I pregiudizi: un errore solo umano? Come i pregiudizi accomunano umani e algoritmi

Fiorella Battaglia

1. Gli atteggiamenti pregiudiziali sono molto diffusi tra i gruppi umani, con rilevanti conseguenze in termini di discriminazioni e di ingiustizia[1]. Il successo delle *Giornate di studio sul razzismo* organizzate dall'Università del Salento e arrivate alla loro quarta edizione testimonia dell'impegno della comunità accademica nel combattere un fenomeno che erode i fondamenti della convivenza e della democrazia[2]. Da quando aziende e governi hanno iniziato a rivolgersi sempre più spesso agli algoritmi di apprendimento automatico per prendere decisioni cruciali, come ad esempio riguardo a chi assumere, o a chi elargire un'assicurazione o un mutuo o a chi deve ricevere benefici governativi e perfino a quale detenuto concedere la libertà vigilata casi di *bias* e di diffusione di pregiudizi mediati da sistemi autonomi sono diventati all'ordine del giorno. Conseguentemente si è aperto un dibattito su diversi fronti. Si indaga non solo sul valore epistemico di tali decisioni ma anche sui profili di ineguaglianza che questi nuovi strumenti pongono in essere, approfondiscono e in alcuni casi creano *ex novo*. In particolare ha anche ricevuto nuovo impulso la ricerca sul carattere stesso dei pregiudizi con domande su se possiamo tracciare dei confini tra ineguaglianze sorte e alimentate in contesti umani, ibridi o solo macchinici.

Questo articolo esamina le preoccupazioni concernenti la diffusione di pregiudizi veicolati dai sistemi di decisione automatizzati, presenta poi i risultati di uno studio del MIT di Boston e dell'Università di Cardiff che ha indagato come sorgono i pregiudizi nella dinamica della comunicazione e tracciato i confini tra errori umani ed errori compiuti dalle macchine. Questo studio ha mostrato una genesi diversa degli errori specifici dei sistemi e degli errori che ripetono quelli compiuti dagli esseri umani.

A partire da alcuni spunti di riflessione sulla natura e sulle dinamiche dei pregiudizi, sosterrò che l'apprendimento automatico del comportamento morale è minacciato (a) dall'essere soggetto a errori, perpetuazione dei pregiudizi e creazione di nuovi pregiudizi; (b) dall'insistenza su un paradigma di apprendimento opaco e

[1] M. Fricker, Epis*temic Injustice. Power and the ethics of knowing*, Oxford University Press, 2007. Vedi anche D. L. Smith, *Making Monsters. The Uncanny Power of Dehumanization*, Harvard University Press, 2021.

[2] T. Christiano, *Democracy*, Stanford Encyclopedia of Philosophy, 2016, http://plato.stanford.edu/entries/democracy/, consultato il 30.01.2023. Vedi anche N. Urbinati, *Democrazia sfigurata. Il popolo fra opinione e verità*, Università Bocconi editore, 2014.

dall'assenza di un sistema di spiegazione; e che (c) una tale comprensione normativa del comportamento e delle pratiche morali esclude i sistemi di apprendimento automatico dal poter essere considerati modelli appropriati di esempio morale e richiede che il processo di apprendimento sia prodotto da sistemi socio-tecnici costituiti sia da esseri umani sia da artefatti. Concluderò con alcune lezioni apprese che evidenziano quali sono le esigenze epistemiche e morali necessarie sia per lo sviluppo di tecnologie degne di configurare dei fidi maestri sostituti sia per i dibattiti pubblici su tali sistemi. La conclusione è che vengono richiesti doveri epistemici e morali a tutti gli attori coinvolti siano essi macchinici o umani.

2. Si potrebbe pensare che il pregiudizio sia un fenomeno legato al modo di pensare e di agire degli esseri umani, che richiede alti livelli di cognizione umana per formare un'opinione o uno stereotipo su una particolare persona o gruppo. Recentemente questa convinzione è stata messa in discussione[3]. A parziale spiegazione si deve considerare che il modo di rappresentare il senso comune è ancora molto deficitario e conduce a risultati che ai nostri occhi appaiono paradossali[4]. In particolare, a dispetto dei differenti tentativi intrapresi per fornire una teoria generale del comportamento morale, si deve ammettere che la morale non è né utilitaristica, né deontologica, ma coinvolge principi deontologici e varie euristiche[5]. Inoltre, le regole e i principi morali sono spesso vaghi e dipendenti dall'ambiente e possono essere in conflitto tra loro[6]. Lo studio citato di Whitaker, Colombo e Rand si è focalizzato sulla genesi dei pregiudizi. Esso ha messo in luce come i robot siano, proprio come gli umani, vittime dei pregiudizi.

Occorre tenere presente che l'approccio basato sul training sui big data è una soluzione adottata per superare i problemi connessi all'assenza di un modello teorico che non fosse controverso. Un approccio in termini di allenamento prevede l'applicazione di tecniche come l'apprendimento automatico per "educare" e "formare" un sistema di intelligenza artificiale a riconoscere le situazioni difficili e a risolvere i conflitti. Questo approccio si ispira a quanto avviene nella vita reale quando gli esseri umani più giovani e inesperti vengono presi in carico dagli adulti che li formano ed educano a diventare attori e attrici morali anche facendo

[3] R. M. Whitaker, G. B. Colombo, D. G. Rand, *Indirect Reciprocity and the Evolution of Prejudicial Groups*, in «Sci Rep», 2018, vol. 8, n. 13247.

[4] W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, 2008.

[5] D. Kahneman, S. Frederick, *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, 2002.

[6] Il tentativo più convincente di dare una rappresentazione unitaria della morale nonostante le diverse sfide all'interno della filosofia morale è fornito da John Doris. Vedi J. M. Doris, *Character Trouble: Undisciplined Essays on Moral Agency and Personality*. Oxford University Press 2022; J. M. Doris, *Talking to Our Selves: Reflection, Ignorance, and Agency*, Oxford University Press, 2015.

riferimento al valore educativo dell'esempio. Due progressi nel campo dell'informatica hanno reso possibile questo passaggio e lo sviluppo di quest'approccio: la disponibilità di una grande quantità di dati (*Big Data*) e l'aumentata potenza di calcolo. Lo schema anche nel caso del comportamento morale non è quindi diverso da quello applicato con successo in altri campi[7]. Concepito in questo modo l'apprendimento di un comportamento morale da parte del sistema è molto simile al modo in cui un sistema di apprendimento bayesiano può essere addestrato a riconoscere cellule tumorali o altri modelli salienti nelle immagini mediche. Le tecniche di apprendimento profondo basate sulle reti neurali riescono a identificare cellule tumorali e loro mutazioni con velocità ed efficienza molto superiori a quelle di un operatore umano, permettendo precisione della diagnosi e accelerazione della ricerca[8].

Allenare i sistemi automatici in rete allora non solo riceve forza dal successo dimostrato in altri campi, ma promette anche di superare i limiti dettati dal fatto che non abbiamo una teoria morale da poter formalizzare. Un database di dimensioni sufficientemente grande di situazioni esemplificative di problemi decisionali morali illustrativi potrebbe costituire l'unica condizione necessaria da soddisfare. Questi i punti di forza di quest'approccio. I punti di debolezza sono purtroppo di due tipi entrambi abbastanza preoccupanti. Innanzitutto i sistemi imparano anche gli errori[9] e sono soggetti a ripeterli senza essere in grado di fornire spiegazioni rispetto alle ragioni che li hanno spinti ad agire così[10].

I sistemi automatici che vengono addestrati in rete e sui social media non possono fare a meno di apprendere comportamenti discriminatori. E perciò nella fase di *training* il sistema prende esempi da dati forniti che sono affetti da distorsioni, errori e mancanza di risorse interpretative per comprendere le ingiustizie compiute ai danni di altri individui o gruppi. Questa è infatti la qualità dei dati che gli umani mettono a disposizione. Lo stesso esperimento ha anche dimostrato che le intelligenze artificiali possono acquisire sia i "bias" dei programmatori sia anche svilupparne in autonomia, con i *chatbot* che imparano/disimparano l'uno dall'altro. Secondo il team di scienziati "gruppi di macchine autonome possono manifestare pregiudizi semplicemente identificando, copiando e imparando questo

---

[7] M. Fisher, C. List, M Slavkovik, A. Winfield, *Engineering Moral Agents - from Human Morality to Artificial Morality*, in «Dagstuhl Seminar», 2016, n. 16222, DOI: 10.4230/DagRep.6.5.114.

[8] Vedi per esempio, N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, *Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC),* arXiv, 1710.05006v3.

[9] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Largo, ML: Crown, 2016.

[10] T. Scanlon, *Being Realistic about Reasons*, Oxford University Press, 2014. Sul tema dell'explanable AI, si veda il progetto ERC di Fosca Giannotti https://xai-project.eu/, consultato il 30.01.2023.

atteggiamento da un'altra macchina"[11]. Le applicazioni matematiche per elaborare gli insiemi di dati si basano su scelte fatte da esseri umani fallibili. Questi database contengono ogni tipo di decisioni e mentre alcune decisioni sono state motivate senza dubbio dalle migliori intenzioni altre invece no. Il problema insorge, perché molti di questi modelli hanno codificato pregiudizi umani e incomprensioni nei sistemi software che sempre più gestiscono le nostre vite.

I *bias* cognitivi sono costrutti fondati, al di fuori del giudizio critico, su percezioni errate o deformate, su pregiudizi e ideologie; utilizzati spesso per prendere decisioni in fretta e senza fatica[12]. Lo psicologo israeliano Kahneman ha lavorato a lungo a una teoria che gettasse luce sulla natura e sul ruolo dell'euristica[13]. Assieme ad altri ricercatori ha teorizzato il modo di funzionare dell'euristica cognitiva. Pare che l'euristica operi per mezzo di un sistema chiamato sostituzione dell'attributo, che avviene senza consapevolezza. Secondo questa teoria, un giudizio complesso da un punto di vista inferenziale può essere sostituito da un'euristica che è un concetto affine a quello precedente, ma formulato più semplicemente. Le euristiche sono, dunque, delle scorciatoie mentali che portano a conclusioni veloci con il minimo sforzo cognitivo. Sulla base di questo resoconto generale possiamo inquadrare concettualmente cosa sono i bias: essi sono particolari euristiche usate per esprimere dei giudizi, solo che alla lunga diventano pregiudizi, su cose mai viste o di cui non si è mai avuto esperienza. Rispetto al contenuto, possiamo dire che in generale quando i giudizi sono veri allora le euristiche funzionano come una scorciatoia mentale e permettono di avere un accesso rapido a informazioni immagazzinate in memoria. In sintesi, se le euristiche rimangono fedeli alle richieste di verità allora si dimostrano essere delle scorciatoie comode e rapide estrapolate dalla realtà che portano a veloci conclusioni. Diverso è il discorso per i bias cognitivi. Questi sono euristiche inefficaci, perché non rispettano i criteri di verità, ma configurano piuttosto pregiudizi astratti che non provengono da dati di realtà, ma si recepiscono a priori senza esercitare nessun tipo di giudizio critico.

Ciò di cui sarebbero capaci gli algoritmi informatici sono delle euristiche inefficaci. In termini nono solo cognitivi ma più moralmente connotati essi si sono dimostrati dei fedeli ripetitori di pregiudizi come il razzismo e il sessismo sulla base dell'apprendimento di dati pubblici e di altri dati generati dalle comunicazioni umane. L'apporto di conoscenza di questo nuovo lavoro mostra la possibilità che l'intelligenza artificiale evolva in modo indipendente producendo nuovi gruppi di pregiudizi. Ci sono già state delle evidenze drammatiche di questa evoluzione. Per esempio, nel 2016 *Microsoft* ha dovuto togliere dalla rete il suo chatbot (software per

---

[11] R. M. Whitaker, G. B. Colombo, D. G. Rand, *Indirect Reciprocity and the Evolution of Prejudicial Groups*, in «Sci Rep», 2018, vol. 8, n. 13247.

[12] D. Kahneman, Amos Tversky, Paul Slovic, *Judgment under Uncertainty. Heuristics and Biases*, Cambridge University Press, 1982.

[13] D. Kahneman, *Pensieri lenti e veloci*, Mondadori, 2019.

simulare una conversazione intelligente come Siri o Cortana) perché si era rivelato un ripetitore dei commenti razzisti, sessisti e xenofobi che si trovano in rete[14]. Il robot Tay (*Thinking about you*) dotato di intelligenza artificiale lanciato su Twitter e altre piattaforme social avrebbe dovuto avviare e sostenere conversazioni con i giovani americani dai 18 ai 24 anni. Ma l'esperimento, che aveva come obiettivo attirare un maggior numero di ragazzi all'universo Microsoft, è fallito. Tay doveva imparare ripetendo e poi rispondere con le sue frasi, cercando di simulare una conversazione normale. L'esordio su Twitter prometteva bene ma è durato meno di 24 ore. L'esperimento oltre ai limiti tecnologici ha anche avuto il merito di mettere in luce i limiti umani e di dare nuovo impulso alle ricerche su ingiustizia e lotta alle ineguaglianze.

Il merito dei ricercatori di informatica e psicologia dell'Università di Cardiff e del MIT è stato quello di dimostrare come gruppi di macchine autonome possono mostrare pregiudizi semplicemente riconoscendo, copiando e imparando gli uni dagli altri questo comportamento, senza la necessità dell'intervento umano. Altri studi avevano già dimostrato che i sistemi autonomi possono apprendere pregiudizi come il razzismo e il sessismo sulla base di dati generati dalla comunicazione umana, questo nuovo lavoro mostra la possibilità di un'evoluzione dei pregiudizi indipendentemente dal contributo umano. I risultati sono stati pubblicati sulla rivista *Scientific Reports*. La metodologia fa ricorso a simulazioni al computer che mostrano come individui o agenti virtuali con pregiudizi simili possano creare un gruppo che esibisce un'interazione interna tendente a escludere individui appartenenti ad altri gruppi. Uno degli autori della simulazione, il professor Roger Whitaker dell'*Institute of Crime and Security Research* e della *School of Computer Science and Informatics* dell'Università di Cardiff, ha dichiarato: "Eseguendo queste simulazioni migliaia e migliaia di volte, iniziamo a capire come si sviluppa il pregiudizio e quali condizioni lo incoraggiano o lo ostacolano"[15]. I risultati hanno anche fornito delle informazioni rispetto alle dinamiche che informano le scelte dei singoli bots. In generale i comportamenti che ottimizzano gli interessi individuali sono quelli preferiti. Un'implicazione importante è che queste scelte non richiedono necessariamente abilità cognitive avanzate. "È ipotizzabile che macchine autonome con la capacità di identificarsi con discriminazione e di copiare gli altri possano in futuro essere suscettibili di fenomeni di pregiudizio che osserviamo nella popolazione umana", ha specificato il professor Whitaker.

Molti degli sviluppi dell'intelligenza artificiale a cui stiamo assistendo

[14] H. Reese, *Why Microsoft's 'Tay' AI bot went wrong,* in «Tech Republic», 24 marzo 2016, https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/, consultato 18.12.2022.
[15] Simone Cosimi, *Ai, i bot sviluppano pregiudizi anche da soli*, Repubblica tecnologia, https://www.repubblica.it/tecnologia/2018/09/07/news/ai_i_bot_sviluppano_pregiudizi_anche_da_soli-205819609/, consultato 18.12.2022.

> riguardano l'autonomia e l'autocontrollo, il che significa che il comportamento dei dispositivi è influenzato anche dal comportamento degli altri che li circondano. Il nostro studio fornisce una visione teorica della situazione in cui gli agenti simulati chiedono periodicamente agli altri un qualche tipo di risorsa.

Dallo studio è emersa un'altra conclusione interessante: all'interno di una popolazione in cui ci sono sottopopolazioni più diverse, è più difficile sviluppare pregiudizi.

> Con un numero maggiore di sottopopolazioni, le alleanze di gruppi che non hanno pregiudizi possono lavorare insieme senza essere sfruttate. Questo riduce anche il loro status di minoranza, riducendo la vulnerabilità allo sviluppo di pregiudizi. Tuttavia, ciò richiede anche una maggiore disponibilità degli attori a interagire al di fuori del proprio gruppo

ha spiegato il professor Whitaker. Questa conclusione assieme alla constatazione che per sviluppare pregiudizi non occorre avere delle capacità cognitive molto evolute in quanto anche delle semplici intelligenze artificiali ne sono capaci sono foriere di indicazioni per le buone pratiche da implementare anche nel mondo *offline*.

3. Il secondo problema che pesa su questi sistemi di apprendimento automatico riguarda i requisiti epistemici della conoscenza che producono. Questi modelli matematici sono opachi, il loro funzionamento è invisibile a tutti[16]. I loro verdetti, anche quando sono sbagliati o dannosi, non possono essere contestati o impugnati. Essi tendono a punire i poveri e gli oppressi della nostra società, rendendo i ricchi più ricchi e i poveri più poveri. E inoltre assomigliano sempre più a modelli di scienza e di razionalità che non sono sopravvissuti. Judea Pearl trova una somiglianza tra i sistemi opachi degli algoritmi attuali e i sistemi di predizione dei sacerdoti babilonesi[17]. L'esperimento del MIT di Boston e dell'Università di Cardiff ha tracciato i confini tra errori umani ed errori compiuti dalle macchine. Tali errori sono in parte specifici dei sistemi e in parte ripetono quelli compiuti dagli esseri umani. In generale, infatti, il modo di rappresentare il senso comune è ancora molto deficitario e conduce a risultati, che ai nostri occhi appaiono paradossali. Inoltre uno studio recente ha messo in luce come i robot siano, in questo caso proprio come gli umani, vittime dei pregiudizi. I sistemi che completano la loro educazione in rete e

---

[16] Sul tema dell'opacità epistemica, si veda P. W. Humphreys, *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press, 2004 e J. M. Durán, N. Formanek, *Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism*. in «Minds & Machines», 2018, vol. 28, pp. 645–666, https://doi.org/10.1007/s11023-018-9481-6, consultato il 30.01.2023.

[17] J. Pearl, *The limitations of opaque learning machines*, in J. Brockman (a cura di), *Possible minds. 25 ways of looking at AI*, Penguin Press, 2019.

sui social media non possono fare a meno di apprendere comportamenti discriminatori. E perciò nella fase di *training* il sistema prende esempi da dati forniti dagli umani che sono distorti. Lo stesso esperimento ha anche dimostrato che le intelligenze artificiali possono acquisire sia i *bias* dei programmatori sia anche svilupparne in autonomia, con i chatbot che imparano/disimparano l'uno dall'altro. Secondo il team di scienziati "gruppi di macchine autonome possono manifestare pregiudizi semplicemente identificando, copiando e imparando questo atteggiamento da un'altra macchina". Vi è la necessità di un'integrazione di nuove scoperte nelle pratiche umane di comunicazione senza restringere l'attenzione a tal punto che l'e*xplanandum* iniziale - la nostra autocomprensione umana - si perda di vista. Cinque domande ci possono aiutare a criticare l'implementazione della moralità artificiale.

1. Le credenze morali sostanziali e il loro rapporto con la sensibilità morale umana richiedono di essere studiate per sé stesse o la teoria morale è secondaria rispetto ad altre teorie filosofiche?

2. Qual è il ruolo costitutivo del dare e dell'assumere ragioni per l'interazione, la comunicazione e l'agente morale umano in generale?

3. In che senso l'articolazione di una prospettiva in seconda persona è fondamentale per un'interpretazione di ciò che significa essere un agente morale umano? E cosa implica per una prospettiva non naturalistica dell'*agency*?

4. In particolare, quali sono le sfide per le letture riduzionistiche delle strutture morali?

5. Come si possono superare le carenze dell'attuale filosofia della mente riduzionista per raggiungere un'interpretazione più completa dell'agente morale nel quadro delle nuove direzioni della filosofia della mente?

Un'ulteriore riflessione procede dal campo della metaetica. Prendendo spunto dai recenti successi dell'IA, è stata sostenuta una nuova articolazione del fisicalismo morale. Mentre il naturalismo morale, cioè la tesi secondo la quale esistono fatti e proprietà morali oggettivi e che questi fatti e proprietà morali sono naturali, gioca un ruolo fondamentale nel dibattito sul fisicalismo, cioè la tesi secondo la quale tutti i fatti sono fatti fisici, solo di recente e in seguito al successo delle indagini neuroscientifiche, solo di recente, e in seguito al successo dei sistemi di IA, nuovi argomenti sono entrati nel dibattito per sostenere la tesi del naturalismo morale e del fisicalismo.

Questo punto di vista non sarebbe influenzato dalle obiezioni consuete al fisicalismo. Pertanto, questo punto di vista deve essere affrontato in modo specifico. Le conseguenze sono molteplici, sia di natura pratica sia teorica. Esse vanno dalla possibilità sostanziale di implementare un codice morale nelle macchine che le renda in grado di decidere, nel peggiore dei casi, sulla vita e sulla morte, fino alla critica delle argomentazioni filosofiche generali. In quest'ottica, ciò di cui abbiamo bisogno per la costruzione di macchine autonome che potrebbero chiarire e, in ultima analisi, risolvere i problemi morali diventa un problema della macchina. I fautori di questa prospettiva non forniscono nessuna giustificazione per l'opinione che la filosofia morale non possa affrontarlo. Al contrario, argomentano facendo ricorso a delle preferenze: "ai filosofi piace pensare in termini di astrazione". Invece "agli ingegneri piace pensare in termini di progetti costruibili": solo questi ultimi sarebbero in grado di rivoluzionare lo studio filosofico della mente[18]. I sostenitori delle macchine morali intendono il loro programma come un'iniziativa di liberazione per emancipare l'innovazione tecnologica dall'essere vincolata da un presunto punto di vista prescrittivo.

Solo con l'avvento delle macchine, secondo i sostenitori di questa visione, sarebbero state poste le condizioni per rispondere in modo efficace alla questione di come si deve configurare un agente morale. Chiamiamo questa una nuova posizione del naturalismo morale in relazione alla macchina, che può essere denominata "visione della macchina morale". È necessario discutere questa nuova forma di naturalismo e metterne in dubbio il successo. Tale critica è necessaria perché ritengo che proprio come è avvenuto per le forme di riduzionismo nelle neuroscienze, anche per questa visione non sia possibile ottenere alcun progresso nella comprensione delle funzioni mentali superiori e in particolare di quelle morali. È infatti noto che la definizione dei fenomeni da studiare con metodi empirici non viene esplicitata allo stesso livello e con lo stesso linguaggio usato nelle scienze empiriche. Quindi, se il fenomeno da indagare non viene individuato a partire da un'indagine che attinga alla descrizione dei sistemi complessi in cui questi fenomeni sono coinvolti, è molto difficile che tali indagini siano all'altezza di ciò che si vuole spiegare. In questi casi allora non è possibile ottenere alcun progresso della conoscenza.

Questa stessa critica vale non solo per le indagini neuroscientifiche, ma anche per la costruzione e l'implementazione di sistemi mentali per quelle che sono state felicemente definite scienze artificiali[19]. In altre parole, esiste una sorta di simmetria tra lo studio in vivo del cervello e l'implementazione *in silico* delle sue funzioni. Questa simmetria implica un'interpretazione fisicalista delle strutture morali.

---

[18] W. Wallach, C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, 2008, p. 59.

[19] V. Schiaffonati, *Explorative Experiments and Digital Humanism: Adding an Epistemic Dimension to the Ethical Debate*, in «Werthner», H., Prem, E., Lee, E.A., Ghezzi, C. (a cura di) *Perspectives on Digital Humanism*. Springer, Cham, https://doi.org/10.1007/978-3-030-86144-5_11, consultato il 30.01.2023.

Rivendicare una visione delle macchine morali significa rivendicare una prospettiva naturalistica e fisicalista. Ciò che questa posizione non prende in considerazione è il fatto che le strutture morali e i loro legami con la sensibilità umana godono di un'indipendenza epistemologica. Solo le prospettive che soddisfano questo vincolo e non si discostano dal terreno per cui sono stati sviluppate sono in grado di sviluppare strutture morali e rendere conto della loro normatività[20]. In questo senso, alcuni autori hanno formulato una versione in termini di fondazionalismo delle ragioni[21]. Con questo approccio hanno voluto sia sostenere l'indipendenza della pratica consolidata di dare e prendere ragioni, che è costitutiva per l'interazione, la comunicazione e la capacità di agire umana in generale, risolvendo al contempo la questione dell'epistemologia morale. Al contrario, i tentativi che si ispirano alle macchine morali sono per loro natura interessati a norme codificate e regole morali, a prescindere dal fatto che la loro normatività sia compresa analizzando la comunicazione e le pratiche comunicative.

4. Gli atteggiamenti pregiudiziali sono molto diffusi tra i gruppi umani, con conseguenze significative[22]. Le azioni alla luce dei pregiudizi si traducono in discriminazione e possono contribuire alla divisione della società e a comportamenti ostili. Nell'esperimento disegnato dai ricercatori del MIT e dell'Università di Cardiff sono definite nuove classi di gruppi, come per esempio, il gruppo pregiudiziale, la cui appartenenza si basa su un comune atteggiamento pregiudiziale nei confronti dei gruppi esterni. Si ipotizza che il pregiudizio agisca come un'etichetta fenotipica, consentendo ai gruppi di formarsi e identificarsi su questa base. Utilizzando una simulazione computazionale, è stata studiata l'evoluzione dei gruppi pregiudiziali, in cui i componenti interagiscono attraverso una reciprocità indiretta. È stato anche possibile osservare come si evolvono la cooperazione e il pregiudizio quando la cooperazione è diretta verso l'interno del gruppo. È stata anche considerata la co-evoluzione di queste variabili quando invece l'interazione è con l'esterno del gruppo emulando il possibile pluralismo di una società. La diversità è influenzata da tre fattori, ovvero l'interazione fuori gruppo, l'apprendimento fuori gruppo e il numero di sottopopolazioni. Inoltre, le popolazioni con una maggiore interazione all'interno del gruppo promuovono sia la cooperazione che il pregiudizio, mentre l'apprendimento esercitato sia all'interno che all'esterno promuove la cooperazione e riduce il pregiudizio. I risultati dimostrano inoltre che il pregiudizio non dipende dalla sofisticata cognizione umana e si manifesta facilmente in agenti semplici con intelligenza limitata, con potenziali implicazioni per i futuri sistemi autonomi e

[20] H. Putnam, *Words and Life*, Harvard University Press, 1995.
[21] T. Scanlon, *Being Realistic about Reasons*, Oxford University Press, 2014.
[22] S. Tiribelli, B. Giovanola, *Weapons of moral construction? On the value of fairness in algorithmic decision-making*, in «Ethics and Information Technology», 2022, vol. 24, fasc. 1.

l'interazione umano-macchinica. Da questi chiarimenti sui meccanismi cognitivi alla base dei pregiudizi possiamo ricavare delle indicazioni in grado di informare non solo l'interazione umano-macchinica ma anche quella tra umani per evitare che le macchine diventino troppo simili a noi nei nostri aspetti più esecrabili e noi veniamo trattati alla stregua delle macchine[23].

[23] D. L. Smith, *On Inhumanity Dehumanization and How to Resist It*, Oxford University Press, 2020 e D. L. Smith, *Less Than Human: Why We Demean, Enslave, and Exterminate Others*, St. Martin Press, 2012.