’uso di robot come insegnanti ed educatori, all’uso di robot come pari nei processi di apprendimento. Questa tripartizione corrisponde a quella proposta da Mubin e colleghi (2013) in una rassegna della letteratura sul tema. I tre scenari possono essere distinti tra di loro lungo le dimensioni (a) della *sostituzione* – ovvero, del fatto che il robot sostituisca o meno un essere umano nel processo di insegnamento-apprendimento; (b) del *ruolo* assegnato al robot nel processo didattico o educativo; (c) della possibilità, da parte degli esseri umani che interagiscono con il robot, di *manipolare* o programmare alcuni aspetti del funzionamento del robot; (d) del livello di *socialità* del robot (ROBOTICA SOCIALE pp…). Iniziamo dal primo degli scenari considerati.

Gli studenti di una classe seconda di scuola primaria seguono un percorso di apprendimento di alcune ore, distribuite in vari giorni, con la continua presenza di un insegnante. In una prima parte del percorso gli studenti lavorano alla costruzione di alcuni robot non umanoidi e non sociali a partire da scatole di montaggio che mettono a disposizione sensori (di prossimità, audio, tattili, visuali) e attuatori, seguendo istruzioni di montaggio fornite dal costruttore. Nella seconda parte del percorso, gli studenti imparano a programmare il robot utilizzando un linguaggio di programmazione imperativo e un’interfaccia di programmazione visuale che “gira” su un calcolatore portatile. Infine, l’insegnante propone un gioco agli studenti: il robot deve essere programmato in modo da muoversi lungo una certa traiettoria tracciata sul pavimento dell’aula. Gli studenti, organizzati in gruppi, elaborano i loro programmi e li mettono alla prova. L’insegnante aiuta gli studenti a elaborare gli errori commessi e ad affrontare i problemi di programmazione che incontrano, evitando di fornire quella che ritiene essere la risposta “giusta”.

Scenari di questo tipo sono generalmente chiamati di robotica educativa (anche se l’espressione assume contorni sfumati: si veda la discussione in Angel-Fernandez, Vincze, 2018). Si caratterizzano, anzitutto, per il fatto che il robot non *sostituisce* alcun essere umano (è la prima delle dimensioni che abbiamo introdotto nel paragrafo precedente). Svolge il *ruolo* di mediatore didattico tra insegnanti e studenti. Le attività educative e didattiche consistono proprio nel *manipolare* la struttura fisica e algoritmica del sistema (ovvero, nel programmarne il comportamento attraverso un linguaggio e un’interfaccia di programmazione) perché esso svolga certi compiti o manifesti certi schemi di comportamento. Il robot viene presentato agli studenti sotto forma di scatola di montaggio e dunque non in veste di robot *sociale*. L’obiettivo dell’attività di programmazione non è quasi mai quello di rendere sociale il robot, ma di programmarlo perché segua certe traiettorie nell’ambiente o manifesti certe reazioni non verbali agli stimoli esterni.

La robotica educativa si basa sul presupposto – oggetto di analisi empirica da parte della comunità di ricerca internazionale – secondo cui le attività di costruzione e di programmazione dei robot, se opportunamente progettate e condotte, possono facilitare l’acquisizione e lo sviluppo di competenze disciplinari (connesse per esempio alle scienze, alla matematica, alla tecnologia) e trasversali rispetto alle singole discipline (per esempio di carattere cognitivo e socio-relazionale, Papert, 1980). I fattori che possono modulare l’apprendimento degli studenti in queste attività sono molti, non ultimo il tipo di guida fornita dagli adulti presenti (Atmatzidou *et al.*, 2018), e lo scenario descritto può variare sotto molti aspetti (che includono il tipo di robot usato, le caratteristiche del linguaggio e dell’interfaccia di programmazione, l’obiettivo di apprendimento, l’età e gli stili cognitivi degli studenti, le loro modalità di apprendimento, la loro motivazione, la loro percezione di autoefficacia, il contesto socio-demografico, lo sfondo culturale degli studenti e degli adulti coinvolti, il ruolo assunto dall’insegnante). Per informazioni aggiornate sulla robotica educativa si vedano (Anwar *et al.*, 2019; Benitti, 2012; Bozzi *et al.*, 2021; Mubin *et al.*, 2013).

* 1. Lo scenario 2: l’insegnante robotico

Lo scenario 1 è attuale: sono molte le scuole e i contesti educativi in cui i robot vengono utilizzati come mediatori didattici nel senso appena descritto. Lo scenario 2 è invece futuribile: non fa parte ancora della quotidianità dell’insegnamento, ma è stato oggetto di vari studi sperimentali.

Un robot umanoide di fattezze femminili, dotato di un viso in grado di esprimere emozioni e di arti superiori, ma non in grado di spostarsi nell’ambiente, è posto in un’aula di fronte a studenti di scuola primaria. L’insegnante è assente. Il robot legge ad alta voce un testo sul principio della leva; durante questa fase non è in grado di interagire in alcun modo con gli studenti. Terminata la lettura, il robot entra in una modalità interattiva in cui mantiene il contatto oculare con i bambini. Quando i bambini si avvicinano, li riconosce e li saluta chiamandoli per nome; rivolge loro domande sul principio della leva, aspetta le risposte, ed esclama “Bravo!” oppure “Sbagliato!” se la risposta è giusta o sbagliata. Se i bambini fanno troppa confusione, produce un’espressione arrabbiata ed esclama: “Fate silenzio!”. Il comportamento del robot è parzialmente teleoperato: un operatore umano controlla alcuni aspetti del comportamento del robot da una stanza vicina, osservando la classe attraverso le telecamere presenti negli occhi del robot. Attraverso dei “click” sul monitor, l’operatore può innescare dei comportamenti standard (gesti di saluto, frasi di circostanza) oppure attivare o disattivare la modalità autonoma di dialogo con gli studenti.

Lo Scenario 2 è simile a quello descritto in (Hashimoto *et al.*, 2011) di cui è protagonista il robot umanoide SAYA. Si differenzia molto dallo scenario 1 lungo le dimensioni che abbiamo introdotto sopra. Anzitutto, contrariamente agli scenari di robotica educativa, il robot è pensato per *sostituire* un insegnante umano o un esperto su un certo tema. Può essere progettato per operare in assenza dell’adulto, come nel caso di SAYA, oppure per lavorare “in compresenza” con l’insegnante umano sostituendo un esperto (Park *et al.*, 2011). Contrariamente alla robotica educativa, il robot è pensato per svolgere in questi casi alcuni *ruoli* svolti dagli insegnanti, da esperti su temi specifici, oppure da educatori o animatori (anche nell’interazione con bambini molto piccoli, come negli scenari discussi da Sharkey, Sharkey, 2010). Il robot non è pensato per essere *manipolabile* dai bambini, ma si presenta già programmato per esibire certi comportamenti o manifestare certe forme di interazione. Generalmente esibisce elevati livelli di *socialità* (contatto oculare, capacità di rispondere selettivamente a persone diverse) e possiede abilità di produzione linguistica e di espressione emotiva. Lo scenario 2 può presentarsi in alcune varianti. Il robot può essere totalmente teleoperato da un essere umano (Yun *et al.*, 2011), oppure possedere gradi di autonomia, come SAYA. Può proporre attività di vario genere, dalle lezioni frontali (come SAYA) alla presentazione di giochi educativi (Movellan *et al.*, 2009), all’interazione linguistica con gli studenti su particolari temi. Può essere impiegato per assistere nell’apprendimento di varie discipline. In letteratura si prospetta l’uso di insegnanti robotici per insegnare inglese: è il cosiddetto RALL (*Robot Assisted Language Learning*, Chang *et al.*, 2010) motivato, in Corea, dalla difficoltà di trovare esperti madrelingua inglesi che vogliano recarsi nelle zone rurali del paese.

Belpaeme e colleghi (2018), Li (2015) e Baxter e colleghi (2017) presentano alcuni studi empirici sugli effetti derivanti dall’uso di insegnanti robotici sull’apprendimento e sulle caratteristiche che tali sistemi dovrebbero possedere per svolgere efficacemente il loro ruolo (torneremo sul tema nel paragrafo 2.4). Belpaeme e colleghi (2018) affermano inoltre che l’introduzione degli insegnanti robotici nei contesti educativi può essere salutata con favore perché – nella loro opinione – essa permette un risparmio economico e maggiori gradi di personalizzazione dei percorsi educativi e didattici, fattori entrambi importanti nel contesto di una generale tendenza all’aumento del numero di studenti per classe.

* 1. Lo scenario 3: “pari” nei processi di apprendimento

Nello scenario 3 (anch’esso non attuale ma futuribile nel senso che abbiamo precisato) al robot viene assegnato il compito di “compagno” durante l’apprendimento.

In una scuola elementare giapponese, due robot umanoidi sono posti ai lati di un corridoio in un’area frequentata dai bambini. I robot possiedono arti superiori e un viso in grado di manifestare semplici espressioni emotive. Sono altamente sensorizzati e programmati per interagire autonomamente, anche verbalmente, con i bambini. Possono distinguere gli interlocutori tra di loro sulla base di dispositivi di identificazione detti RFID (“*radio frequency ID*”) e reagire in modo personalizzato a seconda del bambino con cui interagiscono. Normalmente stanno fermi nella loro collocazione finché un bambino non si avvicina: calcolando la distanza tra di essi e il bambino, classificano quest’ultimo come un partecipante o come un semplice spettatore. A quel punto, iniziano a interagire con lui chiamandolo per nome e invitandolo ad abbracciarlo; dopodiché inizia un semplice dialogo in lingua inglese.

Nello scenario 3, il robot *sostituisce* un essere umano. Viene utilizzato in ambienti in cui sono presenti adulti, ma l’adulto non partecipa direttamente all’interazione: i bambini vengono lasciati da soli in compagnia del robot. Contrariamente allo scenario 2, il *ruolo* assegnato al robot è quello di compagno o pari (per esempio, di un amico madrelingua inglese). Come nello scenario 2, il sistema non è *manipolabile* dagli utenti ed esibisce elevati livelli di *socialità*. Un esempio è il sistema ROBOVIE descritto da Kanda *et al.* (2004). Altri esempi sono descritti in (Mubin *et al.*, 2013; Tanaka *et al.*, 2007; Han *et al.*, 2005). Il robot può svolgere il ruolo di “pari” in vari sensi. Per esempio, può svolgere il ruolo di un compagno novizio che chiede aiuto ai bambini per l’apprendimento dell’inglese (Tanaka, Matsuzoe, 2012) o della scrittura a mano (Hood *et al.*, 2015). In questi scenari, il bambino “insegna” al robot come svolgere un certo compito secondo un approccio chiamato dagli autori *learning by teaching* (Tanaka, Kimura, 2009).

1. Privacy, inganno, responsabilità
	1. Sicurezza e protezione dei dati personali

Dopo avere presentato i principali scenari d’uso educativo e didattico, attuali e futuribili, dei robot, iniziamo ad affrontare alcune delle questioni di carattere etico e sociale che tali scenari sollevano, prendendo le mosse dal tema della tutela dei dati personali. I robot coinvolti in questi scenari possono essere deliberatamente utilizzati per memorizzare e trasmettere a distanza volti, informazioni e parole dei bambini? Questa domanda solleva preoccupazioni etiche e legali soprattutto qualora la trasmissione di dati consenta di risalire all’identità dei bambini coinvolti, violi il diritto alla protezione della vita privata dei bambini sancito dalla Convenzione ONU sui diritti dell’infanzia e dell’adolescenza, possa determinare minacce alla loro integrità fisica o psicologica. Per valutare la fondatezza di questi timori caso per caso può essere utile chiedersi se (a) il sistema in esame possa effettivamente memorizzare o trasmettere a distanza informazioni acquisite dai sensori; se (b) questa possibilità sia controllabile o meno dagli adulti; se (c) la memorizzazione o la trasmissione dei dati sensoriali sia necessaria o meno rispetto alla destinazione d’uso del sistema (in caso contrario, si potrebbe ritenere opportuno disattivarla).

In merito a queste domande, il primo scenario – l’uso di robot come oggetti manipolabili – non solleva preoccupazioni di particolare rilievo. Non sono molti i kit robotici che hanno la possibilità di memorizzare o trasmettere informazioni sui bambini (a). I robot utilizzati in questi contesti sono programmabili dai bambini e dagli adulti, che dunque detengono ampi margini di controllo sulla trasmissione a distanza di eventuali dati memorizzati (b). Se la memorizzazione dei dati sensoriali sia necessaria o meno rispetto alla destinazione d’uso del sistema, è una questione da valutarsi tenendo conto delle abilità e competenze che l’insegnante desidera stimolare attraverso le abilità di programmazione. Se il robot è utilizzato come piattaforma per imparare a programmare algoritmi di *streaming* di dati sul web, la disattivazione delle funzionalità che stiamo esaminando è chiaramente in conflitto con l’uso didattico del robot. Generalmente, tuttavia, nel primo scenario i robot vengono utilizzati per stimolare certe capacità di ragionamento legate alla programmazione. È ragionevole sostenere che i moduli che permettono la memorizzazione e la trasmissione dei dati personali non siano davvero essenziali a questo scopo, e che sia possibile disattivarli – qualora presenti – senza compromettere l’utilità del sistema come mediatore didattico.

La situazione è più complessa negli scenari 2 e 3. I sistemi di robotica sociale impiegati in questi due scenari sono generalmente dotati di sensori in grado di acquisire, e potenzialmente memorizzare e trasmettere, informazioni personali (a); si tratta di sistemi generalmente “chiusi”, sui cui meccanismi di funzionamento adulti e bambini non esercitano un controllo (b). Infine, la possibilità di memorizzare e trasmettere a distanza informazioni personali è essenziale perché questi sistemi svolgano nel modo dovuto il ruolo di “insegnante robotico”. Il ruolo dell’insegnante richiede la capacità di analizzare, memorizzare, interpretare il *feedback* degli studenti (torneremo su questo punto nel paragrafo 2.4). Questa capacità è essenziale per la modulazione dell’intervento comunicativo e, dunque, per la conduzione dell’attività didattica. ROBOVIE (Kanda *et al.*, 2007) personalizza l’interazione con i bambini attraverso la loro identificazione tramite dispositivi RFID. Il robot KindSAR conserva dati anonimizzati sui processi comunicativi e di ragionamento dei bambini (Fridin, 2014) in modo da poter condurre valutazioni in itinere degli effetti dell’interazione tra i bambini e il robot. Queste forme di raccolta di informazioni sul comportamento e sulle attività cognitive dei bambini sono chiaramente insufficienti a garantire il verificarsi di una relazione significativa e didatticamente fruttuosa tra l’insegnante robotico e i bambini; lo sviluppo di “buoni” insegnanti robotici, come torneremo a discutere più avanti, presuppone lo sviluppo di algoritmi che permettano al robot non soltanto di acquisire dati sensoriali “grezzi” sul comportamento dei bambini ma anche di interpretarli e usarli per inferire ipotesi sui loro stati mentali ed emotivi.

Ricapitolando, le preoccupazioni sulla tutela della *privacy* dei bambini possono generalmente essere attenuate disattivando qualsiasi meccanismo di raccolta e trasmissione di informazioni sensoriali negli scenari di tipo 1. Lo stesso tipo di soluzione non può essere adottata negli scenari 2 e 3 in cui l’insegnante robotico, per svolgere bene il compito a esso delegato, *deve* raccogliere informazioni dettagliate, strutturate e interpretate sui bambini con cui interagisce.

* 1. L’inganno della robotica sociale: animismo infantile e teoria della mente

Nel capitolo 9 sono state discusse preoccupazioni di natura etica legate al presunto inganno derivante dall’attribuzione di mente ed emozioni ai robot sociali. Preoccupazioni simili possono sorgere anche in relazione a scenari di interazione tra robot sociali e bambini come quelli di tipo 2 e 3 (Sharkey, 2016). Per valutarne la fondatezza può essere utile porsi le seguenti domande.

1. È vero che, negli scenari di robotica sociale di tipo 2 e 3, i bambini compiono attribuzioni ingannevoli di mente ed emozioni ai robot?
2. Questo inganno, se sussiste, è di natura diversa rispetto a ciò che accade quando i bambini interagiscono con animali o giocattoli?
3. Tale inganno può portare a conseguenze negative nel processo di insegnamento-apprendimento?
4. È un inganno evitabile, e come?

In questa sezione discuteremo la presunta ingannevole attribuzione di mente ai robot sociali da parte dei bambini, riservando la sezione successiva al tema delle emozioni. La prima delle quattro domande può essere a sua volta articolata nelle seguenti: (a) è vero che i bambini attribuiscono stati mentali ai robot con cui interagiscono? In tal caso, (b) tali attribuzioni sono ingannevoli o meno, e in che senso? Entrambe le domande chiamano in causa questioni empiriche e filosofiche di non facile soluzione.

Che i bambini tendano ad attribuire stati mentali ai robot con cui interagiscono è un’ipotesi empirica implicata dagli studi di Piaget sull’evoluzione delle strutture di conoscenza nei bambini. Ne *La représentation du monde chez l'enfant* (1926), Piaget ipotizza che il bambino – dalla fine del primo anno di vita al termine del secondo – non distingua il mondo psichico dal mondo esterno: per questo «ci si deve aspettare che consideri come vivi e coscienti un gran numero di corpi che per noi sono inerti» (ivi, trad. it. p. 175). Durante i primi anni di vita il bambino arriva progressivamente ad attribuire stati mentali soltanto agli animali e agli esseri umani, passando attraverso fasi in cui attribuisce una mente soltanto agli oggetti (per esempio, agli oggetti meccanici) dotati di «moto proprio» e non a tutto ciò il cui moto è causato da «fattori esterni» (ivi, trad. it. p. 187). L’attribuzione di stati mentali a oggetti meccanici da parte dei bambini può essere compresa anche alla luce delle ricerche sulla cosiddetta teoria della mente (Wellman, Liu, 2004; Leslie, 1987; Harris, 1994; Lillard, 1993): si vedano Zhang *et al.* 2019 e Pucciarelli *et al.*, 2020 per alcuni studi empirici sull’attribuzione di intenzionalità e false credenze ai robot da parte di bambini a sviluppo normotipico e nello spettro autistico. Queste considerazioni ci spingono a riflettere sulla possibilità che l’attribuzione di vita e stati mentali ai robot dipenda “intrinsecamente” dalla struttura dei meccanismi cognitivi e conoscitivi dei bambini e non dalle azioni pubblicitarie di chi costruisce, commercializza o utilizza i robot. Attenzione, tuttavia, all’impegno ontologico che accompagna queste attribuzioni. Una cosa è affermare che i bambini si comportano *come se* attribuissero vita e stati mentali al robot; un’altra è affermare che essi *ritengono* che i robot possiedano effettivamente vita e stati mentali. Nella cornice piagetiana, l’animismo infantile viene definito come un “orientamento spirituale”, uno schema di spiegazione, e non come un’attribuzione consapevole di vita e mente agli oggetti (sarebbe dunque un atteggiamento “come se”). Nei bambini molto piccoli, l’interazione con il robot potrebbe essere intesa come gioco di finzione analogo a quello che essi mettono in atto con altri tipi di oggetti inanimati. Nell’ambito di un gioco di finzione, l’attribuzione di vita e stati mentali non implica la convinzione che i robot siano effettivamente vivi e possiedano stati mentali. La plausibilità della tesi dell’inganno, e la fondatezza delle conseguenti valutazioni etiche, dipende anche dalla plausibilità di un’interpretazione realista o strumentalista dell’atteggiamento intenzionale che i bambini adottano nei confronti dei robot (Bechtel, 1985).

La questione introdotta in precedenza – è vero che i bambini attribuiscono ingannevolmente stati mentali ai robot con cui interagiscono in ambito educativo e didattico? – ci spinge anche a riflettere sul senso in cui tali attribuzioni sarebbero *ingannevoli*. Possono esserlo in almeno due sensi. Sono ingannevoli perché è *falso* che i robot possiedono stati mentali e capacità cognitive: questa è una tesi tutt’altro che ovvia e da giustificare identificando criteri plausibili per decidere se un certo oggetto possieda o meno mente o capacità cognitive. Oppure, sono ingannevoli perché attribuiscono ai robot capacità cognitive *diverse* da quelle che essi effettivamente possiedono: questa è una tesi parzialmente empirica che non ammette necessariamente una risposta univoca.

Le considerazioni fin qui introdotte ci inducono a considerare con cautela la tesi secondo cui l’interazione tra bambini e robot sociali riposa su attribuzioni ingannevoli di capacità cognitive ai robot. E ci spingono anche a considerare con cautela la tesi secondo cui i contesti di robotica sociale si basano su un inganno “più grave”, in qualche senso del termine, di quello che si verifica nell’interazione tra bambini e giocattoli oppure tra bambini e animali (Sharkey, Sharkey, 2010). I robot sociali replicano talvolta con grande accuratezza l’aspetto fisico e i movimenti degli esseri umani e possiedono spesso notevoli capacità di interazione verbale e fisica con i bambini, tra cui la capacità di mantenere il contatto oculare. Se queste caratteristiche inducano o meno i bambini ad attribuire una mente ai robot è questione da valutare empiricamente, tenendo però conto delle riserve che abbiamo appena discusso: l’attribuzione potrebbe far parte di un gioco di finzione e dunque assumere una valenza strumentale; potrebbe essere in gran misura dovuta ai meccanismi di mentalizzazione di cui sono dotati i bambini a sviluppo normotipico, e non imputabile alle intenzioni di chi ha costruito o pubblicizzato il sistema.

L’attribuzione ingannevole di stati mentali può generare preoccupazioni nella prospettiva dell’etica delle conseguenze in vista degli effetti collaterali che essa può produrre nel processo di insegnamento-apprendimento. Quali sono questi effetti collaterali? È la terza delle domande che abbiamo introdotto in precedenza. Secondo Sharkey (2016), sovrastimare le capacità cognitive del robot – una delle forme di inganno che abbiamo introdotto poche righe sopra – può generare aspettative infondate rispetto ai compiti che esso può sensatamente svolgere negli scenari 2 e 3, e dunque deleghe troppo arrischiate da parte degli insegnanti umani o del sistema di istruzione. Per esempio, sovrastimare le capacità di interpretazione del comportamento verbale e non verbale dei bambini può indurre gli adulti a delegare al robot il compito di condurre lezioni dialogate oppure di controllare l’interazione sociale dei bambini (per esempio, sgridandoli quando fanno confusione come nel nostro ipotetico scenario 2). False assunzioni sulle capacità cognitive del robot potrebbero indurre gli studenti a interpretare in modo sbagliato il suo comportamento, con conseguenti ricadute negli apprendimenti. Il robot potrebbe dare risposte inappropriate alle domande dei bambini in virtù dei suoi limiti di elaborazione percettiva; mentre i bambini potrebbero “prendere per buone” le sue risposte sovrastimando le sue capacità cognitive.

Vi sono dunque buone ragioni per temere che l’inganno conseguente dall’attribuzione al robot di capacità cognitive diverse da quelle che esso effettivamente possiede possa produrre effetti collaterali negativi nel processo di insegnamento-apprendimento e sollevare preoccupazioni nella prospettiva dell’etica delle conseguenze. Queste preoccupazioni, da valutare assieme agli eventuali benefici derivanti dall’introduzione di insegnanti e “pari” robotici, possono essere affrontate interrogandosi sulla possibilità di attenuare o evitare tali forme di inganno (è la quarta delle domande sollevate all’inizio della sezione) migliorando la comprensione che gli utenti hanno del funzionamento del robot. Una delle strade percorribili vede lo sviluppo di sistemi in grado di produrre spiegazioni del proprio comportamento (Hagras, 2018). Un’altra strada passa attraverso interventi tesi a migliorare il livello di alfabetizzazione informatica degli umani – bambini, insegnanti, educatori – che interagiscono con i robot degli scenari 2 e 3. Entrambe le strade sollevano sfide teoriche ed empiriche di grande rilievo.

* 1. L’emotività ingannevole

Le quattro domande introdotte nella sezione precedente possono guidarci nella valutazione della tesi secondo cui gli scenari di tipo 2 e 3 coinvolgono ingannevoli attribuzioni di *emotività* ai robot. Le prime due possono essere affrontate in modo simile a quello già seguito, e per questo motivo le scorreremo velocemente. Se l’interazione che si verifica negli scenari 2 e 3 sia accompagnata o meno da ingannevoli attribuzioni di emotività ai robot è una questione empirica e filosofica che richiede di valutare (a) se effettivamente, in quei contesti, i bambini attribuiscano emozioni ai robot, e (b) se tali attribuzioni siano ingannevoli. Il primo punto deve essere affrontato tenendo conto della distinzione tra un’attribuzione di tipo strumentale (il “come se” del gioco di finzione) e la convinzione, ontologicamente più impegnativa, che i robot possiedano effettivamente emozioni. Il secondo punto deve essere affrontato ragionando sui criteri in base ai quali è lecito attribuire emozioni a un ente (utili punti di partenza sono Calvo *et al.*, 2015; Caruana, Viola, 2018). Queste domande possono essere utili anche in relazione al proposito di comprendere se l’interazione con i robot sociali si accompagni ad attribuzioni ingannevoli di emotività “più gravi” rispetto a quelli che i bambini a sviluppo normotipico tendono a compiere nei confronti di animali e giocattoli.

La terza tra le domande introdotte verte sulla possibilità che l’attribuzione di emotività ai robot porti conseguenze negative nel processo di insegnamento-apprendimento, ed emerge nella valutazione degli usi educativi e didattici dei robot dal punto di vista dell’etica delle conseguenze. Le preoccupazioni sollevate da Sharkey (2016) riguardano la possibilità che l’attribuzione di emozioni ai robot faciliti lo sviluppo di legami di attaccamento non sicuri con essi. Tali legami, soprattutto se l’interazione non è opportunamente presidiata dagli adulti (come avviene in alcune varianti degli scenari 2 e 3), potrebbero produrre reazioni emotive di tristezza, rabbia, dolore, angoscia e disorientamento nei bambini soprattutto qualora il robot produca comportamenti non appropriati al contesto. Secondo Sharkey, potrebbero educare a forme impoverite o addirittura distorte di relazione affettiva (che potrebbero addirittura contemplare l’abuso sui robot, Nomura *et al.*, 2015); la ristretta gamma comportamentale del robot potrebbe strutturare un tipo di relazione che i bambini preferiscono alla relazione con gli esseri umani.

Se al robot fossero delegati compiti educativi con bambini molto piccoli, le sue capacità interattive potrebbero generare legami di attaccamento tra il bambino e il robot (Bowlby, 1988; Ainsworth *et al.*, 1978). L’attaccamento è un legame duraturo che si stabilisce tra il bambino e i suoi *caregiver*: un legame che porta il bambino a esperire piacere, gioia, senso di sicurezza e comfort quando è in loro compagnia, e angoscia in loro assenza. Sentirsi protetti quando si percepisce un pericolo per la propria incolumità fisica o psichica o in condizioni di stress psicologico, essere confortati e rassicurati, sono gli elementi primari di questa relazione. La teoria dell’attaccamento più attuale sostiene che relazioni di attaccamento sicuro con figure di riferimento principali, anche insegnanti, siano determinanti nello sviluppo e nella socializzazione dei bambini (Verschueren, Koomen, 2012). L’instaurarsi di un legame di attaccamento tra bambino e robot potrebbe determinare effetti collaterali non desiderabili. Secondo (Sharkey, Sharkey, 2010), qualora il robot, a causa delle proprie limitazioni percettive e cognitive, tendesse a produrre comportamenti anomali, imprevedibili e inappropriati al contesto, il legame tra robot e bambino potrebbe degradare verso forme patologiche di attaccamento come quelle che si producono quando un *caregiver* primario non è in grado di interpretare correttamente il comportamento del bambino, reagendo in modo non appropriato a quest’ultimo.

In uno studio mirato allo sviluppo di educatori robotici destinati all’interazione con bambini di 3 anni, le preoccupazioni derivanti dalla possibilità dell’instaurarsi di un legame di attaccamento sono state affrontate nel modo seguente: «per evitare quanto possibile forme di attaccamento emotivo al robot, e sottolineare la differenza tra robot e esseri umani, il robot KindSAR ripeteva spesso di essere un robot» (Fridin, 2014). Questa è una strategia inappropriata per scongiurare le conseguenze negative introdotte sopra. Affermare di essere “solo un robot” non interferisce con i meccanismi responsabili dell’attribuzione di cognizione ed emotività ai robot, verosimilmente collegati sia all’aspetto e al comportamento del sistema, sia alle capacità di mentalizzazione connaturate ai bambini a sviluppo normotipico; potrebbe addirittura generare il sospetto che il robot possieda capacità di autocoscienza verosimilmente assenti in KindSAR. Le preoccupazioni etiche qui accennate si affrontano, in prima battuta, sia attraverso una supervisione adulta dell’interazione tra bambini e robot sia attraverso il miglioramento del livello di comprensione del funzionamento dei robot da parte degli adulti e dei bambini.

* 1. Efficacia e responsabilità dei processi di insegnamento-apprendimento

L’acquisto di robot da destinarsi a finalità educative e didattiche comporta necessariamente costi iniziali e di manutenzione. È lecito chiedersi, assieme alle questioni sollevate nelle precedenti sezioni, se questi costi possano essere giustificati alla luce dell’efficacia dell’impiego educativo-didattico dei robot nei vari scenari che abbiamo considerato. Anche questa domanda chiama in causa questioni empiriche e questioni filosofiche. Le questioni filosofiche sono legate alla riflessione su ciò che intendiamo per “efficacia” nei vari contesti d’uso; una volta definiti dei criteri di efficacia, è necessario utilizzarli per indagare i benefici dell’applicazione dei robot in quei contesti. Sono entrambe domande aperte, tuttora affrontate dalla comunità scientifica internazionale, e di non facile risposta.

Il robot svolge ruoli diversi nei tre scenari che abbiamo considerato. Nel primo scenario, il robot è un oggetto manipolabile (programmabile) da bambini e insegnanti. La questione dell’efficacia d’uso, in questi contesti, può essere in prima battuta formulata nei seguenti termini: i robot, utilizzati come mediatori didattici, facilitano effettivamente l’apprendimento delle abilità e competenze che l’insegnante e l’educatore desidera raggiungere? Questa domanda è affrontata da (Benitti, 2012) in relazione a varie abilità e competenze disciplinari e trasversali. Si tratta tuttavia di un modo troppo astratto e poco utile di porre la questione, perché trascura una moltitudine di fattori che possono modulare i processi e gli esiti dell’apprendimento negli scenari di tipo 1. Esistono diversi tipi di robot programmabili, che differiscono per aspetto esteriore, potenzialità sensoriali e motorie, tipo di linguaggio di programmazione, tipo di interfaccia di programmazione. L’insegnante può proporre attività di programmazione che differiscono su molti piani: tipo di problema, modalità in cui il problema è proposto, modalità di gestione della dimensione comunicativa durante l’attività, modalità di gestione degli spazi, dei tempi e delle relazioni tra studenti, ruolo assunto dall’insegnante – che può intervenire correggendo direttamente gli errori degli studenti oppure accompagnandoli nel processo di riflessione ed elaborazione degli errori (Atmatzidou *et al.*, 2018; Zecca, 2016). Vi sono poi numerosi fattori contestuali e socio-demografici che riguardano l’età degli studenti, le loro conoscenze di sfondo, la loro disposizione a prestarsi ad attività di questo tipo, il loro funzionamento cognitivo, e così via. La valutazione dei benefici deve naturalmente tenere in conto quali obiettivi l’insegnante o l’educatore si proponeva di raggiungere, che possono riguardare l’apprendimento di certe conoscenze e abilità disciplinari (per esempio, informatiche o fisiche) oppure trasversali (per esempio, abilità connesse al ragionamento). Queste considerazioni suggeriscono che la valutazione dell’opportunità di introdurre i robot come strumenti manipolabili e programmabili nelle scuole, a fronte degli spesso elevati costi di acquisto e manutenzione, richiede studi empirici approfonditi che non conducono necessariamente a risposte univoche. Le più o meno aggiornate rassegne sulla robotica educativa citate in precedenza (Anwar *et al.*, 2019; Benitti, 2012; Mubin *et al.*, 2013) possono servire come puntatori a studi che analizzano gli effetti dell’introduzione di attività di robotica educativa in specifici contesti e in relazione a specifici obiettivi di apprendimento.

La valutazione dell’“efficacia” dell’impiego dei robot sociali nello scenario 2 richiede argomentazioni diverse rispetto a quelle necessarie per riflettere sullo scenario 1. Il robot, in questi casi, è chiamato a sostituire un insegnante o un educatore umano nel processo di apprendimento. Se un robot attuale o futuribile possa svolgere il ruolo di “buon” insegnante – in altri termini, se le competenze necessarie per essere un “buon” insegnante possano essere automatizzate al punto da essere implementabili nei robot attuali e futuribili – è una questione che dipende in modo cruciale da *quali* *siano* queste competenze. Gli insegnanti robotici descritti nella letteratura citata sopra – per esempio, il robot SAYA – leggono ad alta voce testi esplicativi su certi argomenti e intrattengono semplici dialoghi di tipo domanda-risposta-valutazione. È lecito chiedersi se sistemi di questo tipo possano essere detti sensatamente “insegnanti robotici”, non soltanto per la relativa rigidità delle relazioni che instaurano (basate sull’ascolto di una lezione frontale o su forme di dialogo estremamente strutturato), ma anche perché essi, allo stato attuale, sono ben lontani dal possedere alcune competenze necessarie per condurre “bene” persino lezioni frontali o dialogate, per non parlare di contesti didattici laboratoriali. Le competenze a cui stiamo facendo riferimento sono relative alla gestione dei processi di comunicazione e, più specificamente, alla calibrazione dei *feedback* forniti dall’insegnante durante i processi di insegnamento-apprendimento.

Larga parte della ricerca in psicologia dell’educazione e in didattica ha focalizzato gli studi sulla comunicazione nei processi di insegnamento-apprendimento, con un fuoco particolare sull’analisi osservativa delle interazioni in classe e del comportamento dell’insegnante (Lumbelli, 1982). In questi contesti, la forma linguistica dell’atto comunicativo non è sempre coerente con il comportamento non verbale: una constatazione o una domanda possono esprimere atteggiamenti di rifiuto, ingiunzione, disapprovazione, a seconda del tono della voce, della mimica facciale, dell’espressione dello sguardo (Watzlawick *et al.*, 1967). Le reazioni degli studenti, dunque la dinamica dell’interazione nel contesto didattico-educativo, dipendono in modo cruciale anche da questi fattori non espliciti, assumendo forme e significati diversi in diversi tipi di attività didattica (lezione, discussione, lavoro di gruppo, laboratorio). Una competenza centrale al ruolo dell’insegnante è dunque quella di riuscire a modulare il proprio comportamento verbale e non verbale in funzione dell’intenzione comunicativa, del *feedback* degli studenti e degli obiettivi di apprendimento che orientano la sua azione. L’insegnante deve anche porsi il problema di capire come i suoi stessi *feedback* vengono recepiti dagli studenti, e come essi modulano i processi cognitivi e socio-relazionali che stanno alla base dell’apprendimento (Hattie, Timperley, 2007). Queste capacità sono verosimilmente al di là della portata delle tecnologie attuali e futuribili di IA.

Alla luce di queste considerazioni è ragionevole dubitare che le riflessioni sull’“efficacia” degli insegnanti robotici (e sull’accettabilità del loro impiego nella prospettiva dell’etica delle conseguenze) possano basarsi unicamente su una valutazione empirica degli apprendimenti degli studenti in contesti sperimentali, come accade in molti degli studi citati da Belpaeme e colleghi (2018) e Li (2015). Riflessioni di questo tipo devono essere accompagnate da valutazioni empiriche dello scarto tra le competenze del robot in contesti di interazione e quelle che caratterizzano un “buon” insegnante. Su quest’ultimo aspetto la ricerca è carente, come ancora si sottolinea in (Belpaeme *et al.*, 2018).

1. Conclusioni

In questo contributo abbiamo fornito una panoramica sugli scenari d’uso attuale e futuribile dei robot per finalità educative e didattiche, e abbiamo discusso alcune questioni etiche relative alla tutela dei dati personali, all’attribuzione ingannevole di stati mentali ed emozioni ai robot sociali, alla possibilità di delegare ai robot sociali il compito di educatore e insegnante. Abbiamo messo in evidenza le questioni empiriche e filosofiche ancora aperte che è necessario affrontare per valutare la liceità di questi scenari d’uso in una prospettiva di etica deontologica e di etica delle conseguenze. Sulla base della serietà delle questioni qui introdotte è ragionevole mantenere attiva e vigile la riflessione etica sugli usi educativi e didattici dei robot, con particolare riferimento allo studio dei loro effetti sull’apprendimento degli studenti e sul loro sviluppo cognitivo, emotivo, relazionale, attraverso studi di natura filosofica ed empirica.

Capitolo 12

Etica, videogiochi e Gamification

Francesca Dagnino, Marcello Passarelli, Donatella Persico

1. Introduzione

I videogiochi sono un *medium* la cui diffusione, negli ultimi anni, ha avuto un’espansione vertiginosa. Secondo il rapporto IIDEA (*Italian Interactive Digital Entertainment Association*) del 2019, il 39% della popolazione italiana tra i 6 e i 64 anni gioca ai videogiochi, e il 26% gioca tutte le settimane (IIDEA, 2019). Contrariamente a quanto si potrebbe pensare, il fenomeno non è limitato ai soli giovani, ma riguarda tutte le fasce di età: dallo stesso rapporto, risulta infatti che il 22% dei giocatori rientra nella fascia d’età dai 45 ai 65 anni. Inoltre, la proporzione di giocatori è ripartita abbastanza equamente tra uomini e donne (53% e 47%, rispettivamente), in contrasto col persistente stereotipo che vede i videogiochi come un passatempo prevalentemente maschile.

Il consumo e la produzione di videogiochi risultano essere particolarmente concentrati negli USA, in Cina, e in Giappone, paesi che beneficiano del 75% dei profitti derivanti dai giochi in commercio[[46]](#footnote-46). Altre zone geografiche sono meno rappresentate: l’Est Europa e il Sud America raccolgono, rispettivamente, il 3% e il 4% dei profitti. Il Medio Oriente e l’Africa, complessivamente, ne raccolgono il 2%[[47]](#footnote-47). Questa sproporzione è più evidente nella produzione dei giochi che nel consumo, col risultato che - come in parte succede nel cinema - la maggior parte della popolazione globale fruisce di contenuti provenienti, in prevalenza, da zone geografiche circoscritte,

L’aumento di diffusione dei videogiochi è andato di pari passo con una loro espansione e diversificazione: oggi è possibile giocare su moltissime piattaforme, dalle console dedicate, con possibilità di realtà virtuale, agli *smartphone* usati tutti i giorni; e se un tempo il giocatore ‘tipico’ era una persona che dedicava al suo hobby una parte consistente del proprio tempo libero, oggi sono diffusi anche i *casual games*, ossia giochi progettati anche per un utilizzo occasionale e di breve durata.

Si osserva, inoltre, l’utilizzo di tecniche mutuate dai videogiochi per scopi non prettamente ludici. È il caso dei giochi educativi, sviluppati ai fini di motivare e coinvolgere i giocatori/studenti nei processi di apprendimento, ma anche della *gamification*. Il termine *gamification* indica l’utilizzo di elementi di *game design* (come trofei, classifiche a punti, l’assunzione di ruoli da parte dell’utente, stimoli alla competizione, elementi narrativi o sfide) al di fuori di veri e propri giochi (Deterding *et al*., 2011). L’utilizzo di tecniche di *gamification* è frequente nel *marketing* e nei contesti aziendali, e ha tipicamente l’obiettivo di promuovere specifici comportamenti da parte degli utenti, siano essi clienti o dipendenti. Ad esempio, l’operatore di *call center* LiveOps utilizza trofei e punteggi per motivare i propri dipendenti a seguire i propri corsi di formazione (Kim, Werbach, 2016).

In questo scenario in rapida evoluzione, è importante comprendere a fondo l’impatto a livello sociale, culturale, psicologico ed educativo dei videogiochi. In questo capitolo, discuteremo le principali problematiche etiche legate ai videogiochi e alla *gamification*. Queste includono problematiche legate al contenuto, quali i contenuti violenti o le rappresentazioni di genere, ma anche problematiche derivanti dalle meccaniche di gioco, quale la promozione della dipendenza dai giochi e lo ‘sfruttamento’ degli utenti. Prima di entrare nel merito degli aspetti etici, tuttavia, è necessario introdurre brevemente il contesto di studio e con esso la terminologia che verrà utilizzata nel capitolo.

1. Videogiochi e Gamification

Come premesso, il settore dei videogiochi ha visto una crescita importante e, di conseguenza, oggi con questo termine si fa riferimento a un insieme ampio e molto diversificato di prodotti nati anche con scopi diversi dal puro intrattenimento. Non entreremo qui nel merito delle varie tipologie di videogioco (per esempio di avventura, sportivi, simulatori, ecc.) ma ci sembra utile introdurre una distinzione tra videogiochi nati esclusivamente per l’intrattenimento ed altri nati con scopi diversi (ad esempio educativo/formativi).

I primi, i cosiddetti *entertainment games,* hanno visto un’evoluzione rapidissima negli ultimi decenni. Dai primi prodotti molto semplici ed essenziali degli anni 70/80 (videogiochi per sale giochi come *Space Invaders* o *Tetris*) si è arrivati a giochi sofisticati nella grafica e nelle meccaniche (cioè nelle regole che guidano le azioni del giocatore e le risposte del gioco) per rendere l’esperienza di gioco (il *gameplay*) sempre più coinvolgente ed immersiva. Nell’ultimo decennio, le opportunità fornite dalla tecnologia (come realtà virtuale e aumentata) combinate ad un ampliamento delle tipologie di giochi (per esempio gli *sports games* su piattaforme con *motion control* o i *casual games*) hanno significativamente ampliato il panorama degli utenti; la diffusione di videogiochi su cellulare, in particolare, ha reso l’esperienza di gioco possibile in ambienti diversi e con meno vincoli. Tutti questi fattori hanno contribuito alla crescita di questa industria che ha raggiunto un fatturato annuo di circa 145 miliardi di dollari nel 2019[[48]](#footnote-48) e hanno fatto dei videogiochi una delle forme di intrattenimento più diffuse. Sebbene creati con finalità ludiche, questi media hanno visto crescere il loro utilizzo anche in contesti diversi, ad esempio nell’educazione o per scopi riabilitativi, proprio per la loro capacità di motivare le persone a svolgere determinati compiti e di coinvolgerle maggiormente, per la loro natura interattiva e la possibilità di raccogliere dati e fornire *feedback* in tempo reale. Tuttavia, occorre tener presente che tra chi si occupa di ricerca nel settore dei giochi e le realtà industriali coinvolte nel loro sviluppo, sembra esistere una distanza difficile da colmare, specie per quanto riguarda le responsabilità sociali ed etiche derivanti dalla diffusione capillare di un medium che ormai può avere una influenza considerevole sulla cultura e sul modo di pensare dei suoi utenti (Passarelli *et al*., 2020; Perrotta *et al*., 2020).

Ciò nonostante, grazie anche alle politiche perseguite dalla Comunità Europea nel promuovere la ricerca e lo sviluppo in questo settore (Perrotta *et al*., 2020), negli ultimi decenni è fiorito l’interesse per quelli che vengono definiti *Serious Games*, *Applied Games* o *Purpose(ful) Games*, cioè giochi sviluppati con un obiettivo primario altro rispetto all’intrattenimento. Questo tipo di giochi viene utilizzato a scopo formativo in tantissimi settori, da quello educativo all’industria della difesa, in medicina, nella gestione delle emergenze, ecc. (Djaouti *et al.*, 2011). Alcuni esempi sono il gioco *Re-mission*[[49]](#footnote-49), sviluppato per incrementare l’aderenza al protocollo terapeutico in giovani pazienti affetti da tumori del sangue, o, in tutt’altro settore, *Merchant*[[50]](#footnote-50), un gioco ambientato nella Venezia del XV secolo che ha l’obiettivo di formare il personale sulla negoziazione e soluzione dei conflitti.

In questi giochi la sfida principale è mantenere in equilibrio la componente educativo/formativa con quella ludica, sfida resa ancor più difficile dal fatto che si tratta di un mercato con risorse nettamente inferiori a quello dell’intrattenimento. Questo aspetto appare di particolare importanza in contesti educativi, per cui un gioco che non risulti “divertente” andrebbe a minare il potenziale motivazionale e di coinvolgimento, limitando i benefici derivanti del suo utilizzo (Persico *et al.*, 2019).

Vi è infine la *gamification*. L'idea alla base è sfruttare il potenziale motivazionale dei videogiochi trasferendo elementi di gioco in contesti del tutto estranei al gioco stesso, ossia nel mondo reale. Come già detto, uno dei contesti di applicazione più comuni è senza dubbio quello del *marketing* (Hamari, 2013; Hamid, Kuppusamy, 2017; Zichermann, Cunningham, 2011). Molte aziende importanti hanno fatto ricorso alla *gamification* nell’ultimo decennio per incentivare diversi tipi di comportamento, dalla condivisione di informazioni con l’azienda stessa alla pubblicizzazione del marchio. Un esempio è il programma *Nike + Run Club (NRC)* condotto tramite l’utilizzo di una app, che stimolava il corridore a registrare dati su di sé e le sue abitudini di corsa fornendo supporto, piani di allenamento e sfide. Anche numerosi programmi di fidelizzazione della clientela sono stati arricchiti con elementi di gioco (*gamified loyalty programs*); Starbucks, per esempio, ha associato un bingo al suo programma *My Starbucks Rewards* (Hwang, Choi, 2019). La *gamification* viene inoltre applicata in contesti educativi (Dichev, Dicheva, 2017) e professionali come nell’esempio del *call center* citato nell’introduzione.

Uno dei temi più dibattuti in relazione alla *gamification* è proprio quello della motivazione. A questo proposito, un aspetto fondamentale è la distinzione tra due diverse tipologie di motivazione (Deci, Ryan, 2000):

* la motivazione intrinseca (o interna), che porta a cercare di raggiungere un obiettivo o a svolgere un’attività perché vissuta di per sé come stimolante, gratificante o perché risponde a un bisogno personale;
* la motivazione estrinseca (o esterna), che viene sollecitata dall‘esterno grazie a un qualche tipo di premio (denaro, gratificazioni, etc.) o alla possibilità di evitare una penalizzazione; è importante notare che questo tipo di motivazione decade al cessare dello stimolo esterno.

Questa distinzione ha dato avvio a studi sul potenziale motivazionale della *gamification* (Hamari *et al.*, 2014), e più nello specifico sui diversi elementi di gioco (punti, classifiche, narrativa, ecc.) in grado di stimolare la motivazione intrinseca (Sailer *et al.*, 2017), aspetto importante soprattutto in contesti formativi. La dicotomia motivazione intrinseca/estrinseca ha implicazioni sul piano etico che verranno presentate nel paragrafo dedicato.

Nei prossimi paragrafi, per motivi di semplicità della trattazione, si distinguerà tra le problematiche etiche legate all’uso di videogiochi e quelle della *gamification*, nonostante il fatto che le caratteristiche comuni ai due concetti facciano sì che anche le problematiche etiche si intreccino indissolubilmente.

1. Problematiche etiche dei videogiochi
	1. Contenuti e comportamenti aggressivi

La stima della proporzione di videogiochi con contenuti violenti varia a seconda della definizione di “contenuto violento”; tuttavia, si ritiene che tra il 50 e l’85% dei videogiochi contengano rappresentazioni realistiche di violenza (Dill *et al.*, 2005). In particolare nel periodo 1980-2000, in cui gli adolescenti - e in particolare gli adolescenti maschi - erano il target di riferimento per il mercato dei videogiochi, l’inclusione di contenuti violenti era usata come vera e propria strategia di *marketing* (es.: *Carmageddon,* la serie *Grand Theft Auto*, *Dragon Age: Origins*). Forse anche a causa di questa pervasività della violenza nei giochi, la possibilità che l’esposizione a contenuti violenti nei giochi possa promuovere comportamenti aggressivi nei giocatori è stata ampiamente discussa e studiata. L’interesse per questo tema non ha riguardato solo l’ambito della ricerca, ma è stata anche oggetto di dibattito da parte dell’opinione pubblica, in particolare in seguito a importanti episodi di cronaca nera (es. Il massacro presso la scuola elementare di Sandy Hook nel 2012 e la sparatoria a Clackamas dello stesso anno). Tuttavia, al momento non c’è consenso nella letteratura scientifica riguardo il rapporto tra videogiochi violenti e comportamenti aggressivi nei giocatori [CAP SEX ROBOTS pp…].

Risultano particolarmente importanti, in questo senso, due meta-analisi svolte con l’intento di raccogliere l’evidenza scientifica di decine di studi per stabilire, una volta per tutte, se questo collegamento causale ci sia. Tuttavia, le meta-analisi hanno dato risultati contrastanti: mentre la prima, condotta da Greitemeyer e Mügge nel 2014, ha concluso che i videogiochi violenti promuovono comportamenti aggressivi, la meta-analisi di Ferguson del 2015 non ha osservato un tale collegamento. Il motivo di questa incoerenza sembra risiedere in parte nella diversità degli studi inclusi nelle due meta-analisi, in quanto la seconda si focalizza solo su ricerche riguardanti bambini e adolescenti. Tuttavia, il fatto che i due studi siano giunti a conclusioni divergenti risulta scoraggiante, perché permane l’incertezza sull’esistenza o meno di questo importante effetto psicologico dei giochi.

Come premesso, parte della difficoltà nel trarre una risposta chiara deriva dall’eterogeneità degli studi su questo tema, che hanno utilizzato differenti definizioni di “contenuto violento” e differenti misure dell’aggressività. Nelle parole di Ferguson (2015), una definizione sufficientemente ampia di “contenuto violento” può portare a classificare come violenti giochi che in base al senso comune risultano innocui (ad esempio: i videogiochi della serie *Pokémon* includono combattimenti, sebbene i contenuti violenti non siano rappresentati con grafica realistica). Inoltre, un ruolo importante è giocato dall’interazione tra contenuto violento, caratteristiche di personalità del giocatore ed elementi narrativi in cui è situato il contenuto violento: nella loro rassegna della letteratura Hartmann, Krakoviak e Tsay-Vogel (2014) evidenziano come i meccanismi di ‘disimpegno morale’ (Bandura, 2002) portino alcuni giocatori, soprattutto di “sparatutto” in prima persona[[51]](#footnote-51), a giustificare la violenza virtuale de-umanizzando gli avversari in gioco pur considerandoli, allo stesso tempo, “antropomorfi” e non semplici insiemi di *pixel*.

Un filone di ricerca che appare particolarmente promettente è quello che ha studiato il rapporto tra meccaniche competitive e comportamenti aggressivi: sembra, infatti, che i giochi basati sulla competizione generino aggressività verbale e fisica, mentre la presenza di meccaniche di cooperazione sembra annullare, o persino invertire, l’effetto potenzialmente negativo dato dall’inclusione di dinamiche competitive (Ewoldsen *et al*., 2012; Velez *et al.*,2014). Alcuni autori arrivano ad affermare che potrebbe essere la competizione, e non la violenza dei contenuti, il fattore chiave che determinerebbe i comportamenti aggressivi dei videogiocatori (Adachi, Willoughby, 2011). Questa ipotesi potrebbe spiegare la forte aggressività verbale osservata nelle comunità online di giochi altamente competitivi ma privi di contenuti violenti graficamente realistici come, ad esempio, *League of Legends* (Kwak, Blackburn, 2014).

* 1. Dipendenza

Il potenziale di intrattenimento offerto dai videogiochi porta talvolta gli utenti a varcare la linea sottile tra il passatempo e la dipendenza. Quando il tempo e le energie dedicate al gioco interferiscono con la vita sociale, scolastica e lavorativa, si può parlare di rischio di dipendenza, e alcune meccaniche di gioco possono esacerbare tale rischio.

Lo studio della dipendenza da giochi è un ramo fiorente della ricerca sugli effetti psicologici dei giochi. In parte, questo interesse deriva dal fatto che nel 2013 l’*American Psychiatric Association* ha incluso l’*Internet Gaming Disorder* (IGD) tra le *Conditions for Further Study* nel *Manuale Diagnostico e Statistico dei Disturbi Mentali 5*. La dipendenza da videogiochi è stata successivamente riconosciuta come disturbo, nel 2018, dall’Organizzazione Mondiale della Sanità, che l’ha inclusa nella nuova edizione dell’*International Classification of Diseases* (ICD-11; WHO, 2020).

Questa decisione si è basata su un corpus di più di 250 studi che mostrano evidenze a favore della necessità di categorizzare come dipendenza comportamentale l’uso eccessivo dei videogiochi e/o di internet (Petry *et al.*, 2014). Tuttavia, la formulazione dei criteri diagnostici di questo disturbo è stata dibattuta, in particolare per quanto riguarda i fenomeni di tolleranza[[52]](#footnote-52) e astinenza (Kaptsis *et al.*, 2016; Kardefelt-Winther, 2014), inclusi tra i criteri dell’IGD per sottolineare il parallelismo con le dipendenze da sostanze psicotrope. Alcuni autori, inoltre, ritengono che i criteri proposti possano portare ad etichettare come patologici comportamenti di gioco nell’intervallo di normalità (Billieux *et al.*, 2015).

Ciò non toglie, tuttavia, che casi in cui il comportamento di gioco assume frequenza patologica siano stati osservati, e allo stesso tempo è evidente come, dal punto di vista di un’azienda produttrice di videogiochi, l’utilizzo di meccaniche di gioco che ne stimolino l’uso compulsivo possa rappresentare un vantaggio (McCaffrey, 2019).

In particolare, le nuove pratiche di monetizzazione sono strettamente legate alla promozione di comportamenti di dipendenza. Per “monetizzazione” intendiamo le modalità con cui un’azienda che produce videogiochi persegue il profitto. Agli albori dell’industria dei videogiochi, gran parte della diffusione dei giochi era effettuata tramite le sale giochi (*Arcade*). In questo sistema di monetizzazione la sala giochi affitta il gioco, lo mette a disposizione della propria clientela, e ha un guadagno *per partita*. Se un gioco è particolarmente avvincente (o genera dipendenza), la sala giochi ne trarrà maggior profitto, ma l’azienda che ha prodotto il gioco ne ricaverà un vantaggio solo indirettamente, tramite il maggior interesse da parte delle sale giochi.

Con il lento declino delle sale giochi, si è affermato come principale sistema di monetizzazione una logica in cui, invece, il gioco è trattato come *prodotto*: il giocatore paga una certa somma per ottenere un gioco (su supporto fisico come un DVD o, più recentemente, in forma puramente digitale), e in seguito a questa transazione il gioco è di sua proprietà. Il giocatore non paga ulteriori costi per il gioco, a meno che non decida di acquistare contenuti di gioco aggiuntivi (nella forma di “espansioni”). In questo caso, l’eventuale vantaggio ricavato dalla promozione della dipendenza è molto limitato, in quanto il guadagno è ottenuto quando il giocatore inizia a giocare, e non quando continua a farlo.

Più recentemente, tuttavia, si sono affermate forme di monetizzazione che ricavano un beneficio diretto dal gioco compulsivo. Una di queste forme è il gioco come *abbonamento*, in cui il giocatore paga periodicamente per continuare a poter giocare; sistema tipicamente usato per alcuni giochi di ruolo online massivi[[53]](#footnote-53) (MMORPG) come *World of Warcraft -* un gioco, nello specifico, per cui sono stati riportati molti casi di dipendenza (Oggins, Sammis, 2012), nonostante la presenza di meccaniche volte a scoraggiare il gioco compulsivo[[54]](#footnote-54). Più di recente, sono stati proposti abbonamenti a interi pacchetti di videogiochi (es. *EA Origin Access*), simili in principio agli abbonamenti a servizi di *streaming* di serie tv o film.

Infine, si osservano sistemi di monetizzazione più flessibili, e potenzialmente più pericolosi, come il *pay-to-win* (transazioni nel gioco per ottenere un vantaggio sugli altri giocatori o per accedere al contenuto del gioco più rapidamente; es. *Rocket League, Star Wars: The Old Republic*) e le *lootboxes* (es.: *Hearthstone*). Queste ultime sono state spesso equiparate al gioco d’azzardo, in quanto prevedono l’acquisto di un “pacchetto” il cui contenuto è sconosciuto al momento della transazione. La modalità è simile all’acquisto di un pacchetto di figurine, nel senso che il giocatore paga il costo senza la garanzia di ottenere il prodotto che effettivamente desidera, e potrebbe pertanto continuare a fare acquisti - compulsivamente - fino a ottenerlo. Il parallelo con il gioco d’azzardo è stato discusso anche a livello europeo senza tuttavia, per ora, giungere a una regolamentazione del fenomeno (Cerulli-Harms *et al.*, 2020).

* 1. Rappresentazioni

Un altro aspetto fondamentale legato al contenuto dei videogiochi è costituito dalle rappresentazioni che essi offrono. Così come la televisione ha un impatto sull’identità sociale e sui modelli cognitivi dei suoi fruitori (Mastro *et al.*, 2007), i videogiochi influenzano il modo in cui interpretiamo noi stessi e il mondo (Williams *et al.*, 2009).

Da questo punto di vista, le rappresentazioni più studiate, e criticate, sono quelle relative al genere. Nei primi decenni dell’industria dei videogiochi, il *target* di mercato erano prevalentemente gli adolescenti maschi, e gli sviluppatori stessi erano in massima parte maschi (Chess *et al.*, 2016). Non dovrebbe sorprendere, quindi, che la maggior parte dei personaggi femminili dei primi videogiochi fossero ipersessualizzati, anche come esplicita strategia di *marketing*. Il personaggio-icona di questo fenomeno è Lara Croft della serie *Tomb Raider*, una protagonista da certi punti di vista positiva - in quanto donna capace che svolge anche attività stereotipicamente maschili - ma la cui sessualizzazione è innegabile (la più recente Bayonetta, dalla serie omonima, segue lo stesso stampo con una sessualizzazione più evidente ed estrema).

Questo fenomeno sta lentamente cambiando, anche in seguito all’aumento in proporzione delle giocatrici donne, per quanto i maschi siano ancora considerati il target di mercato principale (Jussim *et al.*, 2015). Si vedono sempre più personaggi femminili non-stereotipati e non oggettivati (Lynch *et al.*, 2016), anche nel ruolo di protagoniste, come Max (*Life is Strange*) o Ellie (*The Last of Us*). Tuttavia, nelle comunità di giocatori - e in particolare nelle comunità legate a giochi competitivi online - si continuano a osservare frequenti fenomeni di ostilità ed emarginazione verso le donne giocatrici (Paaßen *et al.*, 2017), e i videogiochi che propongono personaggi femminili non sessualizzati sono spesso criticati dagli stessi giocatori (la reazione a *The Last of Us Part 2*, ad esempio, è stata particolarmente accesa; Crossman, 2019). Un evento che ha reso questo atteggiamento particolarmente evidente è la controversia *#GamerGate*, durante e in seguito alla quale professioniste operanti nel mondo dei giochi sono state attaccate e minacciate dalle comunità di giocatori (Huntermann, 2013; Sarkeesian, 2013).

Una simile evoluzione nelle rappresentazioni riguarda i personaggi LGBTQ+: quando il mercato dei videogiochi era meno ampio e diversificato, le rappresentazioni di personaggi LGBTQ+ erano rare, principalmente relegate a ruoli marginali, e spesso stereotipiche o negative (es. Crassius Curio in *Morrowind*; Shaw, Friesem, 2016). Quando le rappresentazioni di personaggi LGBTQ+ hanno iniziato a essere più diffuse, hanno preso prevalentemente la forma di contenuto di gioco opzionale, con cui è il giocatore a decidere se interagire (rientrano in questa categoria i personaggi di supporto nei giochi di ruolo della casa produttrice *Bioware*, che ha incluso la possibilità per i giocatori di perseguire una relazione virtuale con personaggi LGBTQ+ a partire da *Jade Empire*). Ancora più recentemente, hanno iniziato a essere più diffusi personaggi LGBTQ+ come personaggi principali, con un ruolo nella narrazione (es. Ellie di *The Last of Us Part 2*, Undyne e Alphys in *Undertale*). Tuttavia, questo aumento nella rappresentatività riguarda principalmente i personaggi gay, lesbiche e bisessuali, mentre personaggi transessuali sono ancora molto rari (alcuni esempi sono Krem in *Dragon Age: Inquisition* e Madeline in *Celeste*).

Risulta essere meno positiva l’evoluzione della diversità etnica dei personaggi: i personaggi dei videogiochi sono ancora in grandissima parte bianchi (Dill *et al*., 2005; Williams *et al.*, 2009), anche in ambientazioni narrative (come fantasy e fantascienza) in cui il concetto di ‘etnia’ può essere affrontato più liberamente. Questa sproporzione nella rappresentatività riguarda anche i ‘casual games’ (Wohn, 2011), una categoria di giochi che, invece, è piuttosto bilanciata per quanto riguarda le rappresentazioni di genere. La carenza di rappresentazione delle minoranze etniche può stupire, considerato che la percentuale di giocatori appartenenti a minoranze rappresenta un’ampia porzione del mercato (Duggan, 2015). Tuttavia, ciò risulta perfettamente in linea con la proporzione di persone di colore che lavorano nell’industria dei videogiochi (circa il 3%; IGDA, 2015). Inoltre, come menzionato nell’introduzione, c’è una sproporzione geografica nella produzione di videogiochi: Sudamerica, Africa e Medio Oriente sono zone geografiche con una produzione relativamente scarsa (6% della produzione mondiale, complessivamente), e questo esacerba il problema della scarsa rappresentazione di culture non-occidentali e non-eurasiatiche.

* 1. Inclusione e bisogni speciali

Un altro aspetto in cui i videogiochi sollevano questioni etiche è legato all’inclusione in termini di accessibilità. In altre parole: in che misura i videogiochi tendono ad escludere giocatori con deficit percettivi, cognitivi e/o motori? E in che modo sarebbe possibile rendere i giochi più inclusivi?

Questo aspetto è stato particolarmente studiato in relazione ai giochi educativi o, comunque, non di intrattenimento, in cui il problema dell’accessibilità da parte di giocatori con disabilità cognitive è particolarmente sentito. Tuttavia, anche in questo settore la quantità di studi scientifici è relativamente scarsa (Jimenez *et al.*, 2015).

Fondamentalmente, sebbene i diversi tipi di disabilità vengano spesso trattati assieme, ognuno di essi presenta caratteristiche uniche, genera difficoltà differenti e richiede soluzioni specifiche (Yuan *et al.*, 2011). Le disabilità motorie, ad esempio, vengono affrontate principalmente tramite soluzioni hardware: persone con disabilità motorie possono utilizzare *joystick* e altri dispositivi che vengano incontro alle loro necessità. Nel caso di disabilità motorie quasi totali come la quadriplegia, l’uso di *eye tracking* e *joystick* facciali (come *QuadStick*) permette a un giocatore sufficientemente determinato di giocare qualunque tipo di gioco, anche quelli non progettati tenendo conto di queste esigenze. Il principale limite di queste soluzioni è legato al prezzo dei dispositivi, spesso a carico del giocatore. Tuttavia, l’accesso ai videogiochi per persone con disabilità motorie può accrescerne il senso di autoefficacia con ricadute positive sul benessere individuale (Ijsselsteijn *et al.*, 2007), oltre ad essere un possibile fattore motivante all’interno di percorsi di riabilitazione e fisioterapia (de Greef *et al.*, 2013).

La risposta a disabilità di tipo percettivo, invece, passa principalmente tramite soluzioni software, e richiede un’attenzione particolare da parte degli sviluppatori di videogiochi. Ad esempio, per poter essere fruito da non udenti, un videogioco necessita di sottotitoli chiaramente leggibili per i dialoghi, indicatori visivi per gli eventi che si affianchino ai segnali sonori, e sistemi di *feedback* multisensoriali[[55]](#footnote-55). La situazione per quanto riguarda i videogiochi commerciali sarebbe migliorabile: sebbene siano sempre di più i videogiochi che includono sottotitoli - più per rispondere alle esigenze di un mercato internazionale che per considerazioni legate all’accessibilità da parte di non udenti - questi sottotitoli sono spesso implementati in modo inappropriato: troppo piccoli, spezzati troppo poco o male, poco leggibili nel contesto della scena, troppo veloci, o privi di informazioni sugli effetti sonori[[56]](#footnote-56).

Ancora meno soddisfacente è la situazione che riguarda le soluzioni offerte per le disabilità visive. I videogiochi sono un *medium* fortemente visivo, e un giocatore ipovedente o non vedente difficilmente può fruire della maggior parte dei giochi in commercio. Per poter venire incontro alle necessità di persone non vedenti, i giochi dovrebbero cambiare radicalmente il loro modo di presentare le informazioni (anche se aumentare dimensione e contrasto degli stimoli visivi è già d’aiuto per persone ipovedenti). Tuttavia, anche individui con una disabilità lieve e relativamente diffusa come il daltonismo spesso incontrano problemi significativi. Molti giochi utilizzano schemi verde/rosso o rosso/blu per distinguere tra elementi positivi e negativi nel gioco, ma questo elemento distintivo potrebbe essere quasi invisibile per un individuo daltonico (a seconda del tipo di daltonismo). Uno sviluppatore di giochi attento a queste necessità può venire pienamente incontro alle esigenze di giocatori con daltonismo; è il caso di *Hue*, un puzzle game basato esclusivamente sui colori che, tramite il coinvolgimento di comunità di giocatori con daltonismo, è stato adattato per essere fruibile pienamente anche da loro (Plothe, 2018). Il fatto che giochi meno basati sui colori non presentino opzioni per daltonici, anche a fronte della diffusione del daltonismo, è forse indice di una mancanza di sensibilità degli sviluppatori sui temi dell’accessibilità e del *design-for-all* (Heron, 2012).

Diversa, infine, è la situazione per quanto riguarda le disabilità cognitive: in questo caso si osserva una situazione per cui, *by design*, giocatori con disabilità cognitive presentano maggiori difficoltà per certe tipologie di gioco (ad esempio, puzzle o videogiochi strategici, che presentano sfide prevalentemente cognitive) mentre risultano loro più accessibili o anche completamente accessibili altri generi di gioco (ad esempio, *platform[[57]](#footnote-57)* e giochi d’azione che presentino sfide prevalentemente motorie). Proporre diversi livelli di difficoltà, anche molto semplici, può rendere questi giochi ancora più accessibili. Inoltre, ridurre il livello di sovraccarico percettivo può venire incontro a individui con alcuni disturbi (quali i disturbi dello spettro autistico).

1. Problematiche etiche della gamification

La *gamification* presenta problematiche etiche peculiari, in parte legate al suo scopo, ossia promuovere specifici comportamenti nei “giocatori”. La *gamification* non è un gioco, e non ha il fine primario di essere venduta come mezzo di intrattenimento. Al contrario, come abbiamo visto, essa utilizza l’intrattenimento come strumento per coinvolgere gli utenti e “alterarne” il comportamento, tramite tecniche motivazionali generalmente basate sulla motivazione estrinseca (Hamari *et al.*, 2014). Ad esempio, l’uso di premi virtuali e competizioni tra utenti può motivare potenziali clienti a fare acquisti e formandi o dipendenti a svolgere compiti o mansioni ripetitive, poco interessanti e/o poco motivanti.

L’uso della *gamification* è stato fortemente criticato da alcuni autori, che l’hanno descritta come uno strumento manipolatorio che “usa” l’utente per i propri fini, come la raccolta dati a fini commerciali. È celebre il neologismo *exploitationware*, usato da Ian Bogost per descrivere la *gamification* (Bogost, 2013), in riferimento al fatto che il guadagno del “giocatore” è nettamente inferiore a quello di chi “gamifica” e questo squilibrio lo spinge a definirlo sfruttamento. In particolare, sono riportati casi in cui le ricompense virtuali fornite da un sistema *gamificato* sono state utilizzate in sostituzione di compensi finanziari, di fatto evitando di pagare in denaro il lavoro degli utenti (Kim, Werbach, 2016). Questi ultimi autori, in particolare, segnalano come la rapida adozione della *gamification* da parte delle aziende abbia fatto sì che questa diventasse una pratica comune prima ancora che fosse discussa sul piano etico, e ignorando le potenziali divergenze tra gli interessi degli utenti e quelli di chi fornisce il sistema gamificato. Infine, il frequente uso di tecniche motivazionali basate sulla competizione, spesso molto efficaci nell’obiettivo di promuovere il comportamento desiderato, può però provocare emozioni negative negli utenti e favorire un clima di ostilità (si noti il parallelo col fenomeno, discusso sopra, per cui videogiochi molto competitivi favoriscono comportamenti aggressivi nei giocatori).

Vi è poi la questione motivazionale. Sicart (2015), ispirandosi alla filosofia di Aristotele e alla psicologia positiva (Deci, Ryan, 2008) - teorie che condividono il focus sull’*eudaimonia*, sull'arte e la pratica di vivere bene, e sul perché e come si persegue questo obiettivo - avvia una riflessione critica sulla *gamification*. Sicart, partendo proprio da esempi quali il programma *Nike + Run Club*, evidenzia come molti sistemi gamificati trasformino in giochi competitivi (con premi e riconoscimenti) attività che dovrebbero condurre ad avere una miglior qualità di vita, mettendo in secondo piano la riflessione, l’impegno e la spinta interiore che dovrebbero essere alla base di queste attività.

Questo spostamento verso la motivazione estrinseca presenterebbe un’inevitabile conseguenza, evidenziata anche da altri autori (Rigby, 2015), sminuirebbe cioè il valore intrinseco dell’attività stessa spostando il focus dal piacere insito nello svolgere l’attività all’esterno, ossia al perseguimento della ricompensa. È intuibile che questo spostamento non sarebbe affatto desiderabile in alcuni contesti, quale ad esempio quello educativo, in quanto fallirebbe nell’intento di stimolare la motivazione ad apprendere, abituando gli studenti a perseguire la ricompensa, indipendentemente dalla qualità dell’apprendimento.

Una considerazione necessaria è che sicuramente l’analisi degli aspetti etici nell’attuale dibattito dovrebbe essere più articolata perché le pratiche di *gamification* non sono sempre eticamente controverse e non possono essere affrontate come un unicum.

Esistono casi di *gamification* generalmente considerati positivi. Un esempio di particolare successo è l’app per l’apprendimento delle lingue *Duolingo*, che utilizza diverse tecniche (livelli, punteggi, sfide, ecc.) per motivare gli utenti all’apprendimento di una lingua straniera. Gli utenti sono consapevoli del fine di tali tecniche e il comportamento promosso (la ripetizione di esercizi per l’apprendimento di una lingua) è in sintonia con i loro obiettivi. L’app, gratuita, trae profitto dalla vendita di traduzioni effettuate dagli utenti. Non tutti gli utenti, tuttavia, sono consapevoli di produrre traduzioni che vengono vendute[[58]](#footnote-58), così come gli utenti di app gamificate ‘positive’ – quali quelle per motivare all’attività fisica – non sono consapevoli della tipologia e quantità di dati che vengono raccolti. Dunque, sebbene l’influenza dell’app sugli utenti sia positiva, anche in questo caso si osservano potenziali problematicitàetiche. In ogni caso, una *gamification* socialmente responsabile, che miri a promuovere il benessere dell’utente, i comportamenti pro-sociali, o l’automiglioramento, è possibile. Tuttavia, l’utilizzo di *gamification* da parte di aziende è tendenzialmente poco regolamentato e tende a seguire pure logiche di mercato (Thorpe, Roper, 2019). Per evitare che la *gamification* si affermi come tecnica manipolativa e come strumento di sfruttamento occorre prestare particolare attenzione agli aspetti di autonomia, volontarietà, consenso, autoriflessione, trasparenza ed equità (Kim, Werbach, 2016).

1. Promuovere valori e consapevolezza etica attraverso i giochi

Nelle sezioni precedenti si è discusso di come i giochi e le meccaniche da essi implementate sottendano spesso problematiche etiche più o meno esplorate in ambito di ricerca e discusse dal grande pubblico, e di come spesso gli interessi di mercato entrino in conflitto con gli interessi individuali di giocatori e utenti. Tuttavia, i giochi possono anche incontrare l’etica in modo consapevole ed essere usati come strumenti - tramite i contenuti narrativi e le meccaniche di gioco - di riflessione e confronto con problemi etici significativi (Schrier, 2015). Uno dei recenti sviluppi nell’industria dei videogiochi, legato alla differenziazione dei videogiochi negli ultimi anni, riguarda infatti il crescente numero di giochi che non si caratterizzano come pura esperienza ludica, ma che presentano un contenuto artistico o affrontano tematiche delicate o attinenti alla sfera filosofica/etica tramite un forte coinvolgimento emotivo del giocatore (Farber, Schrier, 2017). Perché questo possa avvenire al meglio, occorre: a) sensibilizzare gli sviluppatori di giochi all’intrinseco risvolto etico dello sviluppo dei videogiochi, e b) supportare sviluppatori, educatori, e formatori nell’utilizzo degli elementi di gioco che maggiormente possono promuovere la riflessione etica e valori positivi come la giustizia sociale, l’uguaglianza, o l’etica ambientale (Earp *et al.*, 2018).

Un esempio di iniziativa di questo tipo è *Values at Play*[[59]](#footnote-59), che si occupa di formare e svolgere ricerca sull’uso dei giochi nella riflessione etica (Flanagan, Nissenbaum, 2014). *Values at Play* è nata come cornice metodologica per la promozione della consapevolezza sugli aspetti etici dei giochi, ma è cresciuta fino a diventare un’iniziativa per l’integrazione dei principi di sviluppo eticamente consapevole nei curricula formativi per gli sviluppatori di giochi.

Lo sviluppo eticamente consapevole implica un’attenzione alle meccaniche e ai contenuti del gioco finalizzata ad evitare effetti negativi sul giocatore e sulla società, come alcune delle possibili conseguenze descritte in questo capitolo; tuttavia, può anche significare sviluppare giochi che abbiano l’esplicito intento di apportare cambiamenti benefici sul piano cognitivo, affettivo, o degli atteggiamenti (Formosa *et al.*, 2016). Questo non significa necessariamente sviluppare *serious games* - alcuni esempi di giochi commerciali di intrattenimento che stimolano riflessioni etiche sono già disponibili. Ad esempio, *This War of Mine* affronta temi come la guerra, la sopravvivenza e la solidarietà nel contesto di un assedio visto dalla parte dei civili, facendo leva sul senso di empatia del giocatore. Oppure, *Spec Ops: The Line* propone una riflessione sulla glorificazione e giustificazione della violenza nel contesto di uno sparatutto in prima persona, rendendo evidente come i videogiochi stessi normalizzino fenomeni come la deumanizzazione.

1. Conclusioni

In questo capitolo, abbiamo presentato una panoramica delle principali problematiche etiche riguardanti i videogiochi e la *gamification*. Queste presentano caratteristiche differenti, e sono legate sia al contenuto dei giochi (ad esempio, le rappresentazioni delle minoranze) sia alle loro meccaniche (ad esempio, meccaniche che inducono dipendenza). Per questo motivo, arginare le potenziali conseguenze negative dei giochi richiede la ricerca di soluzioni su più fronti. Di sicuro, un primo passo da compiere è promuovere la consapevolezza dell’esistenza di queste problematiche e delle dinamiche che le favoriscono. I giocatori, infatti, immersi nella dinamica del gioco, si soffermano raramente a riflettere sugli effetti (desiderabili o meno) del loro uso. Persino in ambito educativo, analogamente, l’entusiasmo con cui numerosi docenti hanno abbracciato i giochi e la *gamification* come possibile medium didattico capace di motivare gli studenti, suggerisce una certa superficialità in merito alle implicazioni etiche di questo tipo di strumenti didattici. Al tempo stesso, gli insegnanti e gli educatori non possono ignorare il fatto che i giochi sono ormai parte delle cosiddette “ecologie di apprendimento” dei loro allievi e, di conseguenza, non possono prescindere dal fatto che essi, nel bene e nel male, “apprendono” dai giochi, ossia ne subiscono le influenze, assorbendone spesso inconsapevolmente i valori, siano esse positivi o negativi (Persico *et al*., 2019). In secondo luogo, si dovrebbe considerare come le caratteristiche individuali dei giocatori possano favorire o contrastare effetti negativi dei giochi - ad esempio, alcuni giocatori possono essere particolarmente vulnerabili ai contenuti violenti o all’esasperata competitività dei giochi online, ed essere a rischio di produrre comportamenti aggressivi in risposta.

È importante ricordare, tuttavia, che il mondo dei videogiochi è un campo in continua evoluzione. Negli ultimi decenni, il mercato e l’offerta di videogiochi si sono diversificati enormemente, risolvendo in parte alcuni problemi (ad esempio, le rappresentazioni di minoranze oggi sono tendenzialmente più positive ed equilibrate) ancorché generandone di nuovi (ad esempio, molte forme di monetizzazione eticamente inadeguate sono uno sviluppo relativamente recente). Un monitoraggio attento del *medium* può aiutare a identificare e controllare le problematiche etiche preventivamente, impedendo che le modalità di sviluppo dei giochi e l’uso di tecniche di *gamification* siano lasciati esclusivamente in mano alle logiche di mercato.

*Parte quarta*

Segnavia

Capitolo 13

Codici etici e documenti di indirizzo

Fabio Fossa, Viola Schiaffonati

1. Introduzione

In questo capitolo si presentano i codici etici e i documenti di indirizzo relativi alla robotica e all’Intelligenza Artificiale nella loro connessione all’etica professionale e alla dimensione istituzionale. Il loro scopo è di regolamentare o influenzare la condotta dei diversi attori coinvolti nello sviluppo dei sistemi autonomi e intelligenti, cioè di chiarire quali siano le responsabilità dei professionisti, quali i valori che devono guidarne la condotta e quali le direzioni da impartire all’avanzamento tecnologico perché promuova il benessere del genere umano.

La stesura di codici etici e di comportamento costituisce una pratica fondante della cosiddetta etica professionale. L’etica professionale si occupa degli obblighi, dei doveri e delle responsabilità che attengono ad un individuo in quanto svolge una certa professione. Dal momento che i professionisti sono titolari di competenze specifiche necessarie per l’esecuzione di servizi spesso cruciali per il benessere sociale, è importante che tra i membri della società e coloro che svolgono una data professione si stabilisca un rapporto di fiducia reciproca. Una condizione per cui si sviluppi una solida e duratura fiducia in una professione, però, è che i professionisti siano consapevoli degli impatti etici del loro lavoro e si sforzino di promuovere una condotta che sia, da una prospettiva etica, all’altezza del loro ruolo e delle responsabilità che ne derivano. Per questo motivo, nella seconda metà del secolo scorso[[60]](#footnote-60) molte associazioni di professionisti hanno codificato i valori fondamentali che ogni membro è tenuto a rispettare e promuovere durante lo svolgimento della professione – valori che costituiscono parte della professione stessa e concorrono a formarne l’immagine pubblica. Nel corso degli anni, l’idea di codificare gli aspetti etici relativi ad una certa attività sociale ha superato le barriere della professione intesa in senso stretto per essere recepita in molti e diversi contesti organizzativi (istituzioni, aziende, centri di ricerca, e così via).

Tra i risultati di questo sforzo figurano i codici etici e di condotta, ovvero documenti in cui le diverse organizzazioni indicano quali sono i comportamenti e le linee guida che i loro membri devono seguire per comportarsi in modo responsabile (van de Poel, Royakkers, 2011). I codici di condotta, però, non sono tutti uguali. Innanzitutto, possono essere *codici professionali* e *codici aziendali*. Un esempio del primo tipo è il codice della ACM, che analizzeremo nel paragrafo successivo, mentre un esempio del secondo tipo sono i codici di aziende come Microsoft[[61]](#footnote-61) o IBM[[62]](#footnote-62). Inoltre, in entrambi i casi i codici di condotta possono essere di tre tipi: *aspirazionali*, ossia esprimere i valori morali di una professione o di un’azienda; *consultivi*, ossia aiutare professionisti o dipendenti di un’azienda a esercitare il loro giudizio morale in situazioni concrete; o *disciplinari*, ossia far sì che il comportamento di professionisti e dipendenti si conformi a certi valori o a certe norme. Queste diverse tipologie ci fanno anche capire che i codici sono formulati per ragioni diverse, che vanno dal cercare di aumentare la consapevolezza morale e l’identificazione di una determinata professione con certe norme e valori, fino alla possibilità di migliorare l’immagine di un’associazione o di un’azienda agli occhi esterni. Nei paragrafi successivi prenderemo in considerazioni due esempi particolarmente significativi: il codice etico dell’ACM (*Association for Computing Machinery*), volto a chiarire e promuovere le responsabilità morali e i valori fondamentali che ineriscono alle professioni informatiche, e il testo *Ethically Aligned Design* prodotto dall’IEEE (*Institute for Electrical and Electronical Engineering*), il quale propone una visione ricca e dettagliata delle numerose intersezioni tra le attività dei suoi membri e la dimensione etico-sociale.

Oltre al mondo delle professioni e delle aziende, la codificazione dell’etica riguarda anche la dimensione pubblica e istituzionale. Le tecnologie di cui abbiamo trattato nei capitoli precedenti sono il risultato di articolati processi sociali e, dunque, possono esistere solo nel contesto di complesse strutture politiche e istituzionali che ne normano e sostengono lo sviluppo. Per questo motivo, è di fondamentale importanza che considerazioni di carattere etico siano integrate anche alla regolamentazione dei processi di sviluppo e impiego dei sistemi autonomi e intelligenti; ma in che forme, in che misura, e con quali modalità? Per chiarire questi aspetti e guidare il futuro della tecnologia verso direzioni più eque e sostenibili molte organizzazioni nazionali e internazionali, attraverso la stesura di documenti di indirizzo, hanno contribuito allo sforzo pubblico volto a delineare princìpi e linee guida per promuovere lo sviluppo e l’uso etico delle nuove tecnologie. Per comprendere meglio di che cosa si tratta, nel paragrafo 4 analizziamo *Orientamenti etici per un’IA affidabile,* il più importante documento di raccomandazioni a livello europeo.

Per quanto il numero di simili documenti sia già elevato e in continua crescita, non mancano però dubbi sulla loro efficacia e sul loro reale impatto per quanto riguarda la promozione del comportamento responsabile. Chiudiamo il capitolo con una disamina delle intenzioni, dei limiti e delle critiche che accompagnano questo approccio al fine di evitare sopravvalutazioni della sua portata e di comprenderne così il ruolo e le possibilità effettive.

1. ACM Code of Ethics and Professional Conduct

Per chiarire meglio cosa sia un codice etico e da cosa sia composto, in questo paragrafo ci concentriamo sul codice della ACM, la più grande associazione mondiale di professionisti dell’informatica. Sebbene il codice della ACM abbia una lunga storia (è stato approvato nel 1992), ci concentriamo qui sulla sua ultima versione, messa a punto dopo un lungo e articolato processo di revisione (2018).

In generale, al centro dei codici professionali ci sono *integrità* e *onestà*, ovvero valori che comportano una visione della professione in cui non solo l’aspetto di preparazione professionale è importante, ma anche il fatto che la professione sia svolta in maniera onesta e tale da generare fiducia. Il che significa, per esempio, che un professionista deve rispettare gli accordi e non deve essere manipolato nel suo giudizio. Significa inoltre che un professionista deve evitare il conflitto di interesse, ossia le situazioni in cui la realizzazione di un interesse sia in conflitto con i propri doveri professionali. Un conflitto di interesse è per esempio il caso in cui un professionista svolge il ruolo di consulente per due aziende che sono in competizione fra loro. È evidente che il discorso sui valori essenziali per i codici professionali è legato a doppio filo al modo di intendere una professione sia in generale sia nello specifico. Nel caso dell’informatica, questa riflessione appare particolarmente importante: infatti, oltre alle caratteristiche che sono generalmente menzionate (uso di conoscenza specializzata, monopolio nel gestire la professione, valutazione condotta dai pari, oltre a orientamento al servizio e standard etici), vi sono alcune peculiarità che influenzano il modo in cui la professione è intesa anche nei codici di condotta. Deborah Johnson (2009), fra gli altri, ha soffermato l’attenzione su questi aspetti, a partire dal riconoscimento dell’importanza della cosiddetta prospettiva socio-tecnica per cui le tecnologie informatiche sono sia plasmate dalla società sia contribuiscono a plasmarla. Dato che i professionisti informatici possono ricoprire molti ruoli anche assai diversi fra loro (programmatori, architetti di sistema, gestori di database, ingegneri del software, amministratori di sistema, ecc.), è importante chiedersi quali siano le caratteristiche dell’informatica come occupazione e quali tipi di responsabilità siano associati all’essere esperti di informatica, senza tralasciare di valutare se e come in futuro queste responsabilità cambieranno.

Il codice ACM è rivolto a tutti i professionisti informatici e offre un punto di partenza chiaro per analizzare la struttura, gli obiettivi e le criticità dei codici etici in generale. La ACM è una associazione professionale che, secondo le sue stesse parole, ha l’obiettivo di promuovere l’informatica in linea con i più alti standard professionali in modo tale che il suo esercizio possa avere un impatto positivo sull’intera società. Questo codice etico può quindi essere visto come una sorta di contratto sociale che interessa tutti coloro che fanno parte della ACM e chiarisce quali siano i loro doveri non solo verso l’associazione, ma verso la società in senso lato. È interessante notare che, nelle ragioni avanzate per giustificare la revisione di questo codice, sono enfatizzate alcune delle peculiarità dell’informatica, prima fra tutte il fatto che si tratti di una disciplina che evolve rapidamente (Brinkerman *et al.*, 2016).

Il codice dell’ACM ha svolto un ruolo fondamentale nel guidare i processi di decisione: è stato usato come guida per educare i giovani che si affacciano alla professione informatica, come strumento di supporto per le decisioni degli informatici di professione, come standard usato dal pubblico nel valutare il grado di professionalità degli informatici, e anche come aiuto per dirimere questioni legali e più ampie tensioni etiche. Tuttavia, negli ultimi anni, molte delle questioni sottoposte al COPE (*Committee on Professional Ethics*) dell’ACM relative all’applicazione del codice riguardano proprio l’IA, la robotica e l’apprendimento automatico. Si tratta di cambiamenti straordinari non solo dal punto di vista prettamente tecnologico, ma anche dal punto di vista di come queste tecnologie sono profondamente integrate nelle nostre strutture sociali e nelle nostre vite quotidiane. Proprio da questa consapevolezza è stata avviata la revisione che ha portato alla versione 2018 del codice, alla cui base vi sono i seguenti principi:

* Il codice deve documentare le responsabilità e gli obblighi etici e professionali di tutti i professionisti che lavorano nell’ambito informatico, includendo fra questi i professionisti presenti e futuri, i docenti e tutti coloro che usano l’informatica con impatti significativi.
* Il codice dovrebbe esprimere il consenso della comunità informatica rispetto a certe questioni etiche.
* Il codice dovrebbe essere usato come guida per prendere decisioni.
* Il codice dovrebbe educare sia il pubblico generale sia coloro che aspirano a diventare professionisti informatici rispetto agli obblighi di questa professione.

Il codice è diviso in 4 parti, oltre a un preambolo in cui si spiega a chi è rivolto, perché è strutturato in un certo modo, nonché la sua funzione essenziale nello stimolare un dibattito pubblico rispetto ai problemi etici in gioco. In questo preambolo si sottolinea che il codice non deve essere visto come un algoritmo per risolvere problemi etici, ma come un punto di partenza per un processo di decisione etico rispetto ai problemi morali che si possono incontrare in questa professione. Questo è, a nostro parere, un aspetto molto importante soprattutto quando rivolto a una comunità, come quella informatica, che generalmente non ha una specifica formazione etica. Che i codici etici e di comportamento non siano processi algoritmici e neppure libri di regole ben definite e pronte all’uso è un aspetto decisivo da tenere in costante considerazione non solo nella loro redazione, ma anche nel modo in cui vengono utilizzati.

La sezione 1 delinea i principi etici fondamentali che formano il cuore del codice. Si tratta di principi generali, come per esempio contribuire al benessere delle persone, evitare di causare danni, essere onesti, giusti, rispettare la privacy e il lavoro richiesto per produrre nuove idee e artefatti computazionali. Ogni principio è completato da linee guida che spiegano come applicarlo: per esempio nelle linee guida associate al principio 1.2, che recita «un professionista informatico dovrebbe evitare danni», è illustrato cosa si intende per danni, come possono essere evitati e come possano verificarsi anche quando le intenzioni di partenza sono buone.

La sezione 2 illustra le responsabilità professionali specifiche legate all’informatica: cercare di mantenere una qualità sempre alta sia dei processi sia dei prodotti; consentire valutazioni ad ampio raggio dei sistemi informatici e dei loro impatti, compresa l’analisi dei possibili rischi; promuovere la comprensione dell’informatica, e delle tecnologie a essa collegate, presso un pubblico più ampio; progettare e realizzare sistemi che siano robusti e utilizzabili in modo sicuro. Anche in questo caso al singolo principio sono sempre associate linee guida che aiutino i professionisti a comprenderlo meglio e ad applicarlo in modo significativo. Per esempio il principio 2.7 recita «un professionista informatico dovrebbe promuovere la comprensione dell’informatica, delle tecnologie a essa collegate e delle sue conseguenze presso un pubblico più ampio». Nelle linee guida corrispondenti si spiega come questa comunicazione verso un pubblico più ampio debba essere chiara, rispettosa delle conoscenze di ciascuno e debba tenere in considerazione anche le informazioni inaccurate e inattendibili.

La sezione 3 presenta i principi che si applicano a coloro che hanno un ruolo dirigenziale o di *leadership*. Anche in questo caso viene enfatizzata la centralità del bene pubblico e i principi sono sia di carattere generale (per esempio, gestire il personale e le risorse in modo da migliorare la qualità di vita) sia specifici dell’informatica (per esempio, prestare particolare attenzione allo sviluppo di quei sistemi informatici che sono destinati a diventare parte integrante dell’infrastruttura della società).

Infine, la sezione 4 illustra i principi relativi all’adozione del codice e i casi in cui esso non viene rispettato. I professionisti informatici devono promuovere e rispettare i principi del codice, nonché trattare le sue violazioni come incompatibili con l’appartenenza all’ACM. È interessante notare che nel processo di revisione questa parte è stata ampiamente rivista: se nel codice del 1992 le violazioni del codice avevano un’unica conseguenza, ossia l’espulsione dalla ACM, nella versione del 2018 vengono riconosciuti diversi gradi di violazione e incluse opportunità di rimediare a queste violazioni con provvedimenti meno severi dell’espulsione.

Non è questo il luogo, anche per limiti di spazio, per analizzare nel dettaglio i cambiamenti messi in atto al termine del processo di revisione[[63]](#footnote-63). Ci preme però sottolineare alcuni aspetti generali. Il primo è che nella versione del 2018 sono stati aggiunti nuovi princìpi tesi ad ampliare la visione di responsabilità professionale. Un esempio è il principio 2.9 che sottolinea come i professionisti informatici abbiano la responsabilità di garantire che i sistemi creati siano sicuri. Il secondo è che, nonostante una delle ragioni per la revisione sia dovuta alla massiccia diffusione di tecnologie come l’IA, il codice non fa esplicito riferimento a specifiche tecnologie, dato che ha l’ambizione di usare un linguaggio che possa essere applicato anche alle nuove tecnologie che emergeranno. Il terzo è una costante attenzione all’uso del linguaggio più adatto da usarsi: a titolo di esempio si veda l’interessante riflessione sull’uso di *should* e *must* associati a ciò che gli informatici devono o dovrebbero fare (Gotterbarn *et al.*, 2018). Infine, un ulteriore elemento riguarda il processo stesso di revisione e il modo in cui i cambiamenti sono stati messi in atto. È stata infatti creata un’apposita *task force* che ha prodotto diverse bozze di revisione; ogni bozza è stata sottoposta a una consultazione pubblica (non solo fra i membri dell’ACM); i suggerimenti ritenuti validi sono stati implementati, quelli rifiutati sono stati motivati (ed è spesso interessante capire le ragioni per cui alcune proposte non sono state accettate); ma soprattutto, oltre alla revisione del codice, l’ACM ha messo un a punto un progetto più ampio, denominato “Integrity Project” a supporto della promozione della consapevolezza etica nelle professioni informatiche[[64]](#footnote-64). Si tratta di materiali a supporto del codice, tesi a far sì che il codice non rimanga solo un insieme di ideali di difficile applicazione, ma sia accompagnato da esempi, casi di studio, spiegazioni più dettagliate (inclusa la video serie *Ask an ethicist*[[65]](#footnote-65)) che aiutano a concretizzare i principi del codice e a supportare la loro applicazione.

1. IEEE Ethically Aligned Design

Un altro tentativo interessante da parte di un’associazione di professionisti di considerare le criticità dei sistemi autonomi e intelligenti è il documento a cura della IEEE dal titolo *Ethically Aligned Design: Prioritizing Human Wellbeing with Autonomous and Intelligent Systems* (IEEE, 2019). Si tratta di un documento esplicitamente rivolto ai progettisti di tali sistemi affinché operino in accordo con i valori umani universali. I principi, le linee guida e le raccomandazioni che vengono presentati hanno l’esplicito obiettivo di mettere gli esseri umani al centro della progettazione di tali sistemi, per l’impiego dei quali è necessario recuperare un pieno rapporto di fiducia con gli utenti. Il documento in questione non si limita, infatti, a sottolineare come evitare possibili effetti negativi, ma cerca di promuovere una forma di sviluppo tecnologico che sia a beneficio dell’umanità intera e sostenibile per il pianeta. Un punto molto importante riguarda l’appello all’autoriflessione a cui il documento invita la comunità dei tecnologi, auspicando che la progettazione ingegneristico-tecnologica non sia più considerata una questione puramente tecnica, ma tenga conto, fin dalle sue fasi iniziali, di aspetti più ampi, come per esempio l’impatto sociale. Non è un caso che nell’introduzione l’*eudaimonia* aristotelica, intesa come forma di benessere a livello individuale e collettivo, sia esplicitamente citata come l’obiettivo dello sviluppo tecnologico. Perché questa forma di benessere non sia soltanto intesa in senso occidentale si fa menzione anche ad altre tradizioni che, seppure in modo diverso, condividono la stessa visione, dalle filosofie orientali alle pratiche di saggezza africana.

Il documento della IEEE è suddiviso in 11 capitoli che coprono diversi temi: dai princìpi generali, ai metodi per guidare la ricerca e la progettazione etiche, fino alle politiche da realizzare e agli aspetti legali da tenere in considerazione. Le analisi e le conseguenti raccomandazioni comprendono sia aspetti valoriali sia aspetti implementativi. Si tratta quindi di un documento che ha caratteristiche aspirazionali, ossia delinea ciò che si vorrebbe fosse realizzato, ma anche un impianto pratico che si concentra su cosa la comunità tecnico-scientifica dovrebbe fare per far avanzare positivamente l’intera società.

Al di là di una dettagliata analisi del documento – lungo quasi trecento pagine e, in alcuni casi, molto impegnativo proprio per la visione globale che sta alla sua a base – ci preme qui sottolineare alcuni aspetti rilevanti, soprattutto in funzione della discussione portata avanti in questo capitolo. In primo luogo questo documento, da cui sono sorte una serie di iniziative[[66]](#footnote-66), è stato creato da più di seicento esperti provenienti da tutto il mondo che sono stati progressivamente reclutati attraverso un processo dal basso, via via che le diverse bozze del documento circolavano per essere commentate e riviste. Questo processo di stesura e di reclutamento è assai diverso da come solitamente codici etici e di comportamento vengono redatti: essi sono perlopiù il risultato del lavoro di un ristretto numero di esperti che, solo in un secondo momento e neppure in tutti i casi, sottopongono i loro documenti a una consultazione pubblica. Questo aspetto può essere visto come un punto di forza – nel suo essere inclusivo – ma anche come un punto di debolezza che si riflette nella lunghezza del documento, a cui già prima abbiamo fatto menzione, e nella mancanza di un focus specifico che sia immediatamente chiaro a chi lo vuole utilizzare.

Un altro aspetto rilevante è che tale documento è stato pensato fin dall’inizio in funzione dello sviluppo di standard per la certificazione etica di sistemi autonomi e intelligenti[[67]](#footnote-67). Lo sviluppo e la commercializzazione di standard per la progettazione ingegneristica costituisce infatti una delle principali attività della IEEE. Questo programma di certificazione si occupa in particolare di sviluppare metriche e processi per realizzare una metodologia che certifichi la trasparenza, l’*accountability* e l’assenza di pregiudizi negli algoritmi sviluppati. Quest’ultimo aspetto mette in luce uno degli elementi critici di cui ci occupiamo alla fine di questo capitolo, ossia la difficoltà, se non l’impossibilità, di codificare l’etica. Nel caso del documento dell’IEEE non solo l’etica è codificata, in quanto espressa in principi e raccomandazioni, ma è anche in qualche modo operazionalizzata, ossia concretamente tradotta in precise metriche che i progettisti devono utilizzare nelle loro attività.

1. Orientamenti etici per un’IA affidabile

Come più volte emerso nei capitoli precedenti, da un punto di vista etico è cruciale che si instauri un rapporto di fiducia tra le parti sociali coinvolte nei processi di innovazione tecnologica. Stimolare e mantenere un clima di fiducia è possibile, però, solo con il dovuto sostegno istituzionale a sforzi di regolamentazione che definiscano i valori da rispettare, i rischi da mitigare e le strategie da adottare (Winfield, Jirotka, 2018). Su ciò hanno concordato i 51 membri dello *High Level Expert Group on AI*, un gruppo di esperti incaricati dalla Commissione Europea di produrre un documento di raccomandazioni per uno sviluppo eticamente e socialmente sostenibile dell’IA in Europa.

A seguito di numerose discussioni interne e consultazioni aperte alla cittadinanza, nell’aprile 2019 il gruppo ha presentato il documento *Orientamenti etici per un’IA affidabile* (AI HLEG, 2019). L’approccio a più livelli qui adottato è un esempio significativo di come si persegue, a livello istituzionale, l’integrazione di preoccupazioni etiche e sociali ai vari interessi – scientifici, industriali, finanziari, commerciali, politici – coinvolti nel processo di innovazione tecnologica. Il testo ha lo scopo di offrire una guida teorica e pratica per instaurare un processo di sviluppo in cui i benefici siano promossi, i rischi mitigati, e la loro distribuzione sia equa. Al centro sta la convinzione che un’IA allineata a valori morali possa dare vita ad un settore solido e resiliente capace di attrarre talenti e investimenti. Definire quali siano questi valori e elaborare raccomandazioni che guidino le parti alla loro adozione diventa così un compito centrale.

Nella prospettiva di questo documento, come anticipato, la fiducia sociale è la pietra angolare di un approccio ampio che si declina in considerazioni tecniche e regolative. Nello specifico, tre aspetti rendono le tecnologie di IA degne di fiducia. La prima condizione è tecnica e riguarda la robustezza dei sistemi, cioè la loro capacità di soddisfare le aspettative, di gestire in sicurezza eventuali errori o malfunzionamenti e di resistere ad attacchi esterni o manomissioni. La seconda condizione è di carattere etico e consiste nell’allineamento dei sistemi ai valori morali messi in gioco dal loro funzionamento nei relativi contesti d’uso. Infine, la terza componente ha a che vedere con la dimensione legale e riguarda l’ottemperanza alle regolamentazioni applicabili.

Per garantire che le applicazioni di IA siano degne di fiducia bisogna innanzitutto determinare quali siano i valori morali più importanti in relazione al loro sviluppo e uso. Molte energie sono state dedicate negli ultimi anni a tale compito. Ne sono risultati numerosissimi documenti, provenienti da diverse aree geografiche, che condividono lo scopo di sottolineare la rilevanza dell’impatto etico della tecnologia, di proporre un quadro valoriale adeguato e di tracciare raccomandazioni per gli attori interessati. Sono varie anche le comunità coinvolte: organizzazioni private (come i princìpi della *Partnership on AI*[[68]](#footnote-68)), università (si veda la Dichiarazione di Montreal[[69]](#footnote-69)), centri o gruppi di ricerca (come nei casi dei Princìpi di Asilomar[[70]](#footnote-70), di AI4People[[71]](#footnote-71) o dei Princìpi di Beijing[[72]](#footnote-72)), fondazioni culturali (ad esempio, lo Statuto Etico e Giuridico dell’IA[[73]](#footnote-73)) e commissioni governative (come il cosiddetto Report Villani in Francia[[74]](#footnote-74), *AI in the UK* in Inghilterra[[75]](#footnote-75), o il Libro bianco sull’IA in Italia[[76]](#footnote-76)). I valori via via emersi sono altrettanto numerosi – dignità umana, equità, giustizia, uguaglianza, non-discriminazione, riservatezza, sostenibilità, autodeterminazione, inclusività, sicurezza e così via – tanto che da più parti è stata avvertita l’esigenza di uno studio sistematico che sottolineasse eventuali convergenze e divergenze (Floridi, Cowls, 2019; Jobin *et al*., 2019)[[77]](#footnote-77).

Basandosi su questi studi e sui documenti fondamentali dell’Unione Europea, i quali mettono al centro i diritti umani e il benessere individuale e collettivo, il gruppo ha proposto una lista di quattro valori basilari: 1. Rispetto dell’autonomia umana, 2. Prevenzione dei danni, 3. Equità, 4. Esplicabilità. I quattro princìpi compongono il nucleo normativo dell’intero approccio e stanno alla base delle numerose raccomandazioni pratiche che ne specificano il significato.

Una volta definito il quadro etico di riferimento è necessario muovere dal piano astratto dei princìpi generali ai diversi contesti pratici in cui le tecnologie sono effettivamente sviluppate, prodotte, vendute, usate e regolate. Rendere operativi concetti tanto generali, per quanto significativi, non è mai semplice e richiede una cura particolare (Morley *et al*., 2019). Gran parte del documento è dedicata proprio a numerose raccomandazioni pratiche rivolte ai diversi attori sociali, con l’intento di chiarire in che senso i diversi princìpi siano rilevanti in relazione all’IA e quali azioni possano essere intraprese per assicurarne l’adozione lungo l’intero ciclo di vita della tecnologia. A tal fine, gli autori individuano sette obiettivi più specifici a cui dedicare massima attenzione, già incontrati in altri capitoli del nostro volume: 1. Intervento e sorveglianza degli esseri umani, 2. Robustezza tecnica e sicurezza, 3. Riservatezza e *governance* dei dati, 4. Trasparenza, 5. Diversità, non discriminazione e equità, 6. Benessere sociale e ambientale, 7. *Accountability*.

Per calare i princìpi nella pratica sono necessarie opportune metodologie. Gli autori ne delineano due tipi: tecniche e non tecniche. Sul versante tecnico si esplorano sia metodi di analisi volti a prevedere e correggere impatti etici negativi, sia strategie per integrare valori morali in sede di progettazione – ad esempio, dotare i sistemi di meccanismi *fail safe* che, in caso di arresto forzato dall’esterno, permettano di riprendere le attività in modo sicuro (*security-by-design*) o di specifici accorgimenti volti a evitare la diffusione di dati sensibili (*privacy-by-design*). In più, si propongono indicazioni su come elaborare e testare architetture di IA affidabili assicurando trasparenza, tracciabilità e esplicabilità, così da facilitare la valutazione del prodotto. Infine, si forniscono suggerimenti su come impostare strategie di comunicazione che supportino la trasmissione delle informazioni più opportune riguardo la tecnologia, le sue funzioni e i rischi associati. Metodologie non tecniche includono invece la formulazione e applicazione di politiche e meccanismi legali per la distribuzione equa delle responsabilità e delle compensazioni, nonché il supporto a sforzi di standardizzazione e certificazione, alla costituzione di comitati e codici etici professionali, e ad attività volte a disseminare la coscienza dei rischi etici nei gruppi coinvolti. Il documento si chiude con una lista di controllo per la valutazione dei sistemi di IA e alcuni esempi concreti di opportunità e rischi, ulteriori strumenti volti a supportare la messa in pratica dei princìpi.

Proprio sulla messa in pratica dei principi e delle raccomandazioni è ora concentrata l’attenzione a livello europeo. Si tratta di un passo estremamente delicato, dalla cui attuazione dipende in gran parte il successo dell’intera iniziativa. Nuovi progetti finanziati dall’Unione Europea, come ad esempio AI4EU[[78]](#footnote-78), si occupano di portare avanti il lavoro svolto, coinvolgendo il mondo aziendale e civile nel processo di allineamento dello sviluppo tecnologico ai valori e alle linee guida presentate nel documento. Tra gli obiettivi del progetto figura l’istituzione di un *Observatory on Society and AI* con lo scopo di raccogliere e diffondere informazioni, organizzare iniziative, valutare applicazioni innovative e promuovere la diffusione del dibattito etico in tutti i paesi dell’Unione.

1. Critiche, limiti e senso dell’approccio

I diversi sforzi volti a stilare princìpi e raccomandazioni servono a delineare le responsabilità dei professionisti e a influenzare i futuri sviluppi delle tecnologie digitali, costituendo così un buon punto di partenza per prendere coscienza delle sfide morali in gioco ed esserne all’altezza. Nel corso del tempo, tuttavia, a questo approccio sono state mosse numerose critiche (Floridi 2019). Passiamo ora in rassegna le difficoltà con cui si misura il tentativo di codificare l’etica per promuoverne la diffusione in ambito professionale e tecnologico. Una comprensione più dettagliata dei suoi limiti aiuterà a valutarne meglio il senso e l’effettiva utilità.

* 1. Window dressing e ethical washing

La prima difficoltà da considerare riguarda il cosiddetto *window dressing* o *ethical washing*. Entrambe le espressioni si riferiscono all’esibizione di un impegno etico di sola facciata per ragioni che hanno a che vedere più con l’immagine sociale dell’organizzazione coinvolta che con una sincera convinzione morale. Il problema è evidente nel caso di aziende private che tramite l’adozione di codici aziendali, la formazione di comitati interni, l’assunzione di “eticisti” e l’organizzazione di eventi possono sperare di convincere il pubblico e battere la concorrenza senza dare avvio ad un reale processo di riforma. Anche organizzazioni professionali e politiche sono però esposte allo stesso rischio. Infatti, convincere il pubblico della propria serietà etica ha chiari effetti positivi sulla reputazione di cui un’organizzazione gode e, dunque, sul supporto che può sperare di ottenere nel panorama nazionale e internazionale.

Sono molti gli aspetti intorno a cui si concentra la polemica. Ad esempio, la vaghezza in un certo modo intrinseca ai codici – in particolare a quelli aspirazionali, in cui i valori presentati sono visti più come ideali regolativi che come concrete linee guida per il comportamento – offre il fianco ad attacchi che ne sottolineano l’inconsistenza. Anche la questione della sanzione di comportamenti in disaccordo con i codici è spesso oggetto di contesa. Come abbiamo visto nel caso del codice ACM, ci possono essere diverse violazioni e quindi diverse possibili azioni, oltre all’esclusione dall’associazione. Inoltre, nel caso dei codici aziendali le possibilità di sanzionare sono maggiori rispetto a quelle dei codici professionali. Rimane però il dubbio sull’effettiva attuazione di queste misure e sulla loro efficacia come deterrenti. Il fatto che si tratti poi di ambiti in cui la componente economica gioca un ruolo fondamentale non può che acuire il sospetto circa gli interessi che potrebbero nascondersi alle spalle di attività ed espressioni di intenti centrate sull’etica. L’emanazione di documenti come quelli presentati può sempre nascondere l’intento di dare un’impressione favorevole senza voler prendere le dovute misure sostanziali. Un sospetto acuito anche dalla spettacolarizzazione delle iniziative etiche intraprese in ambito privato e pubblico, le quali sono spesso accompagnate da campagne mediatiche senza che siano chiare le effettive ricadute pratiche e organizzative.

Sarebbe però un errore concludere che gli sforzi che sfociano nella stesura di linee guida e codici etici siano destinati al fallimento. Prendere coscienza degli abusi a cui l’etica professionale e istituzionale è esposta significa anche capire come contrastarli. Se non può essere presupposta, in generale, la buona fede di chi ad ogni livello traccia princìpi, linee guida e raccomandazioni, non resta che organizzare azioni e meccanismi di controllo in grado non solo di vigilare sull’adozione dei princìpi, ma anche di promuovere l’impegno reale su quello fittizio. Spetta alla società in senso lato – ricercatori, progettisti, produttori, utenti, osservatori privati e pubblici, istituzioni e così via – esercitare spirito critico e attento scrutinio così da testare l’effettiva consistenza dell’impegno mostrato dalle parti.

Bisogna inoltre stare attenti a non addossare ai documenti etici funzioni che da soli non possono essere in grado di svolgere. Ad esempio, possiamo affermare che le sanzioni non sono l’obiettivo primario né dei codici professionali, né di quelli aziendali, né dei documenti di indirizzo. Codici e documenti etici hanno fondamentalmente uno scopo diverso: stimolare comportamenti responsabili e la capacità di giudizio morale dei professionisti, dei dirigenti, dei dipendenti e degli utenti ai quali sono rivolti. Badare all’effettiva applicazione di valori e raccomandazioni non è un risultato che un documento, da solo, può ottenere. La serietà dello sforzo etico si misura in funzione delle attività che ne stimolano l’adozione e vigilano sul suo rispetto.

* 1. Problemi di applicazione

Un altro problema riguarda i conflitti tra le raccomandazioni che possono essere espresse da un codice (o da diversi documenti) rispetto al comportamento da tenersi in una determinata situazione, come per esempio quando la fedeltà dei professionisti dipendenti è descritta come un obbligo da ottemperare nei confronti tanto del datore di lavoro quanto della professione. In modo simile, documenti incentrati su valori e linee guida possono dare adito a casi di incertezza, dove la stessa situazione sembra mettere in gioco valori ugualmente importanti, ma che si limitano a vicenda. Un esempio può essere la complessa dimensione della privacy, dove conflitti tra valori quali protezione dell’anonimato e della riservatezza, benessere individuale e comune, sicurezza sociale e autonomia sono comuni. Oltre ai conflitti e alle tensioni interne, anche la vaghezza delle raccomandazioni è una delle critiche spesso rivolte tanto ai codici professionali quanto ai documenti di indirizzo, i quali esprimerebbero valori astratti sui quali sostanzialmente tutti possono essere d’accordo, senza però specificare in maniera chiara come un determinato professionista dovrebbe comportarsi in una determinata situazione. Ciò ha a che fare anche con un’altra obiezione, ossia con la difficoltà, e in alcuni casi l’impossibilità, di mettere in pratica quanto prescritto da un codice, come può essere in generale per tutte quelle azioni che sembrano andare contro gli interessi economici di un professionista o di un dipendente. Ad esempio, attenersi alle molte raccomandazioni proposte in *Orientamenti* può richiedere tempi, risorse e soluzioni che difficilmente collimano con altri interessi commerciali o industriali, con il rischio che le esigenze derivanti da questi ultimi vanifichino ogni reale cambiamento etico.

Contraddittorietà, conflitti, vaghezza e difficoltà pratiche nel tradurre parole in azioni sono limiti effettivi dei documenti che abbiamo preso in considerazione. Ne consegue la loro inutilità? Piuttosto, è necessario regolare le nostre aspettative. Tracciare un quadro valoriale non significa fornire una procedura in grado di risolvere ogni contrasto in modo inappellabile o di produrre soluzioni definitive a tutti i problemi possibili. Abbiamo già avuto modo di ricordare che il ragionamento etico non è riducibile ad un algoritmo che genera risposte univoche e corrette a ogni problema concreto a partire da valori, principi e raccomandazioni di carattere generale. Al contrario, richiede una discussione partecipata e razionale sia, in generale, sui valori rilevanti sia, in particolare, sui casi specifici. In verità, i quadri valoriali sono mappe, carte che possano servire ad orientare i comportamenti dei singoli individui alle prese con l’IA e altre tecnologie. Ed è solo al livello concreto – al livello della progettazione, dello sviluppo, dell’uso, e quindi dell’impegno personale – che l’etica può trasformarsi da parola scritta a azione vera e propria. I quadri generali, se ben intesi e ben usati, hanno il solo scopo di offrire a chi sente l’esigenza di confrontarsi con il problema una serie di linee guida per affrontarlo – non per scavalcare l’atto del giudizio, quindi, ma per innescarlo e stimolarlo. Tutto ciò può servire come cartina al tornasole per valutare sia la consapevolezza di chi propone documenti sui princìpi etici dell’IA sia l’uso che se ne fa nelle diverse sedi di discussione.

* 1. Pluralismo dei valori e autonomia morale

Infine, una critica comune riguarda l’idea che l’etica non possa essere codificata: i codici di condotta e i documenti di indirizzo conterrebbero delle prescrizioni troppo rigide e specifiche, che sarebbero in contraddizione con l’idea secondo cui l’etica sia irriducibile a un insieme di regole e procedure. A supporto di questa critica sono usati diversi argomenti: dal fatto che l’etica richiede la capacità di giudizio morale e non la cieca adozione di regole, per arrivare a discutere della non obbligatorietà delle indicazioni contenute nei codici e, quindi, del loro non essere vincolanti dal punto di vista morale.

A questa obiezione, che fa leva sulla dimensione individuale e privata dell’etica, si collegano anche osservazioni che insistono invece sul pluralismo dei valori in relazione ai contesti sociali e culturali, ovvero sul fatto che le convinzioni razionali circa ciò che è importante e che deve essere tutelato variano da gruppo sociale a gruppo sociale, da cultura a cultura. L’idea stessa di un quadro valoriale che valga per ogni applicazione di IA, in ogni tempo e luogo, è una chimera. Sebbene sia apprezzabile una certa invariabilità transculturale, le convinzioni ragionevoli circa ciò che è buono e degno di essere perseguito variano in numerose circostanze al variare delle coordinate geografiche e temporali. Non solo: anche all’interno di una medesima cultura le opinioni individuali sul bene e sul valore variano significativamente. Diversi esseri umani appartenenti a diverse culture in diverse regioni dello spazio e del tempo nutrono opinioni diverse su ciò che è bene e male, desiderabile o odioso, degno di lode o di biasimo, senza che allo stesso tempo sia sempre possibile squalificare l’altrui convinzione come irrazionale o inaccettabile.

I problemi non mancano, ad esempio, se si limita lo sguardo alla cultura occidentale. Nella nostra tradizione, il diritto dell’individuo ad autodeterminarsi – a farsi un’idea, in autonomia, di cosa sia il bene e il valore – sta alla base del nostro concepirci come persone in grado di compiere scelte considerando ciò che è bene e male; scelte, quindi, di cui essere responsabili. Che ne è dell’autonomia quando sono altri a decidere che cosa è bene e che cosa è male nell’uso dell’IA? E chi mai può arrogarsi il diritto di prendere questa decisione al posto del professionista o dell’utente? Si profila qui anche il rischio della tecnocrazia – di cui già è stato discusso in (ROBOTICA SOCIALE pp…) – soprattutto nel caso in cui valori etici vengano implementati *by-design* nelle tecnologie in modi opachi, ovvero senza che l’utente ne abbia piena coscienza o ne sia adeguatamente informato e senza che la società in senso lato sia coinvolta nei processi decisionali rilevanti. In definitiva, lo stesso concetto di una codificazione che esiga di essere adottata sempre e ovunque sembra intrinsecamente inadeguata e molti dubitano del fatto che lo sforzo di sistematizzare i valori a cui adeguare l’uso dell’IA sia compatibile con la pluralità delle ragionevoli opinioni umane sul tema.

Simili sospetti sono senz’altro parzialmente condivisibili. Uno dei compiti è far sì che non diventino motivo di scontro o disaffezione per l’impegno morale. D’altronde, non è onesto trasformare un quadro valoriale in una tavola dei comandamenti. Tanto nei codici etici quanto nei documenti di indirizzo non si vuole fissare una volta per tutte ciò che è buono o malvagio, giusto o sbagliato, lecito o illecito. Si tratta, al contrario, di stabilire una base di partenza condivisibile e chiara per definire i caratteri morali di una professione o di un’azione sociale, la cui declinazione nei diversi contesti culturali è parte integrante e costitutiva. Spetta ad ogni gruppo sociale passare in rassegna le proposte, valutarle razionalmente e decidere, altrettanto razionalmente, se dare il proprio assenso e impegnarsi a calare valori o raccomandazioni in pratica, oppure continuare la discussione alla ricerca di nuovi equilibri.

Sicuramente, data la dimensione pubblica delle attività normate, ha poco senso affidare esclusivamente al singolo individuo il compito di curarsi dell’aspetto morale della professione e dello sviluppo tecnologico. Date le condizioni di partenza, solo un lavoro comune può ottenere risultati apprezzabili. Per questo motivo è necessario passare dalla dimensione individuale a quella sociale e proporre approcci etici che rappresentino per quanto possibile i valori e le intenzioni condivise delle parti coinvolte. Un risultato, quest’ultimo, che può essere ottenuto solo a livello istituzionale tramite l’esercizio della ragione pubblica, il dibattito scientifico e la rappresentanza, e che tramite questi è sempre passibile di revisioni e modifiche. Lo scopo non è di produrre un perfetto sistema di valori o di raccomandazioni che scavalchi l’autonomia morale del singolo individuo e della cultura a cui appartiene; ma di fornire strumenti utili a stimolare il giudizio, guidare l’impegno e organizzare le attività di tutte le parti coinvolte. Se ben inteso, lo sforzo istituzionale va tutto a beneficio del singolo, il quale nell’esercizio imprescindibile del suo impegno etico può contare sul supporto offerto nei documenti. Il tentativo di guidare lo svolgimento delle professioni e l’innovazione tecnologica anche alla luce di valori morali non è riducibile né alla scrittura di documenti né alle riflessioni individuali o di gruppo su situazioni concrete e casi specifici, ma comprende entrambi i momenti e vive del movimento che li connette.

Capitolo 14

Orizzonti

Fabio Fossa, Viola Schiaffonati, Guglielmo Tamburrini

1. Introduzione

I sistemi autonomi e intelligenti sono tecnologie pervasive: si stanno diffondendo velocemente in ambiti innumerevoli della vita sociale e personale. E sono anche tecnologie dirompenti (*disruptive* in inglese), poiché rimodellano in profondità strutture sociali o economiche preesistenti. È perciò irrealistico cercare di dare conto in maniera esaustiva di tutte le questioni etiche che il loro sviluppo solleva o può sollevare nel futuro imminente. Anche nel progettare un volume come questo – che pure intende fornire una panoramica di ampio respiro su riflessioni e dibattiti etici in corso – è stato necessario operare una scelta di temi e problemi, allo scopo di trovare un ragionevole punto di equilibrio tra esplorazione orizzontale delle questioni etiche in gioco e un loro approfondimento verticale. Alcune questioni notevoli che non sono state qui esaminate riguardano le implicazioni etiche dell’automazione di attività lavorative abilitata dalle tecnologie dell’IA e della robotica; il potenziamento delle capacità mentali e motorie degli esseri umani che potrebbe essere indotto dalle stesse tecnologie; la cosiddetta *machine ethics*, ovverosia il programma di ricerca che si propone di dotare le macchine di qualche competenza morale; l’eventuale attribuzione di diritti e doveri a sistemi autonomi e intelligenti. Ci limitiamo qui a illustrare brevemente questi temi e problemi, fornendo qualche spunto ulteriore di lettura e qualche indicazione sulla possibilità di applicare anche in questi casi le impostazioni metodologiche utilizzate nei capitoli precedenti. Per concludere, accenniamo ai contributi che l’IA e la robotica possono fornire per affrontare sfide etiche primarie del nostro tempo - contrasto delle pandemie, mitigazione del riscaldamento climatico, difesa della democrazia e della pace internazionale - insieme ai nuovi rischi che esse possono porre in tali contesti.

1. Automazione e lavoro

Le questioni etiche che i sistemi autonomi e intelligenti sollevano a proposito del lavoro umano sono casi particolari di due questioni più generali, che si sono presentate a ogni passaggio significativo nella storia dell’automazione dei processi di produzione di beni e servizi. La prima di tali questioni si inquadra nella cornice generale dei problemi di giustizia distributiva, e cioè di che cosa dobbiamo gli uni agli altri, eventualmente con la mediazione delle istituzioni, in base a un dovere morale, piuttosto che in base a una qualche idea di carità o altruismo. La seconda questione riguarda la salvaguardia di condizioni di lavoro dignitose e soddisfacenti per ciascuno.

1. Quali sono e come devono essere affrontati i problemi di giustizia distributiva che l’automazione induce cambiando la disponibilità e la remunerazione del lavoro umano? In particolare, come bisogna affrontare i problemi di disoccupazione tecnologica di breve o di lungo periodo, insieme alle privazioni di beni primari e alle disuguaglianze moralmente inaccettabili che possono colpire i lavoratori sostituiti da macchine nei processi di produzione?
2. Come si deve salvaguardare e promuovere la dignità dei lavoratori e la qualità della loro esperienza lavorativa nei processi produttivi che le nuove tecnologie rimodellano?

Questi interrogativi etici di fondo, i quali calano nel contesto delle attività lavorative il valore della dignità umana e il problema di una distribuzione equa della ricchezza prodotta, hanno attraversato la storia dell’automazione, a partire dall’introduzione dei telai meccanici della prima rivoluzione industriale del XVIII secolo fino alle attuali applicazioni dell’IA e della robotica. Le stesse coordinate etiche sono significativamente incorporate nella lista di obiettivi dell’Agenda 2030 dell’ONU per lo sviluppo sostenibile (ONU, 2015): promozione del pieno impiego e di un lavoro dignitoso per tutti (obiettivo 8); riduzione delle disuguaglianze economiche e sociali (obiettivo 10). Per quanto le questioni etiche in gioco fin dagli albori dell’automazione industriale non siano cambiate nel tempo, è opportuno soffermarsi brevemente sulle novità che le tecnologie dell’IA e della robotica apportano da una prospettiva economica e sociologica, determinando così un nuovo contesto interpretativo per entrambe le questioni (1) e (2).

L’IA e la robotica estendono le possibilità di sostituzione del lavoro umano a compiti che si basano sempre di più sull’esercizio di facoltà intellettuali, di giudizio e di coordinamento intelligente tra percezione e azione. E forniscono inoltre nuovi strumenti di controllo e distribuzione delle attività lavorative. Il fondatore della cibernetica Norbert Wiener mise a fuoco entrambi gli elementi di novità già agli albori dell’IA e della robotica, intorno alla metà del secolo scorso, allorché i primi calcolatori digitali si trovavano ancora nelle fasi iniziali di sviluppo, e i primi robot industriali avrebbero fatto la loro comparsa solo un decennio più tardi. Le sue ipotesi sullo sviluppo dell’automazione erano sostanzialmente basate su modelli teorici che aveva contribuito a elaborare: modelli del controllo adattivo del comportamento delle macchine e modelli dei processi algoritmici di calcolo. Sugli scenari di produzione di beni e servizi prefigurati partendo da tali conoscenze teoriche, Wiener elaborò una riflessione etica originale.

Ricollegandosi alla questione (1), Wiener presagì sia la possibilità di progettare fabbriche senza lavoratori sia l’avanzare di una disoccupazione tecnologica derivante dall’automazione di capacità cognitive e di giudizio. Egli sollecitò i sindacati statunitensi – con una lettera indirizzata nel 1949 al presidente del sindacato dei lavoratori dell’industria automobilistica (*Union of Automobile Workers*) – a interessarsi «alla minaccia assai pressante della sostituzione di massa della manodopera con macchine, non a livello di energia bensì di giudizio, in modo da essere pronti a formulare una politica su questo problema» (trad. it. in Noble, 1993, p. 152). E ricollegandosi alla questione (2) della protezione della dignità dei lavoratori, sottolineò che «qualunque lavoro in concorrenza con il lavoro schiavistico, che gli schiavi siano umani o meccanici, deve accettare le condizioni lavorative del lavoro schiavistico» (ivi, p. 153).

In definitiva, Wiener rielaborò le due questioni fondamentali che caratterizzano fin dagli inizi il dibattito etico sull’automazione, immergendole nel nascente contesto dell’automazione di attività e profili lavorativi che richiedono l’esercizio di capacità di giudizio e di coordinamento intelligente tra percezione e azione. Wiener ripropose in questo contesto anche le questioni relative ai rapporti di potere collegati all’automazione sottolineando come le macchine possano essere utilizzate come strumento di sfruttamento e dominio sui lavoratori, a vantaggio di pochi e a detrimento del benessere collettivo (Tamburrini, 2020).

La previsione più fosca tra quelle avanzate da Wiener - ossia la sostituzione di massa dei lavoratori con macchine - non si è materializzata nel corso dei decenni successivi. Ma il dibattito su questa possibilità è stato ripreso in termini analoghi nel primo scorcio di questo secolo. Nel 2013 Carl Benedikt Frey e Michael Osborne hanno ipotizzato che circa il 50% delle attività lavorative svolte in paesi tecnologicamente avanzati è a rischio di sostituzione da parte di sistemi informatici o robotici. Sarebbero maggiormente a rischio le attività di tipo impiegatizio caratterizzate dallo svolgimento di compiti cognitivi di routine. Frey e Osborne hanno identificato nell’automazione di questo tipo di lavoro una causa centrale della tendenza, rilevata negli ultimi decenni, di una polarizzazione crescente del mercato del lavoro tra occupazioni manuali a reddito basso da un lato e occupazioni intellettualmente più creative e a reddito alto dall’altro lato. (Frey, Osborne, 2013).

Previsioni meno allarmanti sono state elaborate in vari studi successivi. Le stime fornite nel rapporto 2019 sul futuro del lavoro stilato dall’Organizzazione per la cooperazione e lo sviluppo economico (OCSE, 2019) parlano di un 14% di lavori esistenti che potrebbero sparire a causa dell’automazione nel giro di 15-20 anni e di un altro 30% di lavori che subirà una trasformazione radicale. In relazione alla questione (1), il rapporto dell’OCSE sollecita interventi di sostegno, riqualificazione e reinserimento, per rispondere alle esigenze pressanti che riguardano la vita di coloro che sono espulsi dal mondo del lavoro per un arco di tempo più o meno lungo a causa delle nuove forme di automazione.

Il rapporto dell’OCSE sposa anche la previsione di uno svuotamento parziale della fascia di occupazioni caratterizzate dallo svolgimento di compiti cognitivi di routine, accompagnato dalla creazione di nuovi lavori precari e a reddito più basso. Partendo da una tale stima di polarizzazione del mercato del lavoro, il segretario generale dell’OCSE evidenzia una serie di problemi che sono strettamente legati alla questione 2: «Molti sono bloccati in situazioni lavorative precarie con salari esigui, con accesso limitato o inesistente a forme di protezione sociale, di formazione permanente e di contrattazione collettiva» (trad. degli autori).

Alla creazione di nuove forme di precariato contribuiscono le aziende coinvolte nella gestione delle piattaforme digitali e della cosiddetta economia dei lavoretti (*gig economy*). Gli strumenti di controllo e distribuzione delle attività lavorative messe in atto mediante le piattaforme digitali comportano nuove minacce alla dignità dei lavoratori. I sistemi dell’IA e della robotica sono utilizzati diffusamente per offrire e governare su scala globale forme atipiche, precarie e parcellizzate di lavoro, spesso impropriamente presentate come lavoro indipendente (Casilli, 2020).

Un altro aspetto dell’economia digitale al quale riferire le questioni etiche di fondo (1) e (2) riguarda l’estrazione di valore economico dalle attività compiute in rete dagli utenti del web in generale, e utenti delle piattaforme digitali in particolare. Antonio Casilli introduce a tale riguardo un’interessante distinzione tra tipologie di valore estratto: valore di qualificazione, in quanto gli utenti organizzano l’informazione lasciando commenti o valutando beni e servizi offerti dalle piattaforme; valore di monetizzazione, attraverso la cessione ad altri attori economici di dati generati dagli utenti; e valore di automazione, derivante dall’uso dei dati generati dagli utenti per addestrare sistemi basati sull’apprendimento automatico, i quali servono a regime per formulare ipotesi plausibili su profili, propensioni e comportamenti attesi degli stessi utenti (ivi, p. 22). Con l’economia delle piattaforme e dell’informazione nasce la questione di valutare se le attività online degli utenti delle piattaforme debbano essere considerate come una nuova forma di lavoro e se questi ultimi debbano beneficiare in modo diretto o indiretto del valore estratto dalle loro attività in rete. Per approfondimenti sui dibattiti in corso su questo tema, i lettori interessati sono rimandati a (Casilli, 2020) e alla ricca bibliografia in esso contenuta.

1. Potenziamento informatico e robotico

Per potenziamento si intende, in generale, ogni intervento sul corpo umano volto a migliorare le capacità fisiche o mentali al di là di quanto sia necessario per cercare di ristabilire o sostenere lo stato di salute. È stata evocata a più riprese la possibilità di utilizzare le tecnologie dell’IA e della robotica per scopi di potenziamento. Basandosi sullo sviluppo di dispositivi di calcolo sempre più piccoli e potenti, il pioniere dell’IA Marvin Minsky ha immaginato un futuro nel quale sarà possibile sostituire parti del cervello malfunzionanti o di potenziarne le funzioni di elaborazione e di memorizzazione con tecnologie informatiche (Minsky, 1994). Ray Kurzweil ha profetizzato che un giorno si potrà disporre il caricamento della propria mente su un dispositivo informatico e di prolungarne indefinitamente la sussistenza (Kurzweil, 2005). Il movimento postumanista o transumanista, all’interno del quale trova eco anche la profezia di Kurzweil, sposa la prospettiva di utilizzare protesi informatiche e altre tecnologie per superare i limiti della condizione umana e inoltrarsi verso un’ipotetica condizione postumana, caratterizzata dal conseguimento di capacità fisiche e mentali inaccessibili alla specie umana e da nuove forme di simbiosi tra esseri umani e macchine (Bostrom, 2005).

Esempi concreti di esperimenti di potenziamento che coinvolgono tecnologie informatiche o robotiche sono pochi e rudimentali rispetto agli scenari futuribili appena evocati. Una vasta eco ebbe un esperimento bionico compiuto nel 2003 su un soggetto sano. Il ricercatore informatico e cibernetico Kevin Warwick si sottopose volontariamente, attraverso un’operazione chirurgica, all’impianto di microelettrodi all’interno del nervo mediano del braccio sinistro. L’esperimento intendeva fornire una prova di operabilità bidirezionale tra il sistema nervoso periferico di un soggetto sano e un dispositivo robotico, sia per controllare il dispositivo con segnali elettrici trasmessi dal sistema nervoso sia per acquisire da esso segnali di feedback sensoriale (Warwick *et al.*, 2003). L’impianto fu successivamente rimosso anche in considerazione della possibilità di deterioramento funzionale e di infezioni da esso indotte. L’esperimento bionico non ebbe un seguito di rilievo rispetto alla questione del potenziamento effettivo delle capacità fisiche o mentali dell’essere umano.

Anche nel caso delle interfacce cervello-computer (in breve BCI, dall’inglese *brain-computer interfaces*) sono stati ipotizzati usi non terapeutici, soprattutto a fini di potenziamento funzionale delle capacità di comunicazione e controllo. I sistemi BCI sono detti d’ingresso, se producono segnali che afferiscono al sistema nervoso, e di uscita se registrano segnali prodotti dal sistema nervoso centrale dell’utente e li elaborano per controllare un sistema informatico o robotico. BCI di questo secondo tipo sono stati utilizzati per controllare manipolatori robotici, sedie a ruote robotiche e programmi di videoscrittura. Essi consentono di desumere un intento dell’utente (selezionare un’azione di presa di un braccio robotico, una direzione di movimento della sedia a ruote, una lettera dalla videotastiera) in base all’elaborazione dei correlati neurali delle sue attività di pensiero, provvedendo poi a pianificare in ogni dettaglio e a eseguire l’azione corrispondente. I sistemi BCI in uscita possono essere invasivi o non invasivi. I sistemi invasivi utilizzano elettrodi posizionati sulla superficie del cervello o inseriti al suo interno. Per la loro natura invasiva e in funzione di analisi costi-benefici, essi sono stati utilizzati solo da una platea ristretta di utenti, affetti da patologie che ne hanno fortemente ridotto le capacità comunicative e motorie. I sistemi BCI non invasivi sono stati invece utilizzati anche da utenti in buone condizioni di salute. È però importante sottolineare che le attuali limitazioni dei sistemi BCI in uscita e non invasivi (come quelli che elaborano i segnali registrati mediante elettroencefalogramma), soprattutto in termini di velocità e di accuratezza nell’elaborazione dei segnali EEG, non rendono attualmente possibile prefigurarne un uso a fini di potenziamento delle capacità umane (Tamburrini, 2012). Numerose e interessanti sono le questioni etiche legate agli usi terapeutici (Tamburrini, 2009; Mattia, Tamburrini, 2015) e non terapeutici dei sistemi BCI (Ienca, Andorno, 2017), ma nessuna di queste è legata alle questioni del potenziamento e del postumano.

Tra le questioni etiche collegate al potenziamento e a un’ipotetica condizione postumana, ci limitiamo qui a ricordare il problema della persistenza dell’identità personale a seguito degli interventi di potenziamento; il senso di autenticità di un soggetto potenziato e di autorialità rispetto alle scelte e alle azioni che compie; la dissoluzione di caratteristiche fondamentali di un’esistenza genuinamente umana e dei valori ad essa collegati; i problemi di equità che potrebbero derivare dalla difficoltà di accedere al potenziamento a causa di costi elevati o nella competizione per una posizione lavorativa, per esercitare la quale le dotazioni accresciute di un soggetto potenziato conferiscano un chiaro vantaggio (Clarke *et al.*, 2016). Alcune delle principali critiche filosofiche (espresse, per esempio, da Jürgen Habermas, Michael Sandel e Francis Fukuyama) sulle prospettive del potenziamento e di una condizione postumana prendono le mosse da una disamina di queste problematiche etiche (per una approfondita analisi del dibattito, si veda Grion, 2012). Bisogna però tenere presente che uno scenario di potenziamento basato su delle tecnologie dell’IA e della robotica non sembra essere tecnologicamente imminente. Non essendo rintracciabili significativi elementi di supporto negli sviluppi attuali dell’IA, della robotica e della bionica, una discussione meno speculativa può essere vantaggiosamente rinviata al se e al quando degli sviluppi tecnologicamente rilevanti.

1. Etica delle macchine

Un tema controverso che è oggetto di ampio dibattito è la possibilità di dotare i sistemi autonomi e intelligenti di una qualche competenza morale. La questione è al centro di un ambito di ricerche noto come “etica delle macchine” (*machine ethics*: Wallach, Allen, 2009; Anderson, Anderson, 2011). L’etica delle macchine riunisce diverse discipline (informatica, robotica, filosofia, psicologia…) e persegue obiettivi di natura tanto pratica – come la realizzazione di *agenti morali artificiali*, ossia sistemi tecnologici in grado di funzionare secondo valori morali – quanto teorica – come il chiarimento di che cosa si intenda per “agente morale artificiale” e di quali condizioni debbano essere soddisfatte perché un prodotto tecnologico possa essere definito come tale.

La questione degli agenti morali artificiali, per quanto possa sembrare piuttosto astratta, emerge in realtà da un bisogno reale. Secondo i suoi sostenitori, la necessità di dotare gli agenti artificiali di capacità morali deriva direttamente dalla loro autonomia operativa, che come sappiamo rende possibile la delega di compiti complessi a sistemi tecnologici. Questa delega, in alcuni casi, può riguardare compiti che, quando svolti da un agente umano, ne chiamano in causa il giudizio morale. Ad esempio, se si sceglie di delegare l’atto del guidare ad un veicolo autonomo, al contempo vengono delegate anche tutte quelle decisioni morali che stanno alla base di uno stile di guida responsabile e sicuro (per esempio, usare particolare cautela nei pressi di scuole o nel superare ciclisti). Oppure, se si delega ad un algoritmo la raccomandazione di contenuti multimediali a utenti minorenni, si delegano anche tutte quelle decisioni che vertono sulla loro adeguatezza. In casi come questi, l’autonomia di cui godono gli agenti artificiali può portare a conseguenze negative da un punto di vista morale. Nasce così l’esigenza che il funzionamento autonomo dei sistemi tecnologici sia allineato alle aspettative morali che è ragionevole o lecito nutrire a riguardo.

L’etica delle macchine rappresenta una possibile soluzione a questa esigenza (non l’unica: si veda quanto detto ad esempio sul controllo umano significativo in ARMI AUTONOME pp…). L’idea fondamentale consiste nel munire gli agenti artificiali di qualcosa di analogo al giudizio morale umano. Si tratterebbe, cioè, di ideare degli accorgimenti grazie a cui i sistemi, dato un certo input nel contesto di una situazione moralmente rilevante, producano un output sufficientemente simile a quello prodotto nelle medesime circostanze da un agente morale umano o ideale.

Se è vero che la nozione di agente morale artificiale è ormai parte del dibattito, assai meno chiaro è però il suo senso. Le difficoltà nascono dal problema di pensare l’etica delle macchine senza ridurla all’etica degli umani. Da una parte, l’etica degli umani è l’unico paradigma che abbiamo a disposizione per pensare altri tipi di agire morale, compreso quello artificiale. Dall’altra, pensare l’agire morale artificiale sulla falsariga dell’agire morale umano conduce a mettere le due forme sullo stesso piano e ne suggerisce un’identificazione certamente provocatoria, ma debole a livello epistemologico – da cui le aspre critiche che fanno della disciplina un argomento controverso (van Wynsberghe, Robbins, 2017). Tuttavia, le accuse che le sono mosse si applicano spesso solo ad una delle accezioni del concetto, mentre non riguardano altre, più rigorose impostazioni del problema. Per illustrare meglio questo punto, seppur brevemente, prendiamo in considerazione due proposte di classificazione degli agenti morali che hanno lo scopo di includere anche enti artificiali.

Il primo schema è stato proposto da Moor (1995) ed è articolato in quattro classi.[[79]](#footnote-79) Nella prima classe troviamo gli agenti a impatto morale (*ethical impact agents*), ovvero tecnologie la cui mera introduzione sortisce conseguenze di interesse etico. Per quanto importante, si tratta ancora di un livello elementare che non calca la soglia dell’etica delle macchine. Il primo passo si compie invece con gli agenti morali impliciti (*implicit ethical agents*): tecnologie progettate e costruite in modo tale che, durante il loro funzionamento, si adeguino rigidamente a valori morali dati. L’adeguamento della funzione al valore, qui, è predeterminato: i valori sono vincoli che direzionano implicitamente il funzionamento dell’algoritmo verso esiti moralmente apprezzabili. Il gradino successivo della scala è occupato dagli agenti morali espliciti (*explicit ethical agents*) i quali, a differenza dei precedenti, esibiscono la capacità di: a) calcolare esplicitamente diversi scenari di funzionamento alla luce di parametri che corrispondono a valori morali; b) funzionare di conseguenza; e c) fornire informazioni che rendano conto delle computazioni eseguite – il che, con tutte le cautele del caso, costituirebbe un analogo di ciò che per noi umani è la giustificazione. La giustificazione in senso proprio, così come il comportamento morale in senso proprio, si manifestano solo con la classe degli agenti morali pieni (*full moral agents*): esseri dotati di coscienza, intenzionalità e libero arbitrio, capaci di ragionare eticamente e rendere conto delle proprie decisioni.

In modo analogo, Wallach e Allen (2009) propongono di posizionare i diversi agenti morali su un piano cartesiano che ha la qualità dell’autonomia sull’ordinata e la qualità della sensibilità a valori morali sull’ascissa. Sul gradino più basso, dove non si manifestano gradi significativi di alcuna variabile, si situa il livello dell’etica operativa (*operational morality*). La classe include accorgimenti del tutto passivi ed inerti che, però, hanno implicazioni di carattere etico – come le sicure sulle armi da fuoco o, per fare un esempio reso celebre da Latour (1992), i dossi stradali. A gradi elevati di autonomia e sensibilità etica corrispondono gli agenti morali nel pieno senso della parola (*full moral agency*), capaci di responsabilità e degni di fiducia, di cui al momento non sono disponibili esempi tecnologici, ma solo organici – gli esseri umani. Tra i due estremi si estende il dominio dell’etica funzionale (*functional morality*), i cui elementi sono descritti da gradi medi e diversi di autonomia e sensibilità a valori. Secondo gli autori, l’avanzamento tecnologico tende da sé ad agenti artificiali dotati di autonomia sempre maggiore, mentre più complessa è l’implementazione della sensibilità ai valori, che si profila quindi come la missione dell’etica delle macchine.

Entrambe le classificazioni hanno il pregio di impostare un ragionamento che permette di approfondire l’impatto morale delle tecnologie dotate di autonomia operativa. Allo stesso tempo, però, presentano il difetto di insistere eccessivamente sui tratti di continuità tra agire morale artificiale e umano, generando l’impressione che solo una differenza di grado separi i due casi. Sulla base di questa impressione sono state avanzate molte tesi provocatorie. Ad esempio, alcuni sostengono che in futuro saranno le tecnologie di IA a determinare che cosa è buono o malvagio e, in questo modo, sarà raggiunto uno stato di perfezione morale; oppure, che solo i sistemi tecnologici potranno essere pieni agenti morali, in quanto la loro razionalità non sarà inquinata da emozioni e altri elementi spuri che intorbidano la nostra capacità di giudizio; o, infine, che l’essere umano è destinato all’obsolescenza morale, dal momento che non sarà in grado nemmeno di comprendere le teorie etiche elaborate dalle IA del futuro. Simili derive fantascientifiche, per quanto possano risultare interessanti come esperimenti mentali, portano alle estreme conseguenze l’idea per cui ci sia una somiglianza tra agire morale artificiale e umano senza tenere adeguatamente conto delle differenze che distinguono le due forme di agire e che dovrebbero mettere in guardia dall’immaginare il futuro degli agenti morali artificiali come il perfezionamento dell’agire morale umano. Si pensi, ad esempio, alla differenza tra lo svolgere funzioni secondo valori morali stabiliti da altri e l’agire secondo valori morali stabiliti da sé (Fossa, 2018). Una simile differenza, che attualmente distingue l’agire morale artificiale e umano, è un buon punto di partenza per elaborare una teoria dell’agire morale artificiale che si sottragga alle critiche a cui si espongono le derive più fantascientifiche e contribuisca a impostare un discorso sull’etica delle macchine che sia adeguato filosoficamente e fonte di ispirazione per sviluppi tecnologici.

1. Diritti dei robot

Un altro tema controverso del dibattito sull’etica della robotica e dell’intelligenza artificiale è la questione dei diritti dei robot. La domanda fondamentale, a cui si cerca di dare corretta formulazione prima ancora che risposte, chiede se ci siano delle valide ragioni per cui abbia senso riconoscere alcuni diritti ai robot. Si tratta di una domanda che a prima vista può sembrare assurda ma, se affrontata con rigore e senza pregiudizi, rivela una varietà di motivazioni e argomentazioni meritevoli di essere saggiate ulteriormente.

Per orientarsi nell’intrico delle diverse posizioni che sono state sostenute su questo problema è utile rivolgersi al libro di David Gunkel *Robot Rights* (Gunkel, 2018). Il libro offre un’ampia panoramica del dibattito circa i diritti dei robot allo scopo di criticarne i presupposti e cercare un modo migliore di porre una domanda tanto impensabile quanto provocatoria. Il dibattito relativo ai diritti dei robot, che secondo Gunkel ruota intorno a diverse configurazioni del rapporto tra tesi descrittive (su ciò che è) e normative (su ciò che deve essere), è riconducibile in estrema sintesi a quattro posizioni fondamentali.

La prima posizione è basata sull’osservazione di carattere descrittivo per cui i robot non godono delle caratteristiche necessarie per poter essere sensatamente considerati come soggetti di diritto. Al di là di come possano apparire ai propri utenti, i robot sono solo strumenti tecnologici, per quanto molto complessi, e come tali devono essere trattati. Di conseguenza, sarebbe un errore categoriale – una forzatura – ritenerli soggetti di diritto (al di là dei diritti che riguardano la protezione della proprietà). In più, si correrebbe il rischio di svuotare di significato il discorso sui diritti in generale, con potenziali conseguenze negative su quegli enti, come gli esseri umani, che invece si qualificano come soggetti di diritto.

La seconda posizione sostiene invece che, dal momento che i robot possono avere diritti, allora devono averli. Il ragionamento si basa sulla possibilità futuristica dello sviluppo di macchine intelligenti dotate di quelle caratteristiche (coscienza, libertà, ecc.) che fanno degli umani soggetti di diritto. Se si arriverà mai ad una situazione in cui, da questo punto di vista, umani e robot siano indistinguibili, negare diritti ai secondi sarebbe una forma moralmente inaccettabile di razzismo (o meglio, di specismo). Dal momento che l’intera posizione poggia sulla convinzione che il futuro abbia in serbo tecnologie indistinguibili dalla soggettività umana, la sua forza persuasiva è direttamente proporzionale a quella di questa sua premessa.

Le prime due tesi adducono argomenti descrittivi a giustificazione delle loro posizioni normative, riconoscendo la precedenza dell’ontologia sull’etica. Al contrario, le ultime due esplorano la questione dei diritti dei robot mettendo tra parentesi il dato ontologico, cioè sposando la tesi per cui sia errato giustificare ipotesi etiche basandosi su osservazioni ontologiche.

La terza posizione sostiene che, per quanto i robot possano avere diritti, non dovrebbero averli. Il punto di partenza del ragionamento è la considerazione del fatto che l’estensione o meno di diritti ad enti non convenzionali sia una decisione da prendere al di là di asserzioni ontologiche. È il caso, per esempio, dell’estensione di diritti a statue di divinità, ecosistemi o navi. Il discorso sui diritti segue una logica che è determinata da obiettivi e finalità propri, e solo secondariamente da questioni ontologiche. Tutto ciò, però, non dovrebbe costituire una base per l’estensione di diritti ai robot, in quanto questi ultimi sono prima di tutto strumenti – se non, ma con grande cautela, “schiavi” – che noi utilizziamo per il nostro benessere. La connessione tra uso degli strumenti robotici e benessere umano dovrebbe essere il fine del discorso sui diritti, cioè il dato da proteggere tramite la distribuzione di diritti e doveri. L’estensione di diritti ai robot, però, ne ostacolerebbe l’uso, rendendone vana la produzione. Di conseguenza, è bene che il discorso sui diritti dei robot sia funzionale al loro ruolo sociale di strumenti per il benessere umano. Ciò implica, allo stesso tempo, che sarebbe più saggio evitare di progettare sistemi che possano presentare qualità tali per cui l’estensione dei diritti sarebbe un obbligo morale. Gli umani dovrebbero essere sempre al centro nella relazione con i prodotti tecnologici. L’estensione dei diritti ai robot metterebbe a rischio proprio questa centralità e perciò andrebbe evitata.

La quarta posizione asserisce invece che, sebbene i robot non possano avere diritti, dovrebbero averli. La tesi si basa sul fatto che il modo in cui molti utenti fanno esperienza dei robot – in particolare dei robot sociali, come visto nel capitolo 9 – va ben oltre una loro considerazione quali meri artefatti e, tramite fenomeni come l’antropomorfismo, si spinge fino a chiamare in causa tratti tipici dei rapporti tra esseri viventi, quali ad esempio l’emotività. Tale peculiarità nelle interazioni tra umani e robot dovrebbe distinguere questi ultimi, a volte definiti come “quasi-altri” o “quasi-persone”, dai normali strumenti. L’estensione di diritti ai robot dovrebbe essere presa in considerazione proprio sulla base di tale diversità e dell’interesse che noi, in quanti utenti, nutriamo nella protezione di simili artefatti speciali.

Per quanto le posizioni passate in rassegna abbiano i loro meriti e demeriti, si tratta in ogni caso di tesi che prendono le mosse da tesi descrittive per arrivare a tesi normative. Ciò lascia inesplorata la possibilità che sia la dimensione normativa ad essere prioritaria rispetto a quella descrittiva – come suggerisce il pensiero del filosofo Emmanuel Lévinas. Secondo quest’ultima prospettiva, ciò di cui si fa esperienza si determina solo in un secondo momento, e in base all’esperienza stessa: non prima di essa né tantomeno astraendo da essa. Ciò porta ad assumere una posizione attenta a troncare qualsiasi assimilazione dell’alterità al sé, dell’estraneo al noto, e decisa a non inquinare il modo in cui i robot sono esperiti con dati di carattere descrittivo (basati ad esempio, sulla distinzione tra cose e persone o tra viventi e non viventi). Un simile sforzo dovrebbe condurre all’abbandono della prospettiva antropocentrica – dove siamo noi a dotare o meno altre entità di diritti – verso una nuova via che permetta di pensare altrimenti la questione dei diritti non solo dei robot, ma di ogni alterità non-umana che possa essere esperita come si esperisce un altro.

Al netto dello stato preliminare dei risultati raggiunti, il dibattito sui diritti dei robot è un ambito di ricerca interessante sia in sé, sia per la molteplicità di temi che mette in discussione. Da questo punto di vista, il discorso sui diritti dei robot tocca questioni di grande rilevanza non solo per la filosofia e l’etica della robotica e dell’IA – dall’ambiguo influsso della fantascienza all’analogia tra robot e schiavo, dal problema dell’uso del linguaggio all’antropomorfismo – ma anche per il pensiero etico contemporaneo – dalle questioni dell’antropocentrismo e dell’etnocentrismo all’etica degli animali, dalle difficoltà circa il nesso tra tesi descrittive e normative al problema della riduzione dell’alterità a variazione del sé.

1. Automi e sfide etiche globali

Le sfide poste dalla crisi climatica, dalla fragilità delle istituzioni democratiche e dalle minacce alla pace internazionale coniugano una rilevanza etica indiscutibile con una dimensione genuinamente planetaria. L’IA e la robotica possono dare un contributo importante per affrontare le sfide etiche globali del nostro tempo, ma presentano anche rischi significativi in questo contesto globale, come è stato evidenziato in vari modi nei capitoli precedenti ed è ora opportuno riassumere traendo beneficio da uno sguardo d’insieme sul percorso fin qui compiuto.

La pandemia da Covid-19 ha sollevato sfide di grande rilevanza etica per fronteggiare le quali è necessario mettere in campo analisi e interventi di respiro globale che vertono sulla protezione dello stato di salute, del benessere e dei diritti fondamentali delle persone. Le tecnologie dell’IA forniscono nuovi modelli previsionali, opportunamente sviluppati con tecniche di apprendimento automatico, i quali possono essere utilizzati per prevedere l’andamento della pandemia e contrastarne l’ulteriore diffusione. Per esempio, dati relativi alle misure non farmacologiche messe in atto nei primi mesi della pandemia - distanziamento o confinamento sociale, chiusura delle scuole, coprifuoco, e così via - sono stati utilizzati per addestrare modelli dell’IA e fornire su questa base valutazioni comparative della loro efficacia ai fini della riduzione dell’indice Rt di trasmissibilità dell’infezione (Haug et al., 2020). Nel dispiegare questi nuovi strumenti previsionali bisogna però porre la dovuta attenzione ai limiti dai quali essi possono essere afflitti. Poiché i modelli previsionali dell’IA proiettano sul futuro dati e schemi osservati in passato, è possibile che queste informazioni non siano più statisticamente rappresentative dello stato attuale delle cose e di come esso può evolversi (DECISIONE AUTOMATICA pp…). La loro inadeguatezza può dipendere da una varietà di fattori, che comprendono una nuova composizione della popolazione esaminata, modifiche nei metodi di rilevamento dei dati epidemiologici, un cambiamento di politiche e protocolli sanitari e le mutazioni genetiche degli stessi agenti patogeni. Per utilizzare gli strumenti dell’IA a simili scopi predittivi, gli esseri umani devono vagliare con prudenza la validità dei risultati ottenuti. Come ha sottolineato Alessandro Vespignani, «...quando usiamo l’intelligenza artificiale in modo predittivo per i sistemi sociali dobbiamo essere eccezionalmente cauti e approcciare il problema in modo molto diverso da come sviluppiamo algoritmi per il riconoscimento della grafia o il gioco degli scacchi. Abbiamo bisogno di trasparenza, di interpretabilità, ma dobbiamo soprattutto dare priorità a una comprensione profonda dei meccanismi e dei limiti dei sistemi predittivi» (Vespignani, 2019, p. 104). Questo invito alla cautela porta con sé anche una responsabilità morale che riguarda la comunicazione, da parte di scienziati e tecnologi, di informazioni chiare su quanto sia ragionevole oppure irragionevole attendersi dalle tecnologie dell’IA, contrastando il sensazionalismo, la nascita di aspettative eccessive nell’opinione pubblica e la ricorrente ricerca di una panacea tecnologica ai mali dell’umanità.

Analoghe considerazioni valgono a proposito delle tecnologie robotiche, con l’ausilio delle quali si può contrastare la pandemia da Covid-19 e proteggere i valori minacciati dalla sua diffusione e persistenza. In particolare, l’estensione dell’automazione robotica può incrementare la robustezza e la resilienza del sistema sanitario e del sistema produttivo agli *shock* epidemici. Alcuni sistemi robotici sono utilizzati nelle strutture sanitarie per svolgere mansioni logistiche e di sanificazione. Le imprese manifatturiere potrebbero trarre vantaggio dal dispiegamento di robot controllati a distanza per rimanere operative nel corso di una pandemia; i *cobot*, sistemi robotici capaci di interagire in sicurezza in spazi di lavoro condivisi con un essere umano, possono rivelarsi utili per mantenere il distanziamento sociale nelle attività cooperative di produzione industriale. E tuttavia anche visioni di questo genere richiedono la massima cautela comunicativa verso il pubblico dei non specialisti, facendo riferimento al livello di disponibilità tecnologica (o TRL, acronimo di *technological readiness level*) effettivamente conseguito dai sistemi robotici evocati in tali scenari, nel caso in cui essi non abbiano ancora raggiunto un TRL sufficiente a centrare l’obiettivo della produzione in serie e della conseguente commercializzazione.

La questione della crisi climatica è stata diffusamente affrontata nel capitolo 8 e richiamata in altri capitoli. Da un lato, lo sviluppo attuale dell’IA e della robotica introduce nuove opportunità di sviluppo sostenibile, dall’uso di autoveicoli ad autonomia crescente (capitolo 4) all’elaborazione di dati in base a modelli dell’IA per ottimizzare il consumo energetico nel settore dei trasporti o della produzione di energia elettrica. Da un altro lato, la diffusione crescente di sistemi per elaborare dati in formato digitale, anche finalizzati allo sviluppo di modelli dell’IA, comporta un impatto ambientale significativo e ancora poco riconosciuto. È necessario perciò sviluppare politiche etiche di alto livello e distribuire tra gli attori coinvolti le prerogative, i doveri e le responsabilità del caso per guidare le tecnologie dell’informazione, dell’IA e della robotica verso obiettivi concreti di sostenibilità ambientale e mitigazione del riscaldamento climatico.

Nel capitolo 5 sono state esaminate le problematiche etiche e sociali che nascono dalle nuove possibilità di sorveglianza sociale e di indirizzo delle scelte individuali abilitate dalle tecnologie dell’informatica e dell’IA. È stato in particolare evidenziato come un elemento cardine della democrazia, la possibilità di confronto libero e rispettoso dei diritti fondamentali, sia minacciato da usi impropri dell’IA come strumento di moderazione e censura delle discussioni in rete. Il caso Cambridge Analytica (Mainetti, 2018) è un altro esempio concreto di come i sistemi dell’IA possano minare un elemento cardine della democrazia, e cioè l’autonomia delle scelte politiche dei cittadini. I dati personali di un vasto gruppo di utenti di Facebook, pervenuti alla società Cambridge Analytica, sono stati da questa elaborati con strumenti dell’IA per individuare i loro profili psicologici, confezionare su questa base messaggi elettorali personalizzati e indirizzare le opinioni politiche dei destinatari facendo leva sulle inclinazioni e preferenze personali. La difesa dei processi e delle libertà costitutivi della democrazia diventerà plausibilmente un elemento sempre più centrale nella riflessione etica e nello sviluppo di politiche etiche per il governo dell’IA, in quanto la minaccia che essa pone per l’autonomia personale e l’esercizio consapevole delle libertà democratiche potrà solo acuirsi con l’ampliamento atteso della platea di utenti del web, in notevole crescita a livello mondiale anche tra le fasce meno istruite o educate a orientarsi nella selezione e nella valutazione critica delle informazioni.

Nel capitolo 7 sono stati esaminati i problemi di destabilizzazione internazionale e le minacce alla pace posti congiuntamente dalle tecnologie dell’IA e della robotica con lo sviluppo delle armi autonome. Ma le armi autonome, oltre a escludere gli operatori umani dal loop di controllo, sono anche sistemi collegati in rete, per questo motivo esposti ai cyberattacchi (si veda capitolo 6). Nel cyberspazio, i sistemi del l’IA acquisiranno un’importanza crescente come strumenti di difesa, ma anche come strumenti per sferrare cyberattacchi difficilmente prevedibili e realizzabili in autonomia da un sistema tecnologico in un dominio che desta già forti preoccupazioni per il mantenimento della pace e della stabilità internazionali. Bisogna anche evidenziare che un attacco di questo genere potrebbe investire i sistemi informatizzati di comando e controllo delle armi nucleari, provocando la compromissione dei sistemi di comunicazione tra livelli gerarchici di comando o dei sistemi di attivazione di armi nucleari, l’hackeraggio di questi ultimi o false rilevazioni di attacco nucleare. Come ha osservato Alessandro Pascolini, «sarebbe necessaria una maggiore trasparenza per ridurre malintesi e incomprensioni fra gli stati con armi nucleari su questioni relative all'impiego di tecniche di AI, sviluppando varie forme di misure miranti all'aumento di fiducia reciproca. Appare di interesse comune delle potenze nucleari cooperare per risolvere alcuni dei problemi – quali la sicurezza e la protezione dell'AI – che rendono altamente problematico l'uso di tecniche di AI nell'ambito nucleare» (Pascolini, 2021).

Di fronte a minacce esistenziali di tale portata per tutta l’umanità, è fondamentale porre il valore della pace alla base dello sviluppo dell’IA e della robotica, denunciare e frenare la corsa alle armi convenzionali e cyber abilitate da queste tecnologie, promuovere il dialogo tra potenze militari sui rischi ad essa connessi e incoraggiare gli accordi internazionali per una loro riduzione e controllo. Per affrontare questa sfida globale, insieme alle altre sfide etiche globali che significativamente coinvolgono i sistemi autonomi e intelligenti, sarà necessario che le persone - non gli automi - compiano uno straordinario sforzo di intelligenza e onestà, che si colloca alla confluenza di scienza, tecnologia ed etica.

Bibliografia

acm – association of computing machinery (2018), *Code of Ethics and Professional Conduct*, (<https://www.acm.org/code-of-ethics> consultato l’8 Marzo 2021).

adachi, p. j., willoughby, t. (2011), *The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence?*, in “Psychology of violence”, 1, 4, pp. 259-274.

adams b., breazeal c., brooks r. a., scassellati b. (2000), *Humanoid Robots: A New Kind of* Tool, in “IEEE Intelligent Systems and Their Applications”, 15, 4, pp. 25-30.

adar e., tan d. s., teevan j. (2013), *Benevolent Deception in Human Computer Interaction*, in [*CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*](https://dl.acm.org/doi/proceedings/10.1145/2470654), Association for Computing Machinery, s.l., pp. 1863–72, <https://doi.org/10.1145/2470654.2466246>.

ainsworth m. d. s., blehar m. c, waters e., wall s. (1978), *Patterns of attachment: a psychological study of the strange situation*, Erlbaum, Hillsdale, NJ.

allen a. (1998), *Coercing Privacy*, in “William and Mary Law Review”, 40, 3, pp. 723-757.

altmann j., sauer f. (2017), *Autonomous weapons systems and strategic stability*, in “Survival”, 59/4, pp. 117-142.

amann j., blasimme a., vayena e., frey d., madai v. i. (2020), *Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective*, in “BMC Medical Informatics and Decision Making”, 20, pp. 1–9.

amigoni f., schiaffonati v. (2021), *The Importance of Prediction in Designing Artificial Intelligence Systems*, in M. Pelillo, T. Scantamburtlo (eds), *Machines We Trust: Perspectives on Dependable AI*, MIT Press, Cambridge.

amoroso d. (2020), *Autonomous Weapons Systems and International Law*, Edizioni Scientifiche Italiane, Napoli.

amoroso d., tamburrini g. (2017), *The Ethical and Legal Case Against Autonomy in Weapons Systems,* in “Global Jurist”, 18, 1, <https://doi.org/10.1515/gj-2017-0012>.

amoroso d., tamburrini g. (2019), *I sistemi robotici ad autonomia crescente tra etica e diritto: quale ruolo per il controllo umano?*, in “Rivista di BioDiritto”, 1/2019, pp. 33-51.

amoroso d., tamburrini g. (2021), I*n Search of the ‘Human Element’: International Debates on Regulating Autonomous Weapons Systems*, in “The International Spectator”, DOI: 10.1080/03932729.2020.1864995.

ananny m., crawford k. (2017), *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, in “New Media & Society”, 33, 4, pp. 973–989.

anderson m., anderson s. l. (eds.) (2011), *Machine Ethics*, Cambridge University Press, Cambridge.

angel-fernandez j. m., vincze m. (2018), *Towards a definition of educational robotics*, in P. Zech, J. Piater (eds.), *Proceedings of the Austrian Robotics Workshop 2018* (Issue July), DOI 10.15203/3187-22-1-08.

angwin j., larson j., mattu s., kirchner l. (2016), *Machine Bias*, in “ProPublica”, (<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> consultato 26 Gennaio 2021).

anwar s., bascou n. a., menekse m., kardgar a. (2019), *A systematic review of studies on educational robotics*, in “Journal of Pre-College Engineering Education Research”, 9, 2, pp. 19-42.

appel j. m. (2010), *Sex Rights for the Disabled?*, in “Journal of Medical Ethics”, 36, 3, pp. 152-54.

aristotele, *Etica Nicomachea*, Rizzoli, Milano 1986.

arkin r. c. (2009), *Governing Lethal Behavior in Autonomous Robots*,CRC Press, Boca Raton.

id. (2010), *The Ethics of Robotic Deception*, in “IEEE Technology and Society”, (<https://technologyandsociety.org/ethics-of-robotic-deception/> consultato 05 Marzo 2021).

asaro p. (2012), *On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making,* in “International Review of the Red Cross”, 94, pp. 687-709.

atmatzidou s., demetriadis s., nika p. (2018), *How does the degree of guidance support students’ metacognitive and problem solving skills in educational robotics?*, in “Journal of Science Education and Technology”, 27, 1, pp. 70–85.

attfield r. (2013), *Sustainability*, in H. Lafollette (ed.), *International Encyclopedia of Ethics*, <https://doi.org/10.1002/9781444367072.wbiee033>

avgerinou m., bertoldi p., castellazzi l. (2017), *Trends in Data Centre Energy Consumption under the European Code of Conduct for Data Centre Energy Efficiency,* in “Energies”, 10, 1470, pp. 1-18.

badouard r. (2020), *Les nouvelles lois du web. Modération et censure*, Seuil, Parigi.

balistreri m. (2018), *Sex Robot. L’amore al tempo delle macchine*, Fandango, Roma.

bandura a. (2002), *Selective moral disengagement in the exercise of moral agency*, in “Journal of moral education”, 31, 2, pp. 101-119.

barassi v. (2020), *Child Data Citizen: How Tech Companies Are Profiling Us from Before Birth*, MIT Press, Cambridge.

barocas s., nissenbaum h. (2014), *Big Data’s End Run around Anonymity and Consent*, in J. Lane, V. Stodden, S. Bender, H. Nissenbaum (eds.), *Privacy, Big Data and the Public Good. Frameworks of Engagement,* Cambridge University Press, Cambridge, pp. 44-75, <http://dx.doi.org/10.1017/CBO9781107590205.004>

barocas s., hood s., ziewitz m. (2013), *Governing Algorithms: A Provocation Piece*, *Discussion Paper, Governing Algorithms conference*, NYU, May 16-17, 2013, https://governingalgorithms.org/resources/provocation-piece/#:~:text=This%20provocation%20piece%20addresses%20the,assumptions%20that%20characterize%20current%20debates.

baxter p., ashurst e., read r., kennedy j., belpaeme t. (2017), *Robot education peers in a situated primary school study: personalisation promotes child learning*, in “PLOS ONE”, 12, 5, pp. 1-23.

beam a. l., kohane i. (2016), *Translating Artificial Intelligence Into Clinical Care*, in “JAMA”, 316, 22, pp. 2368-69, <https://doi.org/10.1001/jama.2016.17217>.

beam a. l., kohane i. (2018), *Big Data and Machine Learning in Health Care*, in “JAMA” 319, 13, pp. 1317–18, <https://doi.org/10.1001/jama.2017.18391>.

bechtel, w. (1985), *Realism, Instrumentalism, and the Intentional Stance*, in “Cognitive Science”, 9, 4, pp. 473–497.

belkhir l., elmeligi a. (2018), *Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations*, in “Journal of Cleaner Production”, 177, pp. 448-63.

belpaeme t., kennedy j., ramachandran a., scassellati b., tanaka f. (2018), *Social robots for education: a review*, in “Science Robotics”, 3, 21, pp. 1-9.

benbouzid b. (2017), *Des crimes et des séismes. La police prédictive entre sciences, techniques et divination*, in “Réseaux”, 6, 206, pp. 95-123.

benitti f. b. v. (2012), *Exploring the educational potential of robotics in schools: a systematic review*, in “Computers and Education”, 58, 3, pp. 978–988.

benjamin l. t. (1988), *A history of teaching machines*, in “American Psychologist”, 43, 9, pp. 703–712.

bennett c. j. (2017), *Voter databases, micro-targeting, and data protection law: Can political parties campaign in Europe as they do in North America?*, in “International Data Privacy Law”, 6, 4, pp. 261-275.

bentham j. (1789), *An Introduction to the Principles of Morals and Legislation,* T. Payne, Londra (trad. it. *Introduzione ai principi della morale e della legislazione*, a cura di E. Lecaldano, UTET, Torino 1998)

berk r., heidari h., jabbari s., kearns m., roth a. (2021), *Fairness in Criminal Justice Risk Assessments: The State of the Art,* in “Sociological Methods & Research”, 50, 1, p. 3-44, https://doi.org/10.1177/0049124118782533.

berkhout f., hertin j. (2004), *De-materialising and re-materialising: Digital technologies and the environment*, in “Futures”, 36, 8, pp. 903-920.

berlin, i. (1994), *Il legno storto dell’umanità. Capitoli della storia delle idee*, Adelphi, Milano.

berry d. m. (2011), *The philosophy of software: code and mediation in the digital age*, Palgrave Macmillan, New York.

bhuta n., beck s., geiß r., liy h.-y., kreß c. (eds) (2016), *Autonomous weapons systems. Law, ethics, policy*, Cambridge University Press, Cambridge.

big brother watch (2018), *Police use Experian Marketing Data for AI Custody Decisions*, (<https://bigbrotherwatch.org.uk/all-media/police-use-experian-marketing-data-for-ai-custody-decisions/> consultato il 26 Gennaio 2021)

bijker w. e., law j. (eds.) (1992), *Shaping Technology/Building Society: Studies in Sociotechnical Change*, MIT Press Cambridge.

biller-andorno n. (2012), *Care, Ethics of*, in R.F. Chadwick (ed.), *Encyclopedia of Applied Ethics*, Academic Press, San Diego, pp. 390-396.

billieux j., schimmenti a., khazaal y., maurage p., heeren a. (2015), *Are we overpathologizing everyday life? A tenable blueprint for behavioral addiction research*, in “Journal of Behavioral Addictions”, 4, 3, pp. 119-123.

blair g. s. (2020), *A tale of two cities. Reflections on digital technologies and the natural environment,* in “Patterns”, 1, 5, <https://doi.org/10.1016/j.patter.2020.100068>.

blum a. (2012), *Tubes: Behind the Scenes at the Internet*, Penguin Books Limited, Londra.

bogost i. (2013), *Exploitationware*, in R. Colby, M. S. S. Johnson, R. S. Colby (eds.), *Rhetoric/Composition/Play through Video Games*, Palgrave Macmillan, New York. pp. 139–147.

bonnefon j.-f., shariff a., rahwan i. (2016), *The Social Dilemma of Autonomous Vehicles*, in “Science”, 352, pp. 1573-6.

bonnetain x., canteaut a., cortier v., gaudry p., hirschi l., kremer s., lacour s., leurent g., lequesne m., perrin l., schrottenloher a., thomé e., vaudenay s., vuillot c. (2020), *15 Scenarios Describing Misuses of Tracing App*, in “Tracing-Risks.Com” (https://tracing-risks.com/ consultato il 4 maggio 2010).

borenstein j., arkin r. (2016), *Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being*, in “Science and Engineering Ethics”, 22, 1, pp. 31-46.

bostrom n. (2005). *A history of transumanist thought*, in “Journal of Evolution and Technology”,14, pp. 1-25.

boulanin v., verbruggen m. (2017), *Mapping the Development of Autonomy in Weapon Systems*, SIPRI Report, Stoccolma, (<https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems> consultato il 9 Marzo 2021).

boyd d., crawford k. (2012), *Critical Questions for Big Data*, in “Information, Communication & Society”, 15, 5, pp. 662–667.

bowker g.c., star s.l. (1999), *Sorting Things Out: Classification and Its Consequences*, MIT Press ,Cambridge.

bowlby j. (1988), *A secure base: Parent-child attachment and healthy human development*, Basic Books, New York.

bozzi g., zecca l., datteri e. (a cura di) (2021), *Interazione bambini-robot. Riflessioni teoriche, risultati sperimentali, esperienze*, Franco Angeli, Milano.

breazeal c. (2003), *Toward Sociable Robots*, in “Robotics and Autonomous Systems”, 42, 3-4, pp. 167–75.

brennan t., dieterich w., ehret b. (2009), *Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System*, in “Criminal Justice and Behaviour”, 36, 1, pp. 21-40.

brinkman b., gotterbarn d., miller k., wolf m. j. (2016), *Making a Positive Impact: Updating the ACM Code of Ethics*, in “CACM”, 59, 12, pp. 7-9.

brown r., shelling j. (2019), *Exploring the Implications of Child Sex Dolls*, in “Trends & Issues in Crime and Criminal Justice”, 570, pp. 1-13.

buchanan b. g., shortliffe e. h. (eds.) (1984), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, MA.

boulamwini j., gebru t. (2018), *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in “Proceedings of Machine Learning Research”, 81, pp. 1-15.

burr c., cristianini n., ladyman, j. (2018), *An Analysis of the Interaction Between Intelligent Software Agents and Human Users*, in “Minds and Machines”*,* 28, 4, pp. 735-774.

burr c., cristianini n. (2019), *Can Machines Read our Minds?*, in “Minds and Machines”, 29, pp. 461-494.

burrell j. (2016), *How the machine ‘thinks’: Understanding opacity in machine learning algorithms*, in “Big Data & Society”, 3, 1, pp. 1-12.

cabitza f. (2021), *Cobra AI: exploring some unintended consequences of our most powerful technology*, in M. Pelillo, T. Scantamburlo (eds.), *Machines We Trust. Perspective on Dependable AI*, MIT Press, Cambridge.

caliskan a., bryson j., narayanan a. (2017), *Semantics derived automatically from language corpora contain human-like biases*, in “Science”,356, 6334, pp. 183-186.

calvo r. a., d’mello s., gratch j. m., kappas a. (eds.) (2015), *The Oxford Handbook of Affective Computing*, Oxford University Press, Oxford.

caplan r., boyd d. (2018), *Isomorphism through algorithms: Institutional dependencies in the case of Facebook*, in “Big Data & Society”, 5, 1, pp 1-12.

cardon d. (2013), *Présentation:* *Dossier Politique des algorithmes*, in “Réseaux”, 177, pp. 9-21.

carp, a. (2013), *Inside Manhattan’s tower of internet,* in “The awl” (https://www.theawl.com/2013/03/inside-manhattans-tower-of-internet/ consultato il 25 Maggio 2018).

caruana f., viola m. (2018), *Come funzionano le emozioni*, Il Mulino, Bologna.

cerulli-harms a., münsch m., thorun c., michaelsen f., hausemer p. (2020), *Loot Boxes in Online Games and their Effect on Consumers, in Particular Young Consumers*, Policy Department for Economic, Scientific and Quality of Life Policies, Lussemburgo.

casilli a. a. (2020), *Schiavi del clic. Perché lavoriamo tutti per il nuovo capitalismo?*, Feltrinelli, Milano.

chang c. w., lee j. h., chao p. y., wang c. y., chen g. d. (2010), *Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school*, in “Educational Technology and Society”, 13, 2, pp. 13–24.

chess s., evans n. j., baines j. j. (2017), *What does a gamer look like? Video games, advertising, and diversity*, in “Television & New Media”, 18, 1, pp. 37-57.

chouldechova a. (2017), *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments,* in “Big Data”, 5, 2, pp. 153–163, https://doi.org/10.1089/big.2016.0047.

christin a. (2020), *Metrics at Work: Journalism and the Contested Meaning of Algorithms*, Princeton University Press, Princeton.

clarke s., savulescu j., coady c.a.j., giubilini a., sanyal s. (2016) *The Ethics of Human Enhancement. Understanding the Debate*, Oxford University Press, Oxford.

coeckelbergh m. (2020), *AI Ethics*, MIT Press, Cambridge.

codenotti b., leoncini m. (2020), *La rivoluzione silenziosa. Le grandi idee dell’informatica alla base dell’era digitale*, Codice Edizioni, Torino.

contissa g., lagioia f., sartor g. (2017), *The Ethical Knob: Ethically Customisable Automated Vehicles and the Law*, in “Artificial Intelligence and Law”, 25, 3, pp. 365-78.

covington p., adams j., sargin e. (2016), *Deep Neural Networks for YouTube Recommendations*, in S. Sen, W. Geyer (eds.), *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, 15-19 Settembre 2016), Association for Computing Machinery, New York, pp. 191-198.

crossman c. m. (2019), *Endure and Survive: The Last of Us’s Formation of a New Queer Space in Video Games and the Resulting Rhetoric within the Gaming Community*. California State University, Los Angeles.

cummings m. l. (2020), *Rethinking the maturity of artificial intelligence in safety-critical settings*, in “AI Magazine”, in stampa.

dagum p. (2018), *Digital Biomarkers of Cognitive Function*, in “Npj Digital Medicine”, 10, pp. 1-3.

d’aloia a. (a cura di) (2020), *Intelligenza artificiale e diritto. Come regolare un mondo nuovo*, Franco Angeli, Milano.

danaher j. (2016), *Robots, law and the retribution-gap*, in “Ethics and Information Technology”, 18, 4, pp. 299–309.

id. (2017), *Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?*, in “Criminal Law and Philosophy”, 11, pp. 71-95.

dede g., hamon r., junklewitz h., naydenov r., malatras a., sanchez i. (2021), *Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving*, EUR 30568 EN, Publications Office of the European Union, Lussemburgo, doi:10.2760/551271, JRC122440.

dekkers m. (1994), *Dearest Pet. On Bestiality*, Verso Books, Londra.

denardis l., musiani f. (2016), *Governance by infrastructure*, in F. Musiani, D.L. Cogburn, L. DeNardis, N.S. Levinson (eds.), *The turn to infrastructure in Internet governance*, Palgrave Macmillan, New York, pp. 3-21.

dencik l., hintz a., redden j., treré e. (2019), *Exploring Data Justice: Conceptions, Applications and Directions,* in “Information Communication and Society”, 22, 7, pp. 873–81.

dencik l., hintz a., redden j., warne h. (2018), *Data scores as Governance: Investigating uses of citizen scoring in public services project report*, Cardiff University ([http://orca.cf.ac.uk/117517](http://orca.cf.ac.uk/117517/) consultato il 14 Dicembre 2020)

deterding s., sicart m., nacke l., o'hara k., dixon d. (2011), *Gamification. using game-design elements in non-gaming contexts*, in *CHI'11 extended abstracts on human factors in computing systems*, Association for Computing Machinery, New York, pp. 2425-2428.

devlin k. (2018), *Turned on. Science, Sex and Robots*, Bloomsbury, Londra.

dichev c., dicheva d. (2017), *Gamifying education: what is known, what is believed and what remains uncertain: a critical review*, in “International journal of educational technology in higher education”, 14, 9, pp. 1-36.

dijck j. van (2014), *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology*, in “Surveillance & Society”*,* 12, 2, pp. 197-208.

dill k. e., gentile d. a., richter w. a., dill j. c. (2005), *Violence, sex, race, and age in popular video games: A content analysis*, in E. Cole, J. H. Daniel (eds.), *Featuring females: Feminist analyses of media*, American Psychological Association, Washington, DC, pp. 115–130.

dipartimento della difesa degli stati uniti (2012), *Direttiva 3000.09 “Autonomy in Weapons Systems”*, 21 novembre 2012 (<https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf> consultato il 9 Marzo 2021).

domingo-ferrer j., blanco-justicia a. (2020), *Ethical Value-Centric Cybersecurity: A Methodology Based on a Value Graph,* in “Science and Engineering Ethics”, 26, 3, pp. 1267-1285, <https://doi.org/10.1007/s11948-019-00138-8>.

dp-3t/documents. (n.d.), in “GitHub” (https://github.com/DP-3T/documents consultato il 30 Aprile 2020).

dubber m. d., pasquale f., das s. (eds) (2020), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford.

duggan m. (2015), *Gaming and Gamers*, Pew Research Center, (http://www.pewinternet.org /2015/12/15/gaming-and-gamers/ consultato il 2 Maggio 2017)

earp j., persico d., dagnino f. m., passarelli m., manganello f., pozzi f. (2018), *Ethical issues in gaming: A literature review*, in *Proceedings of the 12th European Conference on Game‐Based Learning, Sophia Antipolis, France‐ECGBL 2018*, Academic Conferences and Publishing limited, Reading, pp. 54-61.

ekelhof m. a. (2018), *Moving beyond semantics on autonomous weapons systems: meaningful human control in operation*, in “Global Policy”, 10, 3, pp. 343-348.

elenko e., underwood l., zohar d. (2015), *Defining Digital Medicine*, in “Nature Biotechnology”, 33, 5, pp. 456-61, <https://doi.org/10.1038/nbt.3222>.

elfiky a. a., pany m. j., parikh r. b., obermayer z. (2018), *Development and Application of a Machine Learning Approach to Assess Short-Term Mortality Risk among Patients with Cancer Starting Chemotherapy*, in “JAMA Network Open”, 1, 3, pp. 1-14.

esteva a., krupel b., novoa r. a., ko j., swetter s. m., blau h. m., thrun s. (2017), *Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks*, in “Nature”, 542, 7639, pp. 115-118.

ethik-kommission (2017), *Automatisiertes und Vernetztes Fahren. Bericht Juni* 2017, Bundesministerium für Verkehr und digitale Infrastruktur, Berlin, (<https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile> consultato il 2 Marzo 2021).

ewoldsen d. r., eno c. a., okdie b. m., velez j. a., guadagno r. e., decoster j. (2012), *Effect of playing violent video games cooperatively or competitively on subsequent cooperative behavior*, in “Cyberpsychology, Behavior, and Social Networking”, 15, 5, pp. 277-280.

facchin f., barbara g., cigoli v. (2017), *Sex Robots: the Irreplaceable Value of Humanity*, in “British Medical Journal”, 358,  <https://doi.org/10.1136/bmj.j3790>.

fawcett t. (2006), *An introduction to ROC analysis*, in “Pattern Recognition Letters”, 27, pp. 861–874

fazel s., wolf a., vazquez-montes m., fanshawe t.r. (2019), *Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)*, in “Scientific reports”, 9, 841, pp. 1-9.

ferguson c. j. (2015). *Do angry birds make for angry children? A meta-analysis of video game influences on children’s and adolescents’ aggression, mental health, prosocial behavior, and academic performance*, in “Perspectives on Psychological Science*”*, 10, 5, pp. 646-666.

ferretti a., schneider m., blasimme a. (2018*), Machine Learning in Medicine: Opening the New Data Protection Black Box*, in “European Data Protection Law Review”, 4, 3, pp. 320-32.

flanagan m., nissenbaum h. (2014), *Values at play in digital games*, MIT Press, Cambridge, MA.

floridi l. (2014), *The Fourth Revolution. How the Infosphere is Reshaping Human Reality*, Oxford University Press, Oxford (trad. it. *La quarta rivoluzione. Come l’infosfera sta trasformando il mondo*, Raffaello Cortina, Milano).

id. (2019), *Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical*, in “Philosophy of Technology”, 32, pp.185–193. <https://doi.org/10.1007/s13347-019-00354-x>.

floridi l., cowls j. (2019*), A Unified Framework of Five Principles for AI in Society*, in “Harvard Data Science Review”, 1, 1, https://doi.org/10.1162/99608f92.8cd550d1

fogg b. j., cuellar g., danielson d. (2008), *Motivating, Influencing, and Persuading Users: An Introduction to Captology*, in A. Sears, J. A. Jacko (eds.), *The Human–Computer Interaction Handbook Fundamentals, Evolving Technologies, and Emerging Applications*, Lawrence Erlbaum Associates, New York, Londra, pp. 134-46.

foot ph. (1967), *The Problem of Abortion and the Doctrine of the Double Effect*, in “Oxford Review”, 5, pp. 5-15.

forbes (2016), *How the ‘Niche’ Sex Toy Market Grew Into an Unstoppable $15B Industry*, ([http://www.forbes.com/sites/janetwburns/2016/07/15/adult-expo-founders-talk-15b-sextoy-industry-after-20-years-in-the-fray/$508e13938a19/](http://www.forbes.com/sites/janetwburns/2016/07/15/adult-expo-founders-talk-15b-sextoy-industry-after-20-years-in-the-fray/%24508e13938a19/) consultato 20 novembre 2020).

formosa p., ryan m., staines d. (2016), *Papers, Please and the systemic approach to engaging ethical expertise in videogames*, in “Ethics and Information Technology”, 18, 3, pp. 211-225.

fosch-villaronga e. (2020), *Robots, Healthcare, and the Law. Regulating Automation in Personal Healthcare,* Routledge, Abingdon.

fosch-villaronga e., poulsen a. (2020), *Exploring the Potential Use of Sexual Robot Technologies for Disabled and Elder Care*, in “Paladyn, Journal of Behavioral Robotics”, 11, pp. 1-18.

fossa f. (2018), *Artificial moral agents: moral mentors or sensible tools?* in “Ethics of Information Technology”, 20**,** pp. 115-126, https://doi.org/10.1007/s10676-018-9451-y.

freitag c., berners-lee m., widdicks k., knowles b., blair g., friday a. (2020), *The climate impact of ICT : A review of estimates, trends and regulations* (<https://arxiv.org/abs/2102.02622>, consultato il 5 Febbraio 2021).

fridin m. (2014), *Kindergarten social assistive robot: First meeting and ethical issues*, in “Computers in Human Behavior”, 30, pp. 262–272.

friedman b. (1996), *Value-Sensitive Design*, in “Interactions”, pp. 17–23.

friedman b., kahn p. h. jr. (2008), *Human Values, Ethics, and Design*, in A. Sears, J. J. Jacko (eds.), *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, Lawrence Erlbaum Associates, New York, Londra, pp. 1241–66.

friedman b., kahn p. h. jr., borning a. (2008), *Value Sensitive Design and Information Systems*, in K. E. Himma, H. T. Tavani (eds.), *The Handbook of Information and Computer Ethics*, John Wiley & Sons, Hoboken, pp. 69-101.

fuller m. (2008), *Software Studies: A Lexicon*, MIT Press, Cambridge.

gabrys j. (2013), *Digital Rubbish: A Natural History of Electronics*, University of Michigan Press, Ann Arbor.

gildea f., richardson k. (2017), *Sex Robots—Why We Should Be Concerned*, (https://campaignagainstsexrobots.org/2017/05/12/sex-robots-why-we-should-be-concerned-byflorence-gildea-and-kathleen-richardson/ consultato 20 novembre 2020).

gillespie t. (2018), *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, New Haven.

gómez-gonzález e., gómez e. (2020), *Artificial Intelligence in Medicine and Healthcare: applications, availability and societal impact*, EUR 30197 EN, Publications Office of the European Union, Lussemburgo, doi:10.2760/047666, JRC120214.

gorwa r., binns r., katzenbach c. (2019). *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, in “Big Data & Society”, 7, 1, pp. 1-15.

gotterbarn d., bruckman a., flick c., miller k., wolf m. j. (2018), *ACM Code of Ethics: A Guide for Positive Action*, in “CACM”, 59, 12, pp. 121-124

gray m., suri s. (2019), *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*, Houghton Mifflin Harcourt, Boston.

greef k. de, van der spek e. d., bekker t. (2013), *Designing Kinect games to train motor skills for mixed ability players*, in B. Schouten, S. Fedtke, T. Bekker, M. Schijven, A. Gekker (eds.), *Games for health*, Springer Vieweg, Wiesbaden, pp. 197-205.

greitemeyer t., mügge d. o. (2014), *Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play*, in “Personality and Social Psychology Bulletin”, 40, 5, pp. 578-589.

grion l. (a cura di) (2012), *La sfida postumanista*, Il Mulino, Bologna.

grodshomme-lulin é. (2015), *Gouverner à l’ère du Big Data. Promesses et périls de l’action publique algorithmique*, , Institut de l’entreprise, Parigi.

groll e., francis d. (2016), *FBI: An Account on Clinton’s Private Email Server Was Hacked – Foreign Policy*. (https://foreignpolicy.com/2016/09/02/fbi-an-account-on-clintons-private-email-server-was-hacked/ consultato il 12 Dicembre 2020).

grote t., berens p. (2020), *On the Ethics of Algorithmic Decision-Making in Healthcare*, in “Journal of Medical Ethics”, 46, 3, pp. 205-11.

guerrero s., lópez-cortéz a., indacochea a., garcía-cárdenas j. m., karina zambrano a., cabrera-andrade a., guevara-ramírez p., abigail gonzález d., leone p. e., paz-y-miño c. (2018), *Analysis of Racial/Ethnic Representation in Select Basic and Applied Cancer Research Studies*, in “Scientific Reports”, 8, 1, pp. 1-8.

guerses s., overdorf r., balsa e. (2018), *POTs: The revolution will not be optimized?*, in *11th Hot Topics in Privacy Enhancing Technologies (HotPETs)* (<https://www.petsymposium.org/2018/files/hotpets/3-gurses.pdf>, consultato il 14 Dicembre 2020).

gulshan v., peng l., coram m., stumpe m. c., wu d., narayamaswamy a., venugopalam s., widner k., madams t., cuadros j., kim r., raman r., nelson p. c., mega j. l., wbester d. r. (2016), *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs*, in “JAMA”, 316, 22, pp. 2402-10.

gunkel d. j. (2018), *Robot Rights*, MIT Press, Cambridge-Londra.

gutiu s. (2012), *Sex Robots and Roboticization of Consent*, We Robot Conference, Miami, pp. 1-24.

id. (2016), *The Robotization of Consent*, in R. Calo, A. M. Froomkin, I. Kerr (eds.), *Robot Law*, Edward Elgar Publishing, Cheltenham, pp. 186-212.

haggerty k., ericson r. (2000), *The surveillant assemblage*, in “The British Journal of Sociology”, 51, 4, pp. 605-622.

hagras h. (2018), *Toward human-understandable, explainable AI*, in “Computer”, 51, 9, pp. 28–36.

hamari j., koivisto j., sarsa h. (2014), *Does gamification work?--a literature review of empirical studies on gamification*, in *Proceedings of the 2014 47th Hawaii international conference on system sciences*, IEEE, Washington, DC, pp. 3025-3034.

hamid m., kuppusamy m. (2017), *Gamification implementation in service marketing: A literature review*, in “Electronic Journal of Business & Management”, 2, 1, pp. 38-50.

han j., jo m., park s., kim s. (2005), *The Educational Use of Home Robots for Children*, in J. Adams (ed.), *14th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 378–383.

hansen h. k., flyverbom m. (2014), *The politics of transparency and the calibration of knowledge in the digital age*, in “Organization”, 22, 6, pp. 872-889.

hanson, j. (2018), *Precautionary Principle: Current Understandings in Law and Society*, in D. A. Dellasala, A. Goldstein (eds.), *Encyclopedia of the Anthropocene*, Elsevier,Oxford-Waltham, pp. 361-6.

harris p.l. (1994), *Understanding Pretence*, in C. Lewis, P. Mitchell (eds.), *Children’s Early Understanding of Mind*, Lawrence Erlbaum Associates, Hove, pp. 235-259.

hartmann t. (2017), *The ‘Moral Disengagement in Violent Videogames’ Model*, in “Game Studies”, 17, 2, <http://gamestudies.org/1702/articles/hartmann>.

hashimoto t., kato n., kobayashi h. (2011), *Development of Educational System with the Android Robot SAYA and Evaluation*, in “International Journal of Advanced Robotic Systems”, 8, 3, pp. 51-61.

hattie j., clarke s. (2018), *Visible Learning: Feedback*. Routledge, New York.

hattie j., timperley h. (2007), *The power of feedback*, in “Review of Educational Research”, 77, 1, pp. 81-112.

haug n. geyrhofer l., londei a., dervic e., desvars-larrive a., loreto v., pinior b., thurner s., klimek p. (2020), *Ranking the effectiveness of worldwide COVID-19 government interventions*, in “Nature Human Behaviour”, 4, pp. 1303-12.

hauskeller m. (2014), *Sex and the Posthuman Condition*, Palgrave, London.

heron m. (2012), *Inaccessible through oversight: the need for inclusive game design*, in “The Computer Games Journal”, 1, 1, pp. 29-38.

herweijer c., waughray d. (2018), *Harnessing Artificial Intelligence for the Earth* (<https://www.pwc.com/gx/en/services/sustainability/publications/ai-for-the-earth.html> consultato il 2 Febbraio 2021).

hildebrandt m. (2016), *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology*, Edward Elgar, Cheltenham.

id. (2021), *The issue of bias: the framing powers of Machine Learning*, in M. Pelillo, T. Scantamburlo, *Machines We Trust. Perspective on Dependable AI*, MIT Press, Cambridge.

hilty l. m., aebischer b., andersson g., lohmann w. (2013), *ICT4S 2013*: *Proceedings of the First International Conference on Information and Communication Technologies for Sustainability, ETH Zurich, February 14-16, 2013* (<https://doi.org/10.3929/ETHZ-A-007337628> consultato il 2 Febbraio 2021).

hleg ai - gruppo indipendente di esperti ad alto livello sull’intelligenza artificiale (2019), *Orientamenti etici per un’Intelligenza Artificiale Affidabile*, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> (consultato il 26 Gennaio 2021).

heyns c. (2016), *Autonomous Weapons Systems: Living a dignified life and dying a dignified death*, in Bhuta *et al.* (2016), pp. 3-19.

himmelreich j. (2019), *Responsibility for Killer Robots*, in “Ethical Theory and Moral Practice”, 22, pp. 731-747.

hood d., lemaignan s., dillenbourg p. (2015), *When children teach a robot to write*, in J. Adams, W. Smart (eds.), *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI ’15*, Association for Computing Machinery, New York, pp. 83-90.

horizon 2020 commission expert group to advise on specific ethical issues raised by driverless mobility (2020), *Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility*, Publication Office of the European Union, Lussemburgo.

hosny a., parmar c., quackenbush j., schwartz l. h., aerts h. j. w. l. (2018), *Artificial Intelligence in Radiology*, in “Nature Reviews Cancer”, 18, 8, pp. 500-510.

huckvale k., venkatesh s., christensen h. (2019), *Toward Clinical Digital Phenotyping: A Timely Opportunity to Consider Purpose, Quality, and Safety*, in “Npj Digital Medicine”, 2, 1, pp. 1-11.

huntemann n. (2013), *Introduction: Feminist discourses in games/game studies*, in “Ada: A Journal of Gender, New Media, and Technology”, 2, https://adanewmedia.org/2013/06/issue2-huntemann/.

hursthouse r., pettigrove g. (2018), *Virtue Ethics*, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2018 Edition)*, https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/.

hwang j., choi l. (2020), *Having fun while receiving rewards?: Exploration of gamification in loyalty programs for consumer loyalty*, in “Journal of Business Research”, 106, pp. 365-376.

icrc – international committee of the red cross (2016), *Views on autonomous weapon system*. Paper submitted to the informal meeting of expert on lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW), Ginevra, 11 Aprile 2016.

ieag – united nations independent expert advisory group on a data revolution for sustainable development (2014), *A world that counts: mobilising the data revolution for sustainable development* (<https://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf> consultato il 3 Febbraio 2021).

ieee – institute of electrical and electronical engineering (2019), *Ethically Aligned Design: Prioritizing Human Wellbeing with Autonomous and Intelligent Systems,* First Edition, <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined>

ienca m., andorno r. (2017), *Towards new human rights in the age of neuroscience and neurotechnology*, in “Life Sciences, Society and Policy”, 13, pp. 1-27, <https://doi.org/10.1186/s40504-017-0050-1>

igda – international game developers association (2015), *Developer Satisfaction Survey 2015 - Summary Report* (<https://www.igda.org/?page=dss2015> consultato il 2 Maggio 2017).

iidea (2019), *I videogiochi in Italia nel 2019. Dati sul mercato e sui consumatori*, <https://iideassociation.com/kdocs/1976720/I_videogiochi_in_Italia_nel_2019.pdf>

ijsselsteijn w., nap h. h., de kort y., poels k. (2007), *Digital game design for elderly users,* in *Proceedings of the 2007 conference on Future Play*, Association for Computing Machinery, New York, pp. 17-22.

isaac a. m. c., bridewell w. (2017), *White Lies on Silver Tongues*, in P. Lin, K. Abney, R. Jenkins (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Oxford University Press, Oxford, pp. 157–72.

istat (2020), *Incidenti stradali in Italia* (<https://www.istat.it/it/archivio/245757> consultato il 15 Gennaio 2021).

jain, s. h., powers b. w., hawkins j. b., brownstein j. s. (2015), *The Digital Phenotype*, in “Nature Biotechnology”, 33, 5, pp. 462-63, <https://doi.org/10.1038/nbt.3223>.

jaquet-chiffelle d.-o., loi m. (2019), *Ethical and unethical hacking,* in M. Christen, B. Gordijn, M. Loi (eds.), *The Ethics of Cybersecurity*, Springer International Publishing, (s.l.), pp. 179-204.

jecker n. (2021), *Nothing to Be Ashamed of: Sex Robots for Older Adults with Disabilities*, in “Journal of Medical Ethics”, 47, 1, pp. 26–32.

jiménez m. r., pulina f., lanfranchi s. (2015), *Video games and Intellectual Disabilities: a literature review*, in “Life Span and Disability”, 18, 2, pp. 147-165.

jobin a., ienca m., vayena e. (2019), *The global landscape of AI ethics guidelines*, in “Nature Machine Intelligence”, 1, 9, pp. 389-99.

johnson d. g. (2009), *Computer Ethics*, Fourth Edition, Prentice-Hall, Upper Saddle River.

johnson d. g., miller k.w. (2009), *Information Flow, Privacy, and Surveillance*, in Id., *Computer ethics: analyzing information technology*, Pearson Education, Upper Saddle River NJ, pp. 81-108.

jussim l., crawford j. t., rubinstein r. s. (2015), *Stereotype (in) accuracy in perceptions of groups and individuals*, in “Current Directions in Psychological Science”, 24, 6, pp. 490-497.

just n., latzer m. (2016), *Governance by algorithms: Reality construction by algorithmic selection on the Internet*, in “Media, Culture & Society”, 39, 2, pp. 238-258.

kanda t., hirano t., eaton d., ishiguro h. (2004), *Interactive robots as social partners and peer tutors for children: a field trial*, in “Human-Computer Interaction”, 19, pp. 61–84.

kant i. (1785), *Grundlegung zur Metaphysik der Sitten*, (trad. it. *Fondazione della metafisica dei costumi*, in *Critica della ragion pratica e altri scritti morali*, a cura di P. Chiodi, UTET, Torino, 1995).

kaptsis d., king d. l., delfabbro p. h., gradisar m. (2016), *Withdrawal symptoms in Internet gaming disorder: A systematic review*, in “Clinical Psychology Review”, 43, pp. 58-66.

kardefelt-winther d. (2014), *A conceptual and methodological critique of internet addiction research: Towards a model of compensatory internet use*, in “Computers in Human Behavior”, 31, pp. 351-354.

katzenbach c., ulbricht l. (2019), *Algorithmic governance*, in “Internet Policy Review”, 8, 4, DOI: 10.14763/2019.4.1424.

kelty c. m. (2017), *Too Much Democracy in All the Wrong Places: Toward a Grammar of Participation*, in “Current Anthropology”, 58, 15, pp. 77-90.

kilburn j. (2009), *The Public Health Risks of Media Violence: a Meta Analytic Review*, in “Journal of Pediatrics”, 154, 5, pp. 759-763.

kim t. w., werbach k. (2016), *More than just a game: ethical issues in gamification*, in “Ethics and Information Technology”, 18, 2, pp. 157-173.

kitchin r. (2016), *Thinking critically about and researching algorithms*, in “Information, Communication & Society”, 20, 1, pp. 14-29.

koenig p. d. (2019), *Dissecting the Algorithmic Leviathan. On the Socio-Political Anatomy of Algorithmic Governance*, in “Philosophy & Technology”, 33, pp. 467-485.

kontula o. (2009), *Between Sexual Desire and Reality: The Evolution of Sex in Finland*, The Population Research Institute, Helsinki.

kuhn t. s. (1974), *The essential tension: Selected studies in scientific tradition and change*, University of Chicago Press, Chicago.

kurzweil r. (2005), *The singularity is near. When humans transcend biology*, Viking Press, New York (trad. it. *La singolarità è vicina*, Apogeo, Milano 2008).

kwak h., blackburn j. (2014), *Linguistic analysis of toxic behavior in an online video game*, in L. M. Aiello, D. McFarland, *Social Informatics. SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014,* Springer, Cham, pp. 209-217.

labrique a., agarwal s., tamrat t., mehl g. (2020), *WHO Digital Health Guidelines: A Milestone for Global Health*, in “Npj Digital Medicine”, 3, 1, pp. 1-3.

latour b. (1988), *The Pasteurization of France*, Harvard University Press, Cambridge.

id. (1992), *Where are the missing masses? The sociology of a few mundane artifacts*, in Bijker, Law, (1992), pp. 151-180.

id. (2005). *Reassembling the social: An introduction to actor-network-theory*, Oxford University Press, Oxford.

leslie a. m. (1987), *Pretense and representation: the origins of “theory of mind”*, in “Psychological Review”, 94, 4, pp. 412-426.

li j. (2015), *The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents*, in “International Journal of Human-Computer Studies”, 77, pp. 23–37.

liberati n. (2021), *Phenomenology and sexrobots*, in “International Journal of Technoethics” (in stampa).

lillard a.s. (1993), *Pretend Play Skills and the Child’s Theory of Mind*, in “Child Development”, 64, pp. 348-371.

liu l., stroulia e., nikolaidis i., miguel-cruz a., rios rincon a. (2016), *Smart Homes and Home Health Monitoring Technologies for Older Adults: A Systematic Review*, in “International Journal of Medical Informatics”, 91, pp. 44-59.

liu x., faes l., kale a. u., wagner s. k., fu d. j., bruynseels a., mahendiran t., moraes g., shamdas m., kern c., ledsam j. r., schmid m. k., balaskas k., topol e. j., bachmann l. m., keane p. a., denniston a. k. (2019*), A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and Meta-Analysis*, in “The Lancet Digital Health”, 1, 6, pp. 271–97.loi m., christen m. (2019), *Ethical Frameworks for Cybersecurity,* in M. Christen, B. Gordijn, M. Loi (eds.), *The Ethics of Cybersecurity*, Springer International Publishing, (s.l.), pp. 73-95.

loi m., christen m., kleine n., weber k. (2019), *Cybersecurity in health – disentangling value tensions,* in “Journal of Information, Communication and Ethics in Society”, 17, 2, pp. 229-245, <https://doi.org/10.1108/JICES-12-2018-0095>.

london a. j. (2019), *Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability*, in “Hastings Center Report”, 49, 1, pp. 15-21, https://doi.org/10.1002/hast.97.

lucivero f. (2019), *Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives*, in “Science and Engineering Ethics”, 26, pp. 1009-1030, https://doi.org/10.1007/s11948-019-00171-7.

luck m. (2009), *The Gamer’s Dilemma*, in “Ethics and Information Technology”, 11, 1, pp. 31-36.

lumbelli l. (1982), *Psicologia dell'educazione: comunicare a scuola*, Il Mulino, Bologna.

lupton d. (2012), *M-Health and Health Promotion: The Digital Cyborg and Surveillance Society*, in “Social Theory & Health”, 10, 3, pp. 229–44, <https://doi.org/10.1057/sth.2012.6>.

id. (2013), *The Digitally Engaged Patient: Self-Monitoring and Self-Care in the Digital Health Era*, in “Social Theory & Health”, 11, 3, pp. 256–70, <https://doi.org/10.1057/sth.2013.10>.

id. (2014), *Self-Tracking Modes: Reflexive Self-Monitoring and Data Practices*, SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, http://papers.ssrn.com/abstract=2483549.

lynch t., tompkins j. e., van driel i. i., fritz n. (2016), *Sexy, strong, and secondary: A content analysis of female characters in video games across 31 years*, in “Journal of Communication”, 66, 4, pp. 564-584.

lyon d. (2014), *Surveillance, Snowden, and Big Data: Capacities, consequences, critique*, in “Big Data & Society”, 1, 2, pp. 1-13.

mainetti e. (2018), *Il caso Cambridge Analytica, spiegato bene*, in “Il Post” (<https://www.ilpost.it/2018/03/19/facebook-cambridge-analytica/> consultato il 10 Marzo 2021).

maras m.-h., shapiro l.r. (2017), *Child Sex Dolls and Robots: More than Just an Uncanny Valley*, in “Journal of Internet Law”, pp. 3-21.

mastro d., behm-morawitz e., ortiz m. (2007), *The cultivation of social perceptions of Latinos: A mental models approach*, in *“*Media Psychology”, 9, 2, pp. 347-365.

mattia d., tamburrini g. (2015), *Ethical Issues in Brain–Computer Interface Research and Systems for Motor Control*, in J. Clausen, N. Levy, *Handbook of Neuroethics*, Springer, Berlin Heidelberg, pp. 725-740.

maurer m., gerdes j. c., lenz b., winner h. (eds.) (2015), *Autonomes Fahren. Technische, Rechtliche und Gesellschaftliche Aspekte*, Springer, Berlino, <https://link.springer.com/book/10.1007/978-3-662-45854-9>.

mayer-schönberger v., cukier k. (2013), *Big Data: una rivoluzione che trasformerà il nostro modo di vivere e già minaccia la nostra libertà*, Garzanti, Milano.

maxwell r., miller t. (2012), *Greening the Media*, Oxford University Press, Oxford.

mcarthur n. (2017), *The Case for Sexbots*, in J. Danaher, N. McArthur (eds.), *Robot Sex. Social and Ethical Implications*, MIT Press, Cambridge, pp. 31-45.

mccaffrey, m. (2019), *The macro problem of microtransactions: The self-regulatory challenges of video game loot boxes*, in “Business Horizons”, 62, 4, pp. 483-495.

mcguire p. (2012), *Death by Algorithm*, in “Huffpost” (<https://www.huffpost.com/entry/teacher-evaluations_b_1328456> consultato 26 Gennaio 2021).

mcewan i. (2019), *Macchine come noi*, Einaudi, Torino.

miley j. (2018), *Sex Robot Samantha Gets an Update to Say ‘No’ If She Feels Disrespected or Bored*, in “Interesting Engineering” (<https://interestingengineering.com/> consultato il 27 Gennaio 2021).

mill j.s. (1863), *Utilitarianism*, Parker, Son & Bourn, Londra (trad. it. *La libertà. L’utilitarismo. L’asservimento delle donne*, a cura di E. Lecaldano, Rizzoli, Milano 2016).

minsky m. (1994), *Will robots inherit the world?,* in “Scientific American”, 271, 4, pp. 109-13.

mitchell t. (1997), *Machine Learning*, McGraw-Hill, Singapore.

mitchell m. (2020), *Artificial Intelligence. A Guide for Thinking Humans*, Pelican Books, Londra.

mod – uk ministry of defense (2017), *The UK Approach to Unmanned Aircraft Systems. Joint Doctrine Publication 0-30.2*, 15 settembre 2017.

moen o. m., sterri a.b. (2018), *Pedophilia and Computer-Generated Child Pornography*, in D. Boonin (ed.), *The Palgrave Handbook of Philosophy and Public Policy*, Palgrave Macmillan, Cham, pp. 369-81.

moor j. (1985), *What is computer ethics?,* in “Metaphilosophy”, 16, 4, pp. 266-275.

moor j. h. (2006), *The Nature, Importance, and Difficulty of Machine Ethics*, in “IEEE Intelligent Systems”, 21, 4, pp. 18-21.

morley j., widdicks k., hazas m. (2018), *Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption*, in “Energy Research & Social Science”, 38, pp. 128-37.

morley j., cowls j., taddeo m., floridi l. (2020), *Ethical guidelines for COVID-19 tracing apps,* in “Nature”, 582, 7810, pp. 29-31, <https://doi.org/10.1038/d41586-020-01578-0>.

morley j., floridi l., kinsey l. elhalal a. (2020), *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*, in “Science and Engineering Ethics”, 26, pp.2141-2168, https://doi.org/10.1007/s11948-019-00165-5

morozov e. (2013), *To save everything, click here: The folly of technological solutionism*, Public Affairs, New York.

movellan j., eckhardt m., virnes m., rodriguez a. (2009), *Sociable robot improves toddler vocabulary skills*, in M. Scheutz, F. Michaud (eds.), *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction - HRI ’09*, Association for Computing Machinery, New York, pp. 307-8.

mubin o., stevens c. j., shahid s., mahmud a., dong j.-j. (2013), *A Review of the Applicability of Robots in Education*, in “Technology for Education and Learning”, 1, 1, pp. 1–7.

mueller-birn c., herbsleb j., dobusch l. (2013), *Work-to-Rule: The Emergence of Algorithmic Governance in Wikipedia*, in *Proceedings of the 6th International Conference on Communities and Technologies*, Association for Computing Machinery, New York, pp. 80-89.

müller j. f., gogoll j. (2020), *Should Manual Driving Be (Eventually) Outlawed?*, in “Science and Engineering Ethics”, 26, pp. 1549-1567, <https://doi.org/10.1007/s11948-020-00190-9>.

musiani f. (2013), *Governance by algorithms*, in “Internet Policy Review”, 2, 3, DOI: 10.14763/2013.3.188.

id. (2018), *Prédiction et personnalisation algorithmique: de nouveaux outils pour la police*, in “Revue de la gendarmerie nationale”, 261, pp. 95-100.

nass c., steuer j., tauber e. r. (1994), *Computers Are Social Actors*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI ’94*, Association for Computing Machinery, New York, pp. 72–78, <http://portal.acm.org/citation.cfm?doid=191666.191703>.

nilsson n. (2010), *The Quest for Artificial Intelligence. A history of ideas and achievements*, Cambridge University Press, New York.

nissenbaum h. (1996), *Accountability in a computerized society*, in “Science and Engineering Ethics”, 2, 1, pp. 25-42

id. (2001), *How Computer Systems Embody Values*, in “Computer”, 34, 3, pp. 119–20.

id. (2004), *Privacy as Contextual Integrity*, in “Washington Law Review”, 79, 1, pp. 119–58.

noble d. f. (1993), *La questione tecnologica*, Bollati Boringhieri, Torino.

noble s. u. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, New York.

nomura t., uratani t., kanda t., matsumoto k., kidokoro h., suehiro y., yamada s. (2015), *Why do children abuse robots?*” in J. Adams, W. Smart (eds.), *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 63–64.

nyholm s., smids j. (2016), *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?,* in “Ethical Theory and Moral Practice”, 19, 5, pp. 1275-89.

nyholm s. (2018), *Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci*, in “Science and Engineering Ethics”, 24, pp. 1201-19, <https://doi.org/10.1007/s11948-017-9943-x>

obermeyer z., emanuel e. j. (2016), *Predicting the Future—Big Data, Machine Learning, and Clinical Medicine*, in “The New England Journal of Medicine”, 375, 13, pp. 1216-1219.

ocse (2019), *The future of work. Employment outlook 2019*, Organisation for Economic Cooperation and Development, (https://read.oecd-ilibrary.org/employment/oecd-employment-outlook-2019\_9ee00155-en#page1 consultato l’11 Febbraio 2021).

oggins j., sammis j. (2012), *Notions of video game addiction and their relation to self-reported addiction among players of World of Warcraft*, in “International Journal of Mental Health and Addiction”, 10, 2, pp. 210-230.

o’neil c. (2016), *Weapons of Math Destruction. How big data increases inequality and threatens democracy*, Crown Publishing, New York (trad. it. *Armi di distruzione matematica. Come i big data aumentano la disuguaglianza e minacciano la democrazia*, Bompiani, Milano, 2017)

onu (2015), *Transforming our world: the 2030 Agenda for Sustainable Development*, (<https://sdgs.un.org/2030agenda> consultato l’8 Marzo 2021)

o’reilly t. (2013), *Open Data and Algorithmic Regulation*, in B. Goldstein, L. Dyson (eds.), *Beyond transparency: Open data and the future of civic innovation,* Code for America Press, San Francisco, pp. 289-300.

oswald m, grace j., urwin s., barnes g.c. (2018), *Algorithmic risk assessment policing models: lessons from the Durham HART model and ‘Experimental’ proportionality*, in “Information & Communications Technology Law”, 27, 2, pp. 223-250.

owen r., macnaghten o., stilgoe j. (2012), *Responsible research and innovation: From science in society to science for society, with society*, in “Science and Public Policy”, 39, 6, pp. 751–760, https://doi.org/10.1093/scipol/scs093.

paaßen b., morgenroth t., stratemeyer m. (2017), *What is a true gamer? The male gamer stereotype and the marginalization of women in video game culture*, in “Sex Roles”, 76, 7, pp. 421-435.

papert s. (1980), *Mindstorms*, Basic Books, New York.

parasuraman r., manzey d. h. (2010), *Complacency and Bias in Human Use of Automation: An Attentional Integration*, in “Human Factors”, 52, 3, pp. 381-410.

park s. j., han j. h., kang b. h., shin k. c. (2011), *Teaching assistant robot, ROBOSEM, in English class and practical issues for its diffusion*, in *IEEE Workshop on Advanced Robotics and its Social Impacts*, (s.e.), Menlo Park, pp. 8–11.

pascolini a. (2021), I*ntelligenza artificiale e stabilità strategica*, in “Il Bo Live, Rivista online dell’Università di Padova”, <https://ilbolive.unipd.it/it/news/lintelligenza-artificiale-stabilita-strategica>.

passarelli m., earp j., dagnino f. m., manganello f., persico d., pozzi f., buijtenweg t., haggis m., bailey c., perrotta c. (2020), [*The Distant Horizon: investigating the relationship between social sciences academic research and game development*](https://www.sciencedirect.com/science/article/pii/S1875952119301053), in“Entertainment Computing”, DOI [10.1016/j.entcom.2020.100339](https://doi.org/10.1016/j.entcom.2020.100339).

pasquale f. (2015), *The black box society: the secret algorithms that control money and information*, Harvard University Press, Cambridge.

patrignani n. (2015*), Slow Tech: per un'informatica buona, pulita e giusta*, in “Mondo Digitale”, 14, 59, pp. 1-9.

pelillo m., scantamburlo t. (2013), *How Mature Is the Field of Machine Learning?*, in M. Baldoni, C. Baroglio, G. Boella, R. Micalizio (eds.), *AI\*IA2013: Advances in Artificial Intelligence*, Atti del congresso dell’Associazione Italiana di Intelligenza Artificiale (Torino, 4-6 Dicembre), Springer, Berlino, pp. 121-132.

perrault r., shoham y., brynjolfsson e., clark j., etchemendy j., grosz b., lyons t., manyika j., mishra s., niebles j.c. (2019), *The AI Index 2019 Annual Report*, Human-Centered AI Institute, Stanford University, Stanford (<https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf> consultato il 26 Gennaio 2021)

perrotta c., bailey c., ryder j., haggis-burridge m., persico d. (2020), *Games as (not) culture: a critical policy analysis of the economic agenda of Horizon 2020*, in “Games and culture”, 15, 8, pp. 902-922. DOI [10.1177/1555412019853899](https://doi.org/10.1177/1555412019853899).

persico d., passarelli m., dagnino f., manganello f., earp j., pozzi f. (2018*), Games and learning: Potential and limitations from the players’ point of view*, in *International Conference on Games and Learning Alliance*, Springer, Cham, pp. 134-145.

persico d., passarelli m., pozzi f., earp. j., dagnino f.m., manganello f. (2019). *Meeting Players Where They Are: Digital Games and Learning Ecologies*, in “British Journal of Education Technology”, 50, 4, pp. 1687-1712, doi 10.1111/bjet.12777.

petry n. m., rehbein f., gentile d. a., lemmens j. s., rumpf h. j., mößle t., auriacombe, m. (2014), *An international consensus for assessing internet gaming disorder using the new DSM‐5 approach*, in “Addiction”, 109, 9, pp. 1399-1406.

piaget j. (1926), *La représentation du monde chez l'enfant*, Presses universitaires de France, Parigi (trad. it *La rappresentazione del mondo nel fanciullo,* Bollati Boringhieri, Milano 1966).

piha s., hurmerinta l., sandberg b., järvinen e. (2018), *From Filthy to healthy and Beyond: Finding the Boundaries of Taboo Destruction in Sex Toys Buying*, in “Journal of Marketing Management”, pp. 1078-1104.

platone, *Repubblica*, Bompiani, Milano 2000.

plothe t. (2018), *The Whose View of Hue?: Disability adaptability for color blindness in the digital game Hue*, in “G| A| M| E Games as Art, Media, Entertainment”, 1, 7, <https://www.gamejournal.it/07_plothe/>.

pourhomayoun m., shakibi m. (2021), *Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making*, in “Smart Health”, 20, pp. 1-8, <https://doi.org/10.1016/j.smhl.2020.100178>

pucciarelli m., virgulti l., farina e., datteri e. (2020), *Non-humanoid and non-social robots for the study of the attribution of intentionality in children with autism*, in “Sistemi Intelligenti”, 32, 1, pp.107–122.

quintarelli s., corea f., ferrauto c. g., fossa f., loreggia a., sapienza s. (2020), *Intelligenza Artificiale. Cos’è davvero, come funziona, che effetti avrà*, Bollati Boringhieri, Torino.

rachels j. (1975), *Why Privacy is Important*, in “Philosophy and Public Affairs”, 4, 4, pp. 323-333.

redwood, s., gill p. s., *Under-Representation of Minority Ethnic Groups in Research—Call for Action*, in “British Journal of General Practice”, 63, 612, pp. 342-343.

regan p. (2015), *Privacy and the common good: Revisited*, in B. Roessler, D. Mokrosinska (eds.), *Social Dimensions of Privacy: Interdisciplinary Perspectives*, Cambridge University Press, Cambridge, pp. 50-70. doi:10.1017/CBO9781107280557.004.

richardson k. (2016), *The Asymmetrical ‘Relationship’: Parallels between Prostitution and the Development of Sex Robots*, in “ACM SIGCAS Computers and Society”, 45, 3, pp. 290-293.

rieder g., simon j. (2016), *Datatrust: Or, The Political Quest for Numerical Evidence and the Epistemologies of Big Data*, in “Big Data & Society”, 3, 1, pp. 1-6.

riegel d. l. (2004), *Effects on Boy-Attracted Pedosexual Males of Viewing Boy Erotica*, in “Archives of Sexual Behavior”, 33, 4, pp. 321-23.

rigby c.s. (2015), *Gamification and Motivation*, in S. Deterding, S. Waltz (ed.), *The gameful world: Approaches, issues, applications*,MIT Press, Cambridge, pp. 225–244.

rgdp (2016), *Regolamento (UE) 2016/679 del Parlamento Europeo e del consiglio del 27 aprile 2016 relativo al relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati)*, (<https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016R0679&from=IT> consultato il 22 Febbraio 2021).

ricoeur p. (1990), *Éthique et morale,* in “Revue de l’Institut catholique de Paris”, 34, pp. 131-142 (trad. it. in Id., *Etica e Morale*, Morcelliana, Brescia 2014).

roberts s. t. (2019), *Behind the screen: Content moderation in the shadows of social media*, Yale University Press, New Haven.

rodotà s. (2005), *Intervista su privacy e libertà*, a cura di P. Conti, Laterza, Roma-Bari.

rosenblat a. (2018), *Uberland: How algorithms are rewriting the rules of work*, University of California Press, Oakland.

rouvroy a., berns t. (2010), *Le nouveau pouvoir statistique*, in “Multitudes”, 40, pp. 88-103.

russell s. (2019), *Human Compatible. Artificial Intelligence and the Problem of Control*, Viking, New York.

ryan r. m., deci e. l. (2000), *Intrinsic and extrinsic motivations: Classic definitions and new directions*, in “Contemporary educational psychology”, 25, 1, pp. 54-67.

sae international (2016), *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,* pp. 1-30, (<https://www.sae.org/standards/content/j3016_201609/> consultato in data 26 Gennaio 2021).

sailer m., hense j. u., mayr s. k., mandl h. (2017), *How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction*, in “Computers in Human Behavior”, 69, pp. 371–380. <http://doi.org/10.1016/j.chb.2016.12.033>.

samuel a. l. (1959), *Some Studies in Machine Learning Using the Game of Checkers*, in “IBM Journal of Research and Development”, 3, 3, pp. 210-229.

samuel g., derrick g.e., van leeuwen t. (2019), *The Ethics Ecosystem: Personal Ethics, Network Governance and Regulating Actors Governing the Use of Social Media Research Data*, in “Minerva”, 57, pp 317–43.

samuel g., lucivero f. (2020), *Responsible Open Science: Moving towards an Ethics of Environmental Sustainability*, in “Publications”, 8(4), 54, <https://doi.org/10.3390/publications8040054>.

sandel j.m. (2009), *Justice. What’s the right thing to do*, Farrar, Straus and Giroux, New York (trad. it. *Giustizia. Il nostro bene comune*, Feltrinelli, Milano 2010).

santoni de sio f., van den hoven j. (2018), *Meaningful human control over autonomous systems: A philosophical account*, in “Frontiers of Robotics and AI”, 5, 15, https://doi.org/10.3389/frobt.2018.00015.

sarkeesian a. (2013), *Tropes versus women in video games*, [online video files], (https://goo.gl/9bFRfX consultato il 28 Aprile 2018).

scantamburlo t., charlesworth a., cristianini n. (2019), *Machine decisions and human consequences*, in K. Yeung, M. Lodge (eds.), *Algorithmic Regulation,* Oxford University Press, Oxford, pp. 49-81.

schafer b., frankowski d., herlocker j., sen s. (2007), *Collaborative Filtering Recommender Systems*, in P. Brusilovsky, A. Kobsa, W. Nejdl (eds.), *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer, Berlin, pp. 291-324.

schlosberg d. (2009), *Defining environmental justice: theories, movements, and nature*, Oxford University Press, Oxford.

schrier k. (2015), *EPIC: A framework for using video games in ethics education* in “Journal of Moral Education”, 44, 4, pp. 393-424.

schrier k., farber m. (2017), *The Limits and Strengths of Using Digital Games as “Empathy Machines”*, (https://d1c337161ud3pr.cloudfront.net/files%2Fd61c7672-81d3-4ae0-8cc0-b14f53d1ab01\_Working%20Paper%205.pdf consultato l’8 Marzo 2021).

schumpeter j. a. (1964), *Capitalismo, Socialismo, Democrazia*, a cura di E. Zuffi*,* Edizioni Di Comunità*,* Milano.

schwab k. (2016), *La Quarta Rivoluzione Industriale*, FrancoAngeli, Milano.

selfridge o. g. (1955), *Pattern recognition and modern computers*, in W.L. Martin (ed.), *Proceedings of the western joint computer conference* (Los Angeles, 1-3 Marzo 1955), Association for Computing Machinery, New York, pp. 91-93.

sharkey a. j. c. (2016), *Should we welcome robot teachers ?* in “Ethics and Information Technology”, 18, 4, pp. 283–297.

sharkey n. (2010), *Saying ‘no’ to lethal autonomous targeting*, in “Journal of Military Ethics”, 9, 4, pp. 369-383.

id. (2016), *Staying the Loop: human supervisiory control of weapons,* in N. Bhuta *et al.* (2016), pp. 23-38.

sharkey a., sharkey n. (2010), *The crying shame of robot nannies*, in “Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems”, 11, 2, pp. 161–190.

sharkey a., sharkey n. (2011), *Children, the Elderly, and Interactive Robots: Anthropomorphism and Deception in Robot Care and Companionship*, in “IEEE Robotics and Automation Magazine”, 18, 1, pp. 32–38.

shaw a., friesem e. (2016), *Where is the queerness in games?: Types of lesbian, gay, bisexual, transgender, and queer content in digital games*, in “International Journal of Communication”, 10, pp. 3877-3889.

shmueli g. (2010), *To Explain or to Predict*, in “Statistical Science”, 25, 3, pp. 289-310.

sicart m. (2015*), Playing the good life: Gamification and ethics*, in S. Deterding, S. Waltz (eds.), *The gameful world: Approaches, issues, applications*, MIT Press, Cambridge, pp. 225–244.

siegel m., breazeal c., norton m. i. (2009), *Persuasive Robotics: The Influence of Robot Gender on Human Behavior*, in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, IEEE Press, (s.l.), pp. 2563–68.

singer p. (2001), *Heavy Petting*, (https://www.utilitarian.net/singer/by/2001----.htm consultato l’8 Marzo 2021).

skinner b. f. (1958), *Teaching Machines*, in “Science”, 128, 3330, pp. 969–977.

slote m., pettit ph. (1984), *Satisficing Consequentialism*, in “Proceedings of the Aristotelian Society”, 58, pp. 139-63.

smith, k. (2003), *Wendell Berry and the Agrarian Tradition: A Common Grace*, University of Kansas Press, Lawrence.

soble a. (2007), *Pornografia, sesso e femminismo*, Effepi Libri, Monte Porzio Catone.

somenzi v. (a cura di) (1965), *Filosofia degli automi*, Bollati Boringhieri, Torino.

sparrow r. (2012), *Can Machines Be People? Reﬂections on the Turing Triage Test*, in P. Lin, K. Abney, G. Bekey (eds.), *Robot Ethics: the Ethical and Social Implications of Robotics*, MIT Press, Cambridge, pp. 301– 315.

sparrow r. (2017), *Robots, Rape, and Representation*, in “International Journal of Social Robotics”, 9, 4, pp. 465-477.

id. (2020), *Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might it Be Explained?*, in “International Journal of Social Robotics”, pp. 1-7, DOI: 10.1007/s12369-020-00631-2.

id. (2021), *Sex robot fantasies*, in “Journal of Medical Ethics”, 47, 1, pp. 33-34.

sparrow r., sparrow l. (2006), *In the Hands of Machines? The Future of Aged Care*, in “Minds and Machines”, 16, 2, pp. 141–61.

strikwerda l. (2017), *Legal and Moral Implications of Child Sex Robots*, in J. Danaher, N. McArthur (eds.), *Robot Sex. Social and Ethical Implications*, MIT Press, Cambridge, pp. 133-51.

strubell e., ganesh a., mccallum a. (2019), *Energy and policy considerations for deep learning in NLP*, arXiv preprint (arXiv:1906.02243).

swartout-corbeil d.m. (2006), *Attachment between infant and caregiver*, in *The Gale Encyclopedia of Children’s Health: Infancy through Adolescence*, The Gale Group, (s.l.).

szegedy c., wojciech z., sutskever i., bruna j., erhan d., goodfellow i., fergus r. (2014), *Intriguing properties of Neural Networks*, in Y. Bengio, Y. LeCun (eds.), *Proceeding of the second international conference on learning Representations* (Banff, 14-16 Aprile 2014), https://arxiv.org/abs/1312.6199.

taddeo m. (2019), *Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity,* in “Minds and Machines”, 29, 2, pp. 187–191, <https://doi.org/10.1007/s11023-019-09504-8>.

taeihagh a., lim h. s. m. (2019), *Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks*, in “Transport Reviews”, 39, 1, pp. 103-28, <https://www.tandfonline.com/doi/pdf/10.1080/01441647.2018.1494640?needAccess=true>

tafani d. (2020), *Sulla moralità artificiale. Le decisioni delle machine tra etica e diritto*, in “Rivista di Filosofia”, CXI, 1, pp. 81-103.

taiebat m., brown a. l., safford h. r., qu s., xu **m.**(2018), *A Review on Energy, Environmental, and Sustainability Implications of Connected and Automated Driving*, in “Environmental Science & Technology”, 52, 20, pp. 11449-65.

tamburrini g. (2009), *Brain to computer communication: ethical perspectives on interaction models*, in “Neuroethics”, 2, pp. 137-149.

id. (2012), *La simbiosi cervello-calcolatore e il transumanesimo*, in (Grion 2012), pp. 83-99.

id. (2020), *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma.

tanaka f., cicourel a., movellan j. r. (2007), *Socialization between toddlers and robots at an early childhood education center*, in “Proceedings of the National Academy of Sciences”, 104, 46, pp. 17954–17958.

tanaka f., kimura t. (2009), *The use of robots in early education: A scenario based on ethical consideration*, in *Proceedings of* *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE Press, Piscataway, pp. 558–560.

tanaka f., matsuzoe s. (2012), *Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning*, in “Journal of Human-Robot Interaction”, 1, 1, pp. 78–95.

taylor, l. (2017), *What is data justice? The case for connecting digital rights and freedoms globally*, in “Big Data & Society”, 4, 2, pp. 1-14, <https://doi.org/10.1177/2053951717736335>.

thagard p. (1988), *Computational Philosophy of Science*, MIT Press, Cambridge.

thaler r. h., sunstein c. r. (2008), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, New Haven (trad. it. *Nudge: la spinta gentile. La nuova strategia per migliorare le nostre decisioni su denaro, salute, felicità,* Feltrinelli, Milano 2014).

thorpe a. s., roper s. (2019), *The ethics of gamification in a marketing context*, in “*Journal of business ethics*”, 155, 2, pp. 597-609.

tommaso d’aquino, *La Somma di Teologia*, ESD, Bologna 2014.

treccani (2013), *Nimby, effetto*, in *Lessico del XXI secolo*, 2 voll., Istituto della Enciclopedia Italiana, Roma, (<http://www.treccani.it/enciclopedia/effetto-nimby_%28Lessico-del-XXI-Secolo%29/> consultato il 22 Dicembre 2020).

troncoso, c., payer, m., hubaux, j.-p., salathe, m., larus, j., bugnion, e., lueks, w., stadler, t., pyrgelis, a., antonioli, d., barman, l., chatel, s., paterson, k., capkun, s., basin, d., jackson, d., preneel, b., smart, n., singelee, d., wiegand, t. (2020), *Decentralized Privacy-Preserving Proximity Tracing: Overview of Data Protection and Security,* in “GitHub”, (<https://github.com/DP-3T/documents/blob/master/DP3T%20-%20Data%20Protection%20and%20Security.pdf> consultato l’8 Marzo 2021).

turing a. m. (1950), *Computing Machinery and Intelligence*, in “Mind”, LIX, 236, pp. 433-60 (trad. it. in Somenzi, 1965).

turner r. k. (1979), *Cost-benefit analysis—a critique,* in “Omega”, 7, 5, pp. 411–9.

ulivieri m. (a cura di) (2014) *Loveability. L’assistenza sessuale per le persone con disabilità*, Erickson, Trento.

van de poel i., royakkers l. (2011), *Ethics, Technology and Engineering: an Introduction*, Wiley-Blackwell, Hoboken.

id., (2016), *An Ethical Framework for Evaluating Experimental Technology*, in “Science and Engineering Ethics”, 22, pp 667–686.

van den hoven j. (2008), *Moral Methodology and Information Technology*, in K. E. Himma, H. T. Tavani (eds.), *The Handbook of Information and Computer Ethics*, John Wiley & Sons, Hoboken, pp. 49-67.

van wynsberghe a., robbins s. (2019), *Critiquing the Reasons for Making Artificial Moral Agents*, in “Science and Engineering Ethics”, 3, 25, pp. 719–735.

vaudenay, s. (2020), *Centralized or Decentralized? The Contact Tracing Dilemma,* in “Cryptology ePrint Archive”, 531, (<http://eprint.iacr.org/2020/531> consultato il 12 Dicembre 2020).

vayena e., haeusermann t., adjekum a., blasimme a. (2018), *Digital Health: Meeting the Ethical and Policy Challenges*, in “Swiss Medical Weekly”, 148, pp. 1-9, doi: 10.4414/smw.2018.14571.

vayena e., blasimme a., cohen g. (2018), *Machine Learning in Medicine: Addressing Ethical Challenges*, in “PLoS Medicine”, 15, 11, pp. 1-4, https://doi.org/10.1371/journal.pmed.1002689.

velez j. a., mahood c., ewoldsen d. r., moyer-gusé e. (2014), *Ingroup versus outgroup conflict in the context of violent video game play: The effect of cooperation on increased helping and decreased aggression*, in “Communication Research”, 41, 5, pp. 607-626.

verbruggen m. (2019), *The question of swarms control: challenges to ensuring human control over military swarms*, in “Non proliferation and Disarmament Paper”, 65, pp. 1-16, https://www.sipri.org/sites/default/files/2019-12/eunpdc\_no\_65\_031219.pdf.

verschueren k., koomen h. m. (2012), *Teacher–child relationships from an attachment perspective*, in “Attachment & human development”, 14, 3, pp. 205-211.

vespignani a. (con R. Rijtano) (2019), *L’algoritmo e l’oracolo. Come la scienza predice il futuro e ci aiuta a cambiarlo*, Il Saggiatore, Milano.

viganò e., loi m., yaghmaei e. (2019), *Cybersecurity of critical infrastructure,* in M. Christen, G. Bert, M. Loi (eds.), *The Ethics of Cybersecurity*, Springer, (s.l.), pp. 157-177.

von luxburg u., schölkopf b. (2009), *Statistical Learning Theory: Models, Concepts and Results,* in D.M., Gabbay, S. Hartmann, J. Woods (eds.), *Inductive Logic*, North Holland, (s.l.), pp. 651-706.

vonderau a. (2018), *Scaling the Cloud: Making State and Infrastructure in Sweden*,in “Ethnos Journal of Anthropology”, 84, 4, 698-718.

wallach w., allen c. (2009), *Moral Machines. Teaching Robots Right from Wrong*, Oxford University Press, Oxford.

walzer m. (1977), *Just and unjust wars*, Basic Books, New York (trad. it. *Guerre giuste e ingiuste*, Liguori Editore, Napoli 1990).

warwick k., grasson m., hutt b., goodhew e., kyberd p., andrews b., teddy p., shad a. (2003). *The application of implant technology for cybernetic systems*, in “Archives of Neurology”, 60, pp. 1369-73.

waseem z., hovy d. (2016), *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,* in J. Andreas, E. Choi, A. Lazaridou (eds), *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, pp. 88-93.

watzlawick p., beavin j. h., jackson d. d. (1967), *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes,* WW Norton & Company, New York.

wced – world commission on environment and development (1987), *Our common future*, (<https://www.un.org/ga/search/view_doc.asp?symbol=A/42/427&Lang=E> consultato il 3 Febbraio 2021).

wellman h. m., liu d. (2004), *Scaling of Theory-of-Mind Tasks*, in “Child Development”, 75, 2, pp. 523-541.

whitehead b., andrews d., shah a., maidment g. (2014), *Assessing the environmental impact of data centres part 1: Background, energy use and metrics*, in “Building and Environment”, 82, pp. 151-9.

who – world health organization (2020), *International statistical classification of diseases and related health problems (11th ed.)*, (<https://icd.who.int/> consultato l’8 Marzo 2021).

wilkinson j., arnold k. f., murray e. j., smeden m. van, carr k., sippy r., kamps m. de, beam a., konigorski s., lippert c., gilthorpe m. s., tennat p. w. g. (2020), *Time to Reality Check the Promises of Machine Learning-Powered Precision Medicine*, in “The Lancet Digital Health”, 2, 12, pp. 677-680, [https://doi.org/10.1016/S2589-7500(20)30200-4](https://doi.org/10.1016/S2589-7500%2820%2930200-4).

williams d., martins n., consalvo m., ivory j. d. (2009), *The virtual census: Representations of gender, race and age in video games*, in “New Media & Society”, 11, 5, pp. 815-834.

williams e. (2011), *Environmental effects of information and communications technologies*, in “Nature”, 479, 7373, pp. 354-8.

williamson shaffer d., kigin c. m., kaputt j. j., gazelle g. s. (2002), *What Is Digital Medicine?*, in “Studies in Health Technology and Informatics”, 80, pp. 195-204.

winfield a. f. t., jirotka m. (2018), *Ethical governance is essential to building trust in robotics and artificial intelligence systems,* in “Philosophical Transaction of the Royal Society”, A, 376, pp. 1-13, [http://doi.org/10.1098/rsta.2018.0085](https://doi.org/10.1098/rsta.2018.0085)

winner l. (1980), *Do Artifacts Have Politics?*, in “Daedalus”, 109, 1, pp. 121-136.

wohn d. y. (2011*), Gender and race representation in casual games*, in “Sex roles”, 65, 3-4, pp. 198-207.

wosk j. (2015), *My Fair Ladies. Female Robots, Androids and Other Artificial Eves*, Rutgers University Press, New Brunswick.

wu t. (2010), *The Master Switch: The Rise and Fall of Information Empires*, Random House, New York.

yougov (2017), *1 in 4 Men Would Consider Having Sex with a Robot, September 27, 2017*, (<https://today.yougov.com/news/2017/10/02/1-4-men-would-consider-having-sex-robot/> consultato il 20 Novembre 2020).

young g. (2016), *Resolving the Gamer’s Dilemma*, Palgrave, Cham.

yuan b., folmer e., harris f. c. (2011), *Game accessibility: a survey*, in “Universal Access in the information Society”, 10, 1, pp. 81-100.

yeung k. (2018). *Algorithmic regulation: A critical interrogation*, in “Regulation & Governance”, 12, 4, pp. 505-523.

yun s., shin j., kim d., kim c. g., kim m., choi m.-t. (2011), *Engkey: Tele-education Robot*, in B. **Mutlu**, C. **Bartneck**, J. **Ham**, V. **Evers**, T. **Kanda** (eds.), *Proceedings of* *Social Robotics. Third International Conference, ICSR 2011*, Springer, Berlin-Heidelberg, pp. 142–152.

zagal j. p. (2012), *The Videogame Ethics Reader*, Cognella, San Diego.

zara g. (2018), *La psicologia dei sexbot nel trattamento dei sex offender*, in M. Balistreri, *Sex Robot. L’amore al tempo delle macchine*, Fandango, Roma, pp. 225-282.

zecca l. (2016), *Didattica laboratoriale e formazione. Bambini e insegnanti in ricerca*, FrancoAngeli, Milano.

zhang y., song w., tan z., wang y., lam c. m., hoi s. p., xiong q., chen j., yi l. (2019), *Theory of Robot Mind: False Belief Attribution to Social Robots in Children with and without Autism*, in “Frontiers in Psychology”, 10, p. 1732.

zhu j. s., ge p., jiang c., zhang y., li x., zhao z., zhang l., duong t. k. (2020), *Deep‐learning Artificial Intelligence Analysis of Clinical Variables Predicts Mortality in COVID‐19 Patients*, in “Journal of the American College of Emergency Physicians Open”, 1, 6, pp. 1364-1373.

zichermann g., cunningham c. (2011), *Gamification by design: Implementing game mechanics in web and mobile apps*. O'Reilly Media, Sebastopol.

ziewitz m. (2016), *Governing Algorithms: Myth, Mess, and Methods*, in “Science, Technology, & Human Values”, 41, 1, pp. 3-16.

Le autrici e gli autori

**Maurizio Balistreri** è ricercatore di filosofia morale presso il Dipartimento di Filosofia e Scienze dell’Educazione dell’Università di Torino. È autore di numerosi saggi di bioetica ed etica applicata: gli ultimi libri sono *Superumani: etica e potenziamento* (Espress 2011, 2020), *Il futuro della riproduzione umana* (Fandango 2016) e *Sex robot. L’amore al tempo delle macchine* (Fandango 2018; Malpaso 2021). È uno degli autori di *Biotecnologie. Modificazioni genetiche* (Mulino 2020).

**Alessandro Blasimme**, laureato in Filosofia presso La Sapienza – Università di Roma, ha ricoperto incarichi di ricerca presso l’Istituto Nazionale di Salute e Ricerca Biomedica (INSERM) in Francia, l’università di Harvard negli Stati Uniti (come *Fulbright-Schuman Research Scholar*) e l’Università di Zurigo. Attualmente è *Senior Scientist* presso il Politecnico Federale di Zurigo (ETH Zürich). Le sue competenze vertono in particolare sugli aspetti etici e politici legati all’innovazione in ambito biomedico.

**Francesca Dagnino** è psicologa e psicoterapeuta, lavora presso ITD-CNR dal 2010, dove è diventata ricercatrice nel 2018. Ha partecipato a diversi progetti nazionali ed internazionali nel campo del *Technology Enhanced Learning*. I suoi principali interessi di ricerca sono l'apprendimento basato sui giochi e la gamification, la progettazione didattica, l’inclusione socio-educativa nell’area della scuola in ospedale e dell’istruzione domiciliare.

**Edoardo Datteri** è docente di Logica e Filosofia della Scienza e di Robotica Educativa presso l’Università degli Studi di Milano-Bicocca. Coordina il Laboratorio di Robotica per le Scienze Cognitive e Sociali dello stesso Ateneo. Si occupa dei fondamenti epistemologici e metodologici delle scienze cognitive, con particolare riferimento all’analisi dei ruoli svolti dai robot, dai sistemi bionici e dai sistemi di Intelligenza Artificiale nello studio della cognizione individuale e sociale.

**Fabio Fossa** è assegnista di ricerca presso il Dipartimento di Meccanica del Politecnico di Milano, dove si occupa di filosofia degli agenti artificiali e etica dei veicoli autonomi. La sua ricerca verte su temi di etica applicata, filosofia della tecnologia, etica della robotica e dell’IA, e sul pensiero di Hans Jonas. È direttore della rivista InCircolo – Rivista di filosofia e culture e membro fondatore del gruppo di ricerca Zetesis.

**Giovanni Grandi** è professore associato di Filosofia Morale presso il Dipartimento di Scienze Politiche e Sociali dell'Università di Trieste, dove insegna Etica Pubblica. Si occupa in particolare dei risvolti morali dei processi decisionali e di percorsi decisionali di gruppo. È tra i fondatori dell'iniziativa “Parole Ostili” per la promozione di stili di comunicazione non violenti online. È autore di numerosi studi scientifici e saggi divulgativi in antropologia e filosofia morale.

**Michele Loi** ha conseguito il dottorato di ricerca in teoria politica alla LUISS Guido Carli nel 2007 ed è ricercatore senior presso l’*Institute for Biomedical Ethics and the History of Medicine* all’Università di Zurigo. Il suo ambito di specializzazione sono le teorie della giustizia, bioetica, etica dei dati e degli algoritmi. Michele Loi ha anche contribuito alla scrittura di codici e linee guida etiche, sia per istituzioni internazionali come l’Organizzazione Mondiale della Sanità, sia per privati.

**Federica Lucivero** è ricercatrice in etica e dati presso il centro *Ethox* e il *Wellcome Centre for Ethics and Humanities* al *Big Data Institute* dell'Università di Oxford. Ha svolto ricerche su diversi tipi di biotecnologie (interfacce cervello-computer, diagnostica molecolare, robot e *e-health*) e si è occupata a vario titolo di etica, studi sociali e filosofia della scienza e della tecnologia, di governance dell'innovazione e di bioetica empirica. La sua ricerca integra questioni teoriche e metodologiche con indagini empiriche.

**Francesca Musiani** è ricercatrice titolare (*chargée de recherche*) presso il *Centre National de la Recherche Scientifique* (CNRS) francese, co-fondatrice e vice-direttrice del *Centre Internet et Société*. Ha affiliazioni secondarie al *Centre de Sociologie de l’Innovation* di *MINES ParisTech,* ed all’*Internet Governance Lab* *dell’American University*. Le sue ricerche vertono sulla governance di Internet in una prospettiva STS. È co-presidente della sezione *Communication Policy & Technology* dell’IAMCR e *academic editor* per l’*Internet Policy Review*.

**Marcello Passarelli** è PhD in Psicologia, Antropologia e Scienze Cognitive. Dal 2017 lavora presso l'Istituto Tecnologie Didattiche del Consiglio Nazionale delle Ricerche (ITD-CNR), dove ha collaborato a progetti Europei. In particolare, è stato coinvolto nel progetto *H2020 Gaming Horizons*, che ha studiato l'impatto socio-psicologico dei videogiochi e della gamification. È attualmente coinvolto nel progetto Erasmus+ “*Playful Environment for Inclusive LeArning Design in Europe* (PLEIADE)”.

**Donatella Persico** è dirigente di ricerca presso ITD-CNR e direttrice dell’Italian Journal of Educational Technology. Le sue principali aree di interesse sono la progettazione didattica, l’autoregolazione dell’apprendimento, l’apprendimento collaborativo, il “*game based learning*” e i “*learning analytics*”. È stata responsabile di numerosi progetti di ricerca, nazionali e internazionali, e attualmente coordina il progetto PLEIADE.

**Teresa Scantamburlo** è assegnista di ricerca presso l'European Centre for Living Technology, Università Ca’Foscari di Venezia, dove coordina l’Osservatorio per la Società e l’Intelligenza Artificiale nel contesto del progetto europeo AI4EU. Dopo il dottorato in informatica si è occupata dell’impatto etico-sociale dell’IA svolgendo attività di ricerca presso l’università di Bristol (UK). Con Marcello Pelillo ha curato il volume *Machines we Trust. Perspectives on Dependable AI* per MIT Press.

**Viola Schiaffonati** è professoressa associata di logica e filosofia della scienza presso il dipartimento di elettronica, informazione e bioingegneria del Politecnico di Milano. I suoi principali interessi di ricerca riguardano le questioni filosofiche ed etiche dell’IA e della robotica e l’epistemologia e la metodologia degli esperimenti nell’ingegneria informatica.

**Guglielmo Tamburrini** è professore di filosofia della scienza e della tecnologia all’Università di Napoli Federico II. Si è occupato estesamente di filosofia degli automi: teoria della calcolabilità e problema mente-macchina, autonomia dei sistemi intelligenti ed etica, epistemologia di cibernetica, IA e robotica. Per Carocci ha pubblicato nel 2020 *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*.

**Luisa Zecca** è docente di Progettazione e valutazione di servizi e interventi educativi e di Mediazione didattica e strategie di gruppo. È membro del consiglio scientifico del Crespi, Centro di ricerca educativa sulla professionalità dell'insegnante. Si occupa di modelli di formazione degli insegnanti, di curricolo per competenze, di didattica laboratoriale e dell'educazione alla cittadinanza con particolare attenzione all'ambito della *Social Justice Education* nella scuola dell'infanzia e di base.

1. <http://www.roboethics.org/sanremo2004/> [↑](#footnote-ref-1)
2. [www.roboethics.org/atelier2006/docs/EthicbotsAtelier\_tamburrini.pdf](http://www.roboethics.org/atelier2006/docs/EthicbotsAtelier_tamburrini.pdf) [↑](#footnote-ref-2)
3. Sin dagli inizi il progetto di macchine intelligenti è considerato un’impresa complessa che riguarda l’automatizzazione di diverse abilità umane, come la capacità di dedurre delle conclusioni da principi generali (ragionamento deduttivo) e la facoltà di seguire regole e criteri per esplorare un certo numero di soluzioni tratte da un insieme più grande così da limitare la ricerca (la ricerca per euristiche). La deduzione e le euristiche sono problematiche affrontate soprattutto nei primi anni di sviluppo dell’IA e hanno trovato ampia applicazione nei giochi e nelle dimostrazioni di teoremi. Per una trattazione dei problemi affrontati dall’IA e una ricostruzione storica dei metodi proposti si veda Nilsson (2010). [↑](#footnote-ref-3)
4. Un altro problema simile alla classificazione è quello della regressione. La differenza tra i due è che nel primo caso lo scopo è predire una variabile discreta, ossia una classe o una categoria, mentre nel secondo caso l’obiettivo è predire una variabile continua, generalmente un numero reale. Un esempio di regressione è la predizione del prezzo di un immobile in base alla zona, l’anno di costruzione e i metri quadrati. [↑](#footnote-ref-4)
5. L’esempio è tratto da un saggio di Kuhn (1974, pp. 309-312), successivo alla sua maggiore opera “La struttura delle rivoluzioni scientifiche”, dove approfondisce il significato di paradigma e la sua intima relazione con la vita della comunità scientifica. In questo saggio, Kuhn mette in parallelo la capacità di acquisire esempi di problemi e soluzioni all’interno di una comunità scientifica e l’apprendimento dei concetti. Kuhn stesso descrive come il processo di apprendimento possa essere eseguito da una macchina e la descrizione offerta non si discosta molto dall'approccio del *machine learning.* [↑](#footnote-ref-5)
6. In statistica, gli attributi possono essere chiamati anche variabili indipendenti, mentre l’output da predire (la classe o il numero reale nel caso della regressione) può essere chiamato variabile dipendente. Si noti che la relazione funzione tra le variabili dipendenti e indipendenti non implica alcun nesso logico-causale. [↑](#footnote-ref-6)
7. I concetti di sensibilità e specificità sono comuni nei test clinici e si comprendono meglio nella classificazione binaria. Supponiamo di dover predire i pazienti malati su una popolazione. In questo caso, la sensibilità corrisponde alla percentuale di inferenze positive corrette (es. la proporzione di pazienti classificati correttamente come malati rispetto alla popolazione realmente malata), mentre la specificità si riferisce alle inferenze negative corrette (es. la proporzione di pazienti classificati correttamente come negativi rispetto alla popolazione realmente sana). L’accuratezza invece è una misura complessiva che riflette la proporzione di predizioni positive e negative corrette sul totale della popolazione. [↑](#footnote-ref-7)
8. Le competizioni sono eventi in cui algoritmi diversi vengono testati su particolari problemi e dati per metterne a confronto le *performance*. Per esempio, nell’ambito della visione artificiale, si vedano i risultati raggiunti dalle reti neurali “profonde” (*deep neural network*) nella competizione *ImageNet* (<http://www.image-net.org/challenges/LSVRC/>). Per una panoramica sulle performance raggiunte da alcuni popolari metodi diIA si vedano le seguenti risorse: Perrault *et al.* (2019) e la piattaforma AI Collaboratory: <http://dmip.webs.upv.es/AICollaboratory/> [↑](#footnote-ref-8)
9. Questi strumenti sono elaborati da un’area di studi dell’apprendimento automatico chiamata *Statistical Learning Theory*. Questa si occupa delle fondamenta teoriche dell’apprendimento e le condizioni che ne garantiscono la consistenza. Per un’introduzione accessibile si veda von Luxburg and Scholkopf (2009). [↑](#footnote-ref-9)
10. Da un punto di vista tecnico questo principio è implementato da una tecnica nota come *Collaborative Filtering*. Questo approccio elabora previsioni sulla base delle opinioni, preferenze e comportamenti di un gruppo di utenti ed è molto comune nei sistemi di raccomandazione. Per un’introduzione al tema si veda Schafer *et al*. (2007) [↑](#footnote-ref-10)
11. Per maggiori informazioni si veda il sito dell’azienda: <https://www.hirevue.com/> [↑](#footnote-ref-11)
12. Questi algoritmi si basano su metodi statistici e costituiscono la base delle scienze attuariali. [↑](#footnote-ref-12)
13. Un esempio classico di distribuzione identica e indipendente è quella generata dal lancio ripetuto di una monetina (non manipolata). In questo caso l’evento “testa” avrà la stessa probabilità dell’evento “croce” (50%) e l’uno sarà indipendente dall’altro, perché la monetina non ha “memoria” dell'ultimo lancio eseguito, quindi sarà non condizionato dagli eventi precedenti. [↑](#footnote-ref-13)
14. L’idea si basa sul concetto aristotelico di causa formale, ossia quando uno o più elementi costituiscono la forma o l’essenza di una determinata cosa. [↑](#footnote-ref-14)
15. Sull’approccio essenzialista nel *machine learning* si veda Pelillo e Scantamburlo (2013). [↑](#footnote-ref-15)
16. L’*overfitting* può verificarsi quando il training set è piccolo e le correlazioni individuate nella fase di addestramento sono frutto di coincidenze o significative solo per quel campione di dati. [↑](#footnote-ref-16)
17. L'apprendimento può anche essere visto come un processo di esplorazione di uno spazio che contiene diverse funzioni ciascuna corrispondente ad un’ipotesi sull'identità della classe da predire. Questo spazio potrebbe essere molto esteso e per questo è importante che il programmatore definisca una preferenza, “forzando” così l’algoritmo a cercare in una porzione di spazio. In filosofia della scienza questa prospettiva ha suggerito un’analogia tra il *machine learning* e la scoperta e la valutazione delle teorie scientifiche (Thagard, 1988). [↑](#footnote-ref-17)
18. Per verificare se la classe di funzioni preferite è corretta sarà importante guardare all’errore commesso in fase di addestramento. Se questo sarà limitato, ci si aspetta che la funzione da apprendere sia “contenuta” nello spazio di funzioni preferite e che l’errore continuerà ad essere limitato anche con dati sconosciuti. Per una discussione dell’*inductive bias* si veda Mitchell (1996). [↑](#footnote-ref-18)
19. La difficoltà nell’interpretare la logica dell’algoritmo è una sfida per l’applicazione del Regolamento Generale per la Protezione dei Dati (si vedano articoli 13-15). [↑](#footnote-ref-19)
20. Per una spiegazione rigorosa ma accessibile alle reti neurali profonde si veda Codenotti e Leoncini (2020) o Mitchell (2019). [↑](#footnote-ref-20)
21. L’effetto discriminatorio degli algoritmi è legato al concetto di *bias*, un termine inglese che può avere diversi significati a seconda del contesto considerato. Nel dibattito contemporaneo sull’impatto dell’IA il *bias* è inteso come una preferenza ingiustificata dell’algoritmo verso una determinata categoria di soggetti che può sfociare in una discriminazione sociale. Ma nell’apprendimento automatico il termine ha una connotazione tecnica e, come visto nel paragrafo 2.3, esso fa parte delle assunzioni base dell’inferenza statistica e identifica una componente essenziale al funzionamento dell’algoritmo (Hildebrandt, 2021). In questo paragrafo per evitare la confusione tra le diverse accezioni preferiamo parlare semplicemente di pregiudizio. [↑](#footnote-ref-21)
22. La psicometria è l’area della psicologia che si occupa di valutare in modo quantitativo fenomeni psicologici come l’intelligenza e la personalità. Il test utilizzato da Cambridge Analytica si ispira al modello delle *Big Five*, che classifica la personalità in 5 tratti fondamentali: estroversione, amicalità, coscienziosità, nevroticismo, apertura mentale. Per una rassegna degli studi empirici su come inferire qualità psicologiche dalle tracce digitali lasciate dagli utenti su Internet si veda Burr e Cristianini (2019) [↑](#footnote-ref-22)
23. OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. [↑](#footnote-ref-23)
24. https://op.europa.eu/it/publication-detail/-/publication/ac957f13-53c6-11ea-aece-01aa75ed71a1 [↑](#footnote-ref-24)
25. https://read.oecd-ilibrary.org/social-issues-migration-health/health-data-governance\_9789264244566-en#page1 [↑](#footnote-ref-25)
26. https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf [↑](#footnote-ref-26)
27. I *Google car* sono stati coinvolti in vari sinistri e i veicoli autonomi della Tesla anche in incidenti mortali. Uno di questi è stato originato da condizioni di illuminazione poco favorevoli per l’occhio umano, ma particolarmente avverse per il modulo percettivo dell’autoveicolo. E il collaudatore a bordo di un’automobile a guida autonoma della società Uber non è riuscito a riprenderne il controllo prima che questi investisse e uccidesse una ciclista a Tempe, in Arizona. [↑](#footnote-ref-27)
28. Entrambi gli scenari somigliano all’esperimento mentale del carrello ferroviario (*trolley problem*, cfr. Foot, 1967; Tafani, 2020). Ci sono però differenze cruciali tra il *trolley problem* e le collisioni inevitabili, tra cui la ripetizione della scelta morale, che non si pone nel caso del problema tradizionale del carrello, gli effetti presumibili della ripetizione sulle scelte successive, nonché la dimensione sociale, anziché puramente individuale, delle scelte che riguardano la programmazione dei VA (Nyholm, Smids, 2016). Per un repertorio variegato di scenari di collisione inevitabile per VA, cfr. Maurer *et al.*, 2015). [↑](#footnote-ref-28)
29. Per retribuzione morale si intende la possibilità di attribuire con giusta causa la responsabilità di un danno di carattere morale a qualcuno. Il processo di attribuzione di responsabilità è complesso e anch’esso passibile di analisi morale. Secondo alcuni (Danaher, 2016), attribuire la responsabilità morale in modo corretto è un’esigenza importante che concorre a mantenere sana e coesa la società. L’autonomia operativa dei sistemi tecnologici, complicando ulteriormente un processo già di per sé piuttosto delicato, rappresenta quindi un’importante sfida da gestire adeguatamente. [↑](#footnote-ref-29)
30. Semplificando, per paternalismo morale si intende la pratica di decidere per altri cosa è meglio per loro contro la loro volontà, pur godendo costoro in principio del pieno diritto di deciderlo in autonomia. Per una discussione più estesa si veda ROBOTICA SOCIALE pp….. [↑](#footnote-ref-30)
31. Di fatto mantenuto in inglese nell’ambito della lingua italiana; parole equivalenti sono invece in uso in altre lingue, come il francese *gouvernance* e lo spagnolo *gobernanza*. [↑](#footnote-ref-31)
32. Il capitolo riprende ed aggiorna alcuni scritti in inglese ed in francese, in particolare (Musiani, 2013; DeNardis, Musiani, 2016; Musiani, 2018). [↑](#footnote-ref-32)
33. Questa prospettiva viene invece spesso adottata nei media e nel dibattito pubblico. [↑](#footnote-ref-33)
34. https://www.torproject.org [↑](#footnote-ref-34)
35. Un caso del genere è testimoniato da Benjamin Kunz Mejri nel MOOC del progetto CANVAS sull’etica della sicurezza cyber, accessibile all’indirizzo https://canvas-project.eu/results/mooc.html. [↑](#footnote-ref-35)
36. La Carbon Footprint è una misura utilizzata per quantificare l’impatto ambientale di un prodotto, un processo, un’organizzazione o un individuo e corrisponde alla quantità (in tonnellate) di emissioni di gas serra associate all’attività in questione. È espressa in tonnellate di CO2 equivalente (tCO2e), che prende come riferimento l’effetto serra associato alla CO2, considerato pari a 1. Le emissioni di altri gas vengono contabilizzate rispetto a questo parametro <https://www.minambiente.it/pagina/cose-la-carbon-footprint>; <http://lca.jrc.ec.europa.eu/Carbon_footprint.pdf>; consultati il 18 febbraio 2021) [↑](#footnote-ref-36)
37. Questa considerazione rimanda al principio di precauzione, che è centrale nella riflessione ambientalista. Secondo la formulazione della Dichiarazione di Rio (1992) questo approccio ha lo scopo di proteggere l’ambiente e consiste nell’adottare, in caso di rischio di danno grave o irreversibile, misure cautelative per prevenire il degrado ambientale, anche in assenza di un consenso della comunità scientifica sull’entità dei rischi: “Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation”. Il principio viene spesso criticato perché in una sua formulazione più rigida richiederebbe una prova certa dell’assenza di rischio prima dell’approvazione di qualunque innovazione, causando così un vistoso rallentamento del progresso scientifico-tecnologico; tuttavia, in una sua formulazione più morbida, esso sottolinea la responsabilità per Stati e parti interessate di valutare i possibili rischi e costi per l’ambiente e le generazioni future. Per un approfondimento sul principio di precauzione si veda Hanson (2018). [↑](#footnote-ref-37)
38. Nonostante non ci sia consenso sul fatto che l’attività umana sia l’unica causa del riscaldamento globale, il rapporto pubblicato dall’*Intergovernmental Panel on Climate Change* (IPPC) nel 2013 stabilisce che è estremamente probabile che il cambiamento climatico sia antropogenico (causato dall’attività umana). [↑](#footnote-ref-38)
39. Un esempio di questo approccio è quello adottato da alcuni movimenti politici negli anni ’60 e ’70, in cui motivazioni ambientaliste erano associate al rifiuto dell’ideologia capitalista e degli eccessi della rivoluzione industriale (Smith, 2003). [↑](#footnote-ref-39)
40. Gli scienziati ambientali si riferiscono alle metodologie che valutano gli impatti ambientali associati ai diversi stadi del ciclo di vita di un servizio, processo o prodotto come “*Life Cycle Assessment*” (LCA) o “*cradle-to grave*”. [↑](#footnote-ref-40)
41. È opportuno sottolineare che esistono diversi parametri per valutare l’impatto ambientale. Per esempio, insieme alla *carbon footprint* (discussa nella nota 1), l’*ecological* *footprint* è un altro parametro per misurare l’impatto delle attività umane sull’ambiente. Se la *carbon footprint* è una misura del peso (in tonnellate o chilogrammi) delle emissioni di gas responsabili del riscaldamento globale, l’*impronta* *ecologica* (o ambientale) ha un significato più letterale, in quanto è un indicatore di spazio che converte il consumo di risorse naturali e la produzione di risorse in unità di misura di superficie. Gli ettari globali (gha) esprimono la quantità di spazio ambientale (area biologicamente produttiva di mare e di terra) che sarebbe necessaria perché tutta la popolazione mondiale mantenesse un determinato stile di vita. È stato calcolato che ci vorrebbero cinque “pianeti Terra” per sostenere lo stile di vita dell’individuo medio statunitense, tre e mezzo per sostenere uno stile di vita italiano e mezzo per sostenere uno stile di vita indiano (<https://data.footprintnetwork.org/#/>?; consultato il 18 febbraio 2021). [↑](#footnote-ref-41)
42. United States of America Environmental Protection Agency (https://www.epa.gov/environmentaljustice; consultato il 19 febbraio 2021) [↑](#footnote-ref-42)
43. <http://www.greenpeace.org/usa/globalwarming/click-clean/>. [↑](#footnote-ref-43)
44. Per quanto riguarda i codici di autoregolamentazione degno di nota e’ quello *dell’International Telecommunication Union* (ITU) (https://www.itu.int/en/mediacentre/Pages/PR04-2020-ICT-industry- to-reduce-greenhouse-gas-emissions-by-45-percent-by-2030.aspx). Per quanto riguarda le normative europee, lo *European Green Deal* in parte discute la necessità di ridurre le emissioni dei CED (https://ec.europa.eu/info/sites/info/files/european-green-deal- communication\_en.pdf). Le questioni relative alle politiche di economia circolare per i rifiuti elettronici sono trattate nel documento: *European Commission, 2020d. Changing how we produce and consume: New Circular Economy Action Plan shows the way to a climate-neutral, competitiveeconomy of empowered consumers.*

(https://ec.europa.eu/commission/presscorner/detail/en/ip\_20\_420). [↑](#footnote-ref-44)
45. https://www.youtube.com/watch?v=aFuA50H9uek [↑](#footnote-ref-45)
46. https://newzoo.com/insights/rankings/top-10-countries-by-game-revenues/ [↑](#footnote-ref-46)
47. https://newzoo.com/insights/articles/newzoos-top-100-countries-by-2015-game-revenues/ [↑](#footnote-ref-47)
48. <https://www.digitalbros.com/il-gruppo/il-mercato-dei-videogiochi/?lang=it#:~:text=Il%20fatturato%20mondiale%20dei%20software,per%20console%20(%2445%2C2> [↑](#footnote-ref-48)
49. https://hopelab.org/product/re-mission/ [↑](#footnote-ref-49)
50. https://www.game-learn.com/game-based-learning-corporate-training/serious-game-negotiation-and-conflict-resolution/ [↑](#footnote-ref-50)
51. Giochi, prevalentemente di guerra, in cui il giocatore comanda un personaggio in prima persona e in cui la principale attività è sparare ad avversari virtuali. [↑](#footnote-ref-51)
52. Il fenomeno della tolleranza si verifica quando, dopo ripetute somministrazioni di una sostanza, questa produce un effetto inferiore rispetto a quello iniziale, o, in altre parole, quando si debba aumentare la dose della sostanza per avere gli stessi effetti esperiti in precedenza. [↑](#footnote-ref-52)
53. Giochi in cui il giocatore impersona un personaggio in un mondo virtuale *multiplayer* di grandi dimensioni [↑](#footnote-ref-53)
54. https://www.techradar.com/news/gaming/blizzard-discourages-wow-poopsocking-486618 [↑](#footnote-ref-54)
55. http://gameaccessibilityguidelines.com/full-list/ [↑](#footnote-ref-55)
56. https://www.md-subs.com/what-game-subs-got-wrong-in-2017 [↑](#footnote-ref-56)
57. Giochi che presentano sfide principalmente legate alla navigazione di uno spazio 2D o 3D. Esempi sono *Super Mario* o *Crash Bandicoot*. [↑](#footnote-ref-57)
58. https://www.forbes.com/sites/alexkonrad/2014/02/18/language-learning-app-duolingo-raises-20m-in-race-to-teach-english/ [↑](#footnote-ref-58)
59. http://valuesatplay.org/ [↑](#footnote-ref-59)
60. La nascita dei primi codici professionali in epoca moderna può essere fatta risalire al periodo successivo alla Seconda Guerra Mondiale, quando furono formulati per esempio i primi codici professionali degli ingegneri per cercare di riabilitare l’immagine sociale della scienza e della tecnologia uscita in parte compromessa dal conflitto. [↑](#footnote-ref-60)
61. <https://www.microsoft.com/en-us/legal/compliance/default.aspx> [↑](#footnote-ref-61)
62. <https://www.ibm.com/partnerworld/program/code-of-conduct> [↑](#footnote-ref-62)
63. Per la disamina dei cambiamenti occorsi si vedano Brinkman *et al*., 2016 e Gotterbarn *et al.,* 2018.. [↑](#footnote-ref-63)
64. Per ulteriori dettagli si veda <https://www.acm.org/code-of-ethics>. [↑](#footnote-ref-64)
65. <https://ethics.acm.org/integrity-project/ask-an-ethicist/> [↑](#footnote-ref-65)
66. Per ulteriori dettagli si può consultare <https://ethicsinaction.ieee.org/> [↑](#footnote-ref-66)
67. <https://standards.ieee.org/industry-connections/ecpais.html> [↑](#footnote-ref-67)
68. <https://www.partnershiponai.org/tenets/> [↑](#footnote-ref-68)
69. <https://www.montrealdeclaration-responsibleai.com/the-declaration> [↑](#footnote-ref-69)
70. <https://futureoflife.org/ai-principles/> [↑](#footnote-ref-70)
71. [https://www.eismd.eu/wp-content/uploads/2019/11/AI4PeoplÈs-Ethical-Framework-for-a-Good-AI-Society\_compressed.pdf](https://www.eismd.eu/wp-content/uploads/2019/11/AI4People%E2%80%99s-Ethical-Framework-for-a-Good-AI-Society_compressed.pdf) [↑](#footnote-ref-71)
72. <https://www.baai.ac.cn/blog/beijing-ai-principles> [↑](#footnote-ref-72)
73. <https://fondazioneleonardo-cdm.com/site/assets/files/2553/fle1_booklet_conferenza_ita_ibm_111119.pdf> [↑](#footnote-ref-73)
74. <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>); [↑](#footnote-ref-74)
75. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> [↑](#footnote-ref-75)
76. https://ia.italia.it/assets/librobianco.pdf [↑](#footnote-ref-76)
77. Oltre ai testi citati, su ciò si veda il comodo strumento sviluppato dall’AI Ethics Lab (http://aiethicslab.com/big-picture/). [↑](#footnote-ref-77)
78. <https://www.ai4eu.eu> [↑](#footnote-ref-78)
79. Per alcuni esempi sulle diverse classi di agenti morali rimandiamo al saggio di Moor. [↑](#footnote-ref-79)