Aldo Pisano

Interrompere l'umano
Bias*, responsabilità e autonomia nell'utilizzo bellico dell'IA*

Abstract: This paper explores ethical implications of AI use in warfare, focusing on two key issues: responsibility and potential limits to human autonomy. The design of autonomous weapons should prioritize transparency and autonomy in decision-making, especially in morally challenging situations. The automation bias highlights the risks of relying on AI as infallible due to its mathematical programming. This bias undermines human deliberation and violates ethical theories at both the metaethical and normative levels. Starting from the HCI model, an ethics by design for AI is necessary, providing support while allowing users to maintain responsibility. Trusting autonomous weapons requires ensuring that autonomy does not compromise human decision-making, preserving the value of ethical choices' complexity and avoiding the reduction of ethics to mathematical generalizations. Users should have the freedom to disregard AI advice and act according to the situation, thereby assuming responsibility.

"Non far sapere mai ad un soldato quanto è disumana la guerra"
Gavin Hood

1. Etiche antiche per guerre moderne

Si potrebbe sostenere che l'IA sia nata in guerra, anzi *per* la guerra. È noto che Alan Turing – il 'padre' dell'Intelligenza Artificiale – abbia inventato una macchina per decriptare codici nazisti durante la seconda guerra mondiale. Ad oggi, lo sviluppo dell'IA vede un'impennata storica. Mentre scriviamo è in corso una vera e propria corsa agli armamenti che oggi vede l'IA come principale simbolo di affermazione politica, sociale, economica. Una guerra che, ormai, avviene fra aziende e a cui recentemente abbiamo assistito con ChatGPT, la volontà di Elon Musk di bloccarne per sei mesi lo sviluppo, le attuali linee UNESCO per lo sviluppo etico dell'IA, il nuovo lancio di Google.

Sono tutti eventi sintomatici di come l'informazione, combinata con l'IA, diventi sempre più capillare con relativo immagazzinamento di enormi quantità di dati[1].

1 Cfr. L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina, Milano 2017; Id., *Pensare l'infosfera. La filosofia come design concettuale*, Raffaello

Mechane, n. 5, 2022 • Mimesis Edizioni, Milano-Udine Web: mimesisjournals.com/ojs/index.php/mechane • ISBN: 9791222307961 • ISSN: 2784-9961

L'elaborazione dei *Big Data* tramite algoritmi è finalizzata a produrre previsioni e raggiungere determinati obiettivi, soprattutto nel caso di impiego da parte di aziende e/o istituzioni sociali e politiche. La precedente affermazione, in qualche modo, è già una definizione di Intelligenza Artificiale che include i tre elementi che la contraddistinguono: dati, obiettivi, azione. Più in generale, L'IA è l'evoluzione più recente dello strumento tecnologico che svolge un ruolo di supporto e/o sostituzione rispetto ad azioni umane noiose, ripetitive o, nel caso dell'ambito bellico, pericolose per la vita umana. Ciò che differenzia l'IA da uno strumento tecnologico tradizionale è il suo essere adattiva, ossia in grado di auto-modificarsi e produrre soluzioni in base a specifici *dataset* che, se non disciplinati, possono produrre discriminazioni e mettere a rischio i principi di pluralismo e giustizia. Recentemente, un consiglio di alti esperti sull'IA della Commissione Europea ha prodotto delle linee guida per un'IA affidabile[2] perché non più considerabile come mero strumento. Si rende dunque necessaria una differenza fra delega (temporanea di un'azione) ed esonero (totale affidamento dell'azione alla macchina). Già a questo punto, entrano strettamente in gioco le armi autonome.

Il nostro punto di discussione è se si possa lasciare completa autonomia alle macchine in ambito bellico e *come e quando* l'IA possa intervenire interrompendo l'azione dell'essere umano[3]. Per farlo, si considerino due aspetti preliminari:

i. l'ambito in cui collochiamo la riflessione è quello del *decision making* e che chiama in gioco non solo teorie etiche, ma anche teorie euristiche dell'azione e dell'intelligenza.

ii. ne deriva una seconda precisazione: i contesti complessi (non-controllati) in cui avviene l'azione *non sono* contesti di rischio (controllati). Nel primo caso aumenta l'incertezza perché aumenta il numero di variabili e l'imprevedibilità dello scenario[4].

Partendo da queste due prime constatazioni, tenteremo di comprendere perché la supervisione umana si rende necessaria. Partiamo dal caso esemplificativo del riconoscimento dell'obiettivo da parte di un drone-killer. Assumiamo una situazione

Cortina, Milano 2020. Si veda anche: L. Manovich, *Cultural Analytics. L'analisi computazionale della cultura,* Raffaello Cortina, Milano 2023.

2 *Orientamenti etici per un'IA affidabile*, Commissione Europea, 8 aprile 2019. Si rinvia anche a: S. Zuboff, *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*, LUISS, Roma 2019. Il tema è qui affrontato sotto il profilo socio-politico e nei rischi sulla privacy e profilazione insiti nei sistemi di sorveglianza e riconoscimento facciale.

3 Il lavoro di McFarlane, seppur datato, risulta essere ancora un punto di riferimento essenziale a misurare l'interazione e l'interruzione uomo-macchina. Si vedano: D. C. McFarlane, K. A. Latorella, "The Scope and Importance of Human Interruption in Human-Computer Interaction Design", *Human-Computer Interaction*, 17(1), 2002, pp. 1–61. D. C. McFarlane, "Coordinating the Interruption of People in Human-Computer Interaction", *Human–Computer Interaction* — INTERACT'99, IOS Press 1999, pp. 295-303.

4 A questo proposito, tornano utili gli studi condotti da Gerd Gigerenzer. Il testo a cui si fa qui riferimento è: G. Gigerenzer, *Perché l'intelligenza umana batte ancora gli algoritmi*, Milano, Raffaello Cortina 2023. Per un *focus* maggiore sui processi decisionali si rinvia a: G. Gigerenzer, *Imparare a rischiare. Come prendere decisioni giuste*, Raffaello Cortina, Milano 2015.

di incertezza con molte variabili: nebbia, distanza etc. etc. In tali situazioni emergono problemi percettivi in cui la presenza di un minimo rumore/interferenza comporta una difficoltà di riconoscimento degli oggetti da parte dell'IA (fig.1). Questo perché l'IA manca di concettualizzazione, non conferisce significato all'immagine, non lavora per schemi simbolici, ma riconosce in base a un calcolo che è solo ed esclusivamente basato su elementi grafici a-semantici: pixel e colori. L'IA distingue tra due oggetti come un pulmino scolastico e uno struzzo, ma *non sa* cosa siano. La mancanza di consapevolezza[5] estromette la macchina dai processi di significazione. Con l'intervento del rumore sull'immagine, l'IA potrebbe *scambiare* il pulmino per uno struzzo. Va da sé che, nell'ambito del *decision making* in un contesto bellico, questo tipo di interferenze comportano danni irreversibili, perché la macchina non 'vede come'[6], ma semplicemente individua obiettivi e riconosce in base ai set di dati con cui è allenata. L'intelligenza umana, d'altra parte, operando per mezzo del funzionamento vicario[7] e grazie a processi di concettualizzazione, riesce a individuare con chiarezza un obiettivo e per cui un pulmino rimane sempre un pulmino.

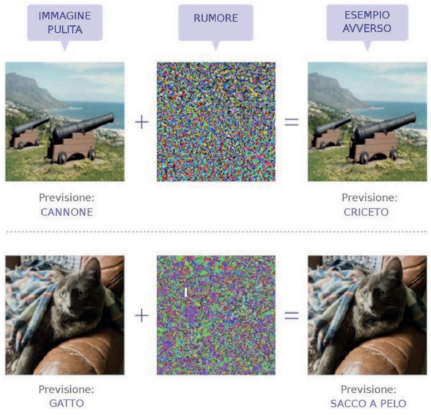

Figura 1 – Esempio di riconoscimento da parte dell'IA di alcune immagini prima e dopo l'introduzione di un rumore. Nel primo caso un cannone viene scambiato per un criceto. Nel secondo caso un gatto è scambiato per un sacco a pelo. Questo per assenza di abilità semantiche e di ca-

5 "la consapevolezza è estremamente importante quando questi sistemi sono autorizzati a prendere automaticamente decisioni che hanno conseguenze sulla vita o sulla morte, come i droni militari, i soldati robot e altre armi letali autonome. Una macchina può sapere benissimo come fare a uccidere, ma non sa cosa sta facendo e perché" [Gigerenzer, *op. cit.,* p. 140].

6 Cfr. L. Wittgenstein, (1953) *Philosophische Untersuchungen*, a cura di P.M.S.-Hacker e J. Schulte, Blackwell, Oxford; tr. it. a cura di M. Trinchero, Ricerche filosofiche, Einaudi, Torino 2009.

7 Gigerenzer sul funzionamento vicario scrive che rappresenta la "flessibilità del cervello nel basarsi su indizi mutevoli, a seconda di cosa sia disponibile". [Gigerenzer, *Perché l'intelligenza umana batte ancora gli algoritmi*, cit., p. 131].

pacità di astrazione che non permette all'IA di concettualizzare. Una sorta di dis-abilità semantica che costringe l'IA a operare un calcolo statistico di volta in volta per riconoscere l'oggetto[8].

La supervisione umana e l'interazione uomo-macchina risultano essere essenziali per arginare i rischi suddetti, e permette di comprendere come sia più funzionale l'utilizzo di un'IA bellica semi-autonoma, anziché autonoma. Si considerino, quindi, le due seguenti tipologie di impiego bellico dell'IA:

I) le armi semi-autonome[9] sono sempre controllate da un operatore umano, quindi godono di un'interazione uomo-macchina continua. Qui la catena di responsabilità risulta chiara anche se i droni rimangono pericolosi, soprattutto nel momento in cui coinvolgono civili[10]. Permane, inoltre, il problema di un *ethics by design*, cioè di una progettazione etica a monte, che non comprometta l'autonomia di scelta di chi sta manipolando il drone. L'essere umano deve poter interrompere l'azione della macchina, la macchina non deve interrompere la scelta dell'essere umano che viene qui assunta in virtù di parametri ben più complessi. Il pericolo che una progettazione non-etica possa far sì che la macchina comprometta e interrompa l'autonomia umana, in forza dei criteri di ottimizzazione, efficienza, correttezza è il problema sollevato proprio da McFarlane[11]. La progettazione di un'Intelligenza Artificiale semi-autonoma passa da alcuni criteri che ne disciplinano i livelli e le modalità di interruzione dell'agire umano. Questo problema è correlato al *bias* di automazione e alla falsa credenza che l'IA riesca sempre nelle scelte eticamente più corrette, perché fondate sulla computazione, quindi su un modello matematico infallibile. Questione che, a sua volta, deriva dalla mancata differenziazione fra rischio e incertezza. In un sistema stabile[12], in cui il rischio può essere calcolato e le variabili poche, la possibilità che l'IA raggiuga efficacemente l'obiettivo è più alto. Il problema rimane traslare l'IA dal sistema stabile al sistema instabile, e mantenendo un accurato livello di previsione e precisione.

II) le armi autonome configurano molti problemi etici mentre aumenta il livello di impiegabilità[13]. Sono meno costose, proprio perché non prevedono una supervisione umana, e sotto il profilo politico costituiscono un'affermazione di forza a

8 L'immagine è disponibile sulla pagina web: https://www.quantamagazine.org/ai-researchers-fight-noise-by-turning-to-biology-20211207/. Ultima consultazione in data 20 aprile 2023.

9 Qui i sistemi ATR (*Automatic Target Recognition*) costituiscono un ottimo compromesso perché individuano direttamente gli obiettivi, ma senza prendere decisioni se non per scelta dell'operatore umano.

10 Cfr. G. Bangone, *La guerra al tempo dei droni. Da Falluja ai terroristi dell'ISIS. La nova frontiera dei conflitti*, Hoepli, Milano 2014.

11 A questo proposito, Floridi ridiscute la posizione di Mafarlane in L. Floridi, *Etica dell'intelligenza artificiale*, cit.

12 Cfr. G. Gigerenzer, *op. cit.*

13 L'impiego onnipervasivo dell'IA, oggi necessita di una riconfigurazione degli ambienti in cui esse riescano a essere produttivamente impiegate. Per questo costruiamo sempre più spazi a misura di IA in un processo che Floridi definisce di 're-ontologizzazione', cioè di riedificazione degli ambienti di vita finalizzati alla cooperazione uomo-macchina. Il che significa creare mondi stabili in cui l'IA operi il più autonomamente possibile.

livello internazionale in riferimento alla lotta agli armamenti[14]. Per introdurre le questioni etiche relative alle armi autonome, è utile Recuperare il *trolley problem* che pone il dilemma morale se sia giusto uccidere cinque persone anziché una tirando la leva di un treno: qui emerge un primo conflitto fra due differenti teorie normative. La teoria deontologica afferma la non-negoziabilità del valore della vita umana, anche di fronte a un elemento quantitativo (salvare più vite è meglio che salvarne una); la teoria consequenzialista indirizzerebbe l'azione verso la scelta che *quantitativamente* soddisfa il maggiore bene possibile (per cui è possibile scegliere di uccidere cinque persone anziché una soltanto). Questo problema diviene evidente nel caso delle armi autonome come nel caso di un drone killer il cui *target* è un terrorista che potrebbe uccidere centinaia di civili (dilemma attorno a cui ruota *Il diritto di uccidere*, film del 2016 citato in esergo al presente articolo). Il dilemma si pone nel momento in cui il drone deve calcolare i limiti di attuazione dell'obiettivo e quindi di esecuzione dell'azione. Questo dipende strettamente dalla programmazione e dalle modalità con cui il drone percepisce l'area circostante, gli eventuali civili coinvolti e altre variabili che compaiono in uno scenario ad alta complessità[15]. Il tutto considerando che nella macchina non intervengano rumori e distorsioni (fig. 1). Vengono qui violate alcuni principi etici sull'uso dell'IA:

a) responsabilità, per cui non sappiamo chi tra 'le molte mani'[16] sia il responsabile dell'arma autonoma;

b) trasparenza, per cui non sapremo perché la macchina ha deciso di eseguire il compito, nonostante fossero implicati i civili (fermo restando che sia stato in grado di riconoscerli).

c) la supervisione umana è venuta meno, esonerando completamente il soggetto dalla responsabilità e quindi privandolo della propria funzione di agente morale.

L'assenza di una situazione che coinvolga direttamente due soggetti umani, un agente e un paziente morale, determina un'asimmetria che riconfigura lo *ius in bello*, non solo perché c'è uno sganciamento dalla responsabilità, quanto anche l'assenza di qualsiasi condizione umana – fondata su empatia e co-appartenenza alla specie – che potrebbe evitare di prendere decisioni impulsive o di eseguire ciecamente un ordine senza valutare il mutare delle condizioni specifiche di complessità. L'IA rimane indifferente alla condizione di sofferenza dell'altro decidendo della sua vita: si pensi a un soldato nemico che si arrende o ferito brutalmente, quindi inabile a portare avanti un combattimento e che, probabilmente, da un agente umano verrebbe risparmiato in quanto non considerato più una minaccia.

14 Cfr. G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma 2020.

15 È il caso dell'italiano Giovanni Lo Porto che, preso in ostaggio da al-Qaida in Pakista, è rimasto vittima di un missile aria-terra lanciato un drone degli Stati Uniti contro un edificio dove si trovava insieme ad altri ostaggi e ai terroristi di al-Qaida (il vero obiettivo dell'attacco).

16 Tamburrini definisce "il problema delle molte mani" l'impossibilità di attribuire responsabilità [Tamburrini, *op. cit.*, p. 90].

In entrambi i casi, armi autonome e armi semi-autonome, si presenta il problema della depersonalizzazione dell'operazione bellica. Allo stesso tempo, un esonero dall'agire è correlato al *bias* di automazione che rischia di accecare l'esercizio del buonsenso e di declassare l'intelligenza umana che, di fatto, rappresenta un sistema di gestione dell'incertezza. Questo esula da modelli matematici, applicabili a condizioni di rischio in cui tutte le variabili sono calcolabili. La confusione fra rischio e incertezza è quello che produce il pericolo di una matematizzazione dell'etica, ossia una riduzione dell'agire a modelli formali in una prospettiva di 'ragione calcolante' che l'IA riporta *in auge*. E lo fa privilegiando una posizione di *etica dimostrativa,* fondata su presupposti matematico-naturalistici, rispetto a un'*etica dialettica* che, al contrario, parte da premesse condivise, utili a preservare legittimamente i principi di pluralismo e democrazia. Rilevare questa differenza è utile per evitare di affidare la nostra società e le nostre scelte ad 'armi di distruzione matematica'[17] in cui saremo sempre più condannati dalle macchine.

2. Matematizzare l'etica: algocentrismo ed euristica

Nel cult cinematografico del 1957 di Sidney Lumet, *La parola ai giurati (12 Angry Men*), dodici uomini si riuniscono per decidere della vita di un imputato. Il giudizio di undici di loro è immediato e opta per una condanna a morte. Solo uno (Henry Fonda) decide di porre un ragionevole dubbio con l'esito di un capovolgimento del verdetto e facendo lentamente emergere gli errori di valutazione e di *bias* degli altri membri della giuria. A compromettere inconsciamente il giudizio dei giurati intervengono anche conflitti, rabbia, ansie e paure personali. Ulteriori aggravanti che intaccano la possibilità di ponderazione lucida della scelta sono: una notevole fretta nello scegliere della vita o della morte, secondo un principio di ottimizzazione dei tempi e del tutto legato a interessi personali; il distanziamento empatico dall'azione e dalla persona pregiudizievolmente condannata. I giurati, infatti, compiono una vera e propria scelta "da remoto" che pretende di essere lucida e generalizzabile. Così, il film di Lumet mette in scena agenti umani e condizioni depersonalizzanti (la distanza dai fatti e dal condannato) che dovrebbero garantire obiettività e certezza nella scelta etica, facilitando il verdetto. Un complesso di tematiche utile non solo a riallacciare la riflessione al pericolo della scelta da remoto già analizzata introducendo le armi autonome, ma anche ad aprire la riflessione sul pericolo di matematizzazione dell'etica.

Ricostruiamo la questione partendo da un punto di vista logico. Come scrivono Boniolo e Vidali "L'argomentazione dialettica nasce dalla necessità di affrontare, con una discussione razionale, ambiti conoscitivi in cui la verità delle premesse non è riconosciuta. Non si tratta di una discussione qualunque, quindi, ma di un con-

17 Cfr. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Penguin, London 2017.

fronto tra posizioni, nel quale si rispetta i principi logici – primo tra tutti il principio di non contraddizione – e non si fa ricorso a scorrettezze argomentative"[18]. Alla luce degli sviluppi legati alla crisi del modello neopositivista di conoscenza e scienza, dalla crisi delle ideologie che "ha riproposto la necessità di discutere più a fondo premesse e valori altrimenti acquisiti come scontati o non bisognosi di valutazione razionale"[19], la democratizzazione dei processi decisionali "che prevede la discussione e la deliberazione basate su argomenti e contrapposizioni di tesi"[20], lo sviluppo comunicativo, la globalizzazione "dei processi di scambio e delle conoscenze"[21] sono tutti elementi che hanno riabilitato il valore della dialettica, qui da intendersi come incontro/scontro argomentativo su soluzioni/conclusioni impossibili da inquadrare come universali e necessarie, ma di volta in volta rispondenti alle esigenze della singola situazione.

Il paradigma della complessità sostiene il valore della scelta nel quadro del situazionismo. Per cui, l'intelligenza umana, riferita alla virtù dianoetica della *phrónesis*, è un sistema di adattamento della risposta allo scenario specifico che non può essere pregiudicata da un insieme di informazioni limitate e ripetitive – cosa che avviene nel caso dell'IA che utilizza *dataset* storici. Dunque, un sistema etico basato sulla dimostrazione logico-razionale, su proposizioni prescrittive, non fornisce risposte adeguate al contesto. Per converso, un'etica basata sulla dialettica, quindi sulla flessibilità del pensiero come apertura a nuove risposte, non condizionata da dati storici preconfezionati, offre l'elemento di discrimine tra intelligenza artificiale e intelligenza umana.

Il rischio dell'IA sta in un ipotetico atteggiamento pregiudizievole da parte dell'essere umano (utente, operatore, programmatore) che potrebbe esonerarsi dall'azione promuovendo l'idea di un'IA infallibile, perché adattiva e autonoma, che opera nel *giusto* in virtù di calcoli *corretti*. C'è qui un problema di fallacia naturalistica. Una proposizione prescrittiva (giusto/sbagliato) è derivata da una proposizione matematico-descrittiva (corretto/non-corretto). L'intelligenza artificiale è al corrente di fatti immagazzinati in un *database*, ma da questi non possiamo derivare dei valori. La proposta antinaturalistica e intuizionistica di Moore[22] tor-

18 G. Boniolo, P. Vidali, *Strumenti per ragionare. Logica e teoria dell'argomentazione*, Bruno Mondadori, Milano 2011, p. 15.

19 Ivi, p. 20.

20 *Ibidem.*

21 *Ibidem.*

22 G. E. Moore, *Principia Ethica*, Cambridge University Press, Cambridge 1993. Un precedente nella storia dell'etica lo troviamo in Hume che, seppure sostenga il ruolo dell'autonomia dell'etica e la non derivabilità dei valori dai fatti, rispetto a Moore mantiene una posizione non-cognitivista. A questo proposito si rinvia a: G. Caracaterra, *Il problema della fallacia naturalistica. La derivazione del dover essere dall'essere*, Giuffré, Milano 1969; G. Siniscalchi, *Esistenza e dovere in G.E. Moore*, Editrice Adriatica, Bari 2004. Una buona ricostruzione sul naturalismo in etica si trova in: G. Bongiovanni (a cura di), *Oggettività e morale. La riflessione etica del Novecento*, Bruno Mondadori, Milano 2007. L'opposizione fatto-valore, descrizione-prescrizione, autonomia-non autonomia dell'etica (nella prospettiva cognitivista o non cognitivista) ha avuto

na utile al mantenimento di un'autonomia della disciplina rispetto al pericolo di riduzionismo a cui oggi andiamo incontro, sbilanciando troppo la nostra fiducia sull'IA. La *open question* mooriana può essere riattualizzata in ottica pluralista, proprio perché permette di non esaurire un valore etico universalizzandolo in una definizione univoca, una descrizione o un calcolo. Se il 'bene' – come qualità e che qui utilizziamo per definire un processo decisionale ottimale per risolvere un conflitto morale – si esaurisse in una in una sola formulazione, decadrebbe il senso stesso dell'agire etico come 'deliberazione intelligente' rispetto a condizioni variabili di incertezza e che assume un soggetto situato storicamente. Affidare in maniera completamente autonoma all'IA decisioni che hanno delle implicazioni etiche (il caso delle armi autonome) è un passaggio critico di chi detiene un tale potere decisionale e che rischia di non scardinare il *bias* di automazione, ma anzi di ridurre la scelta etica a una formulazione logica. L'azione che ricuce insieme esperienza pregressa, conoscenza e analisi specifica della situazione concretizza l'idea stessa della libertà come 'cominciamento'[23] e a cui si accompagna la responsabilità nella non-ripetizione di soluzioni programmate.

Il panorama della matematizzazione – che è quindi un rischio di riduzionismo implementato e sostenuto da una pregiudizievole fiducia nelle macchine – trova anche un altro elemento problematico dal punto di vista etico: i *bias* insiti negli algoritmi, inseriti in fase di progettazione e programmazione dell'IA[24]. Per misurare il livello di pericolo dell'IA come 'arma di distruzione matematica', O'Neil introduce le tre categorie da tenere in considerazione che sono: opacità, scala e danno. Anche in questo caso, torna un approccio essenzialmente basato sul rischio. Tuttavia, questa è una considerazione che seppure nasca da una forma di previsione dei potenziali danni, comunque necessita di una progettazione etica pre-applicazione che:

a) consideri determinati principi inviolabili nella programmazione, quindi con una curvatura deontologica.

b) lasci libera scelta all'essere umano nella fase di interazione, evitando interruzioni dell'autonomia che, in virtù del principio umano-centrico, rimane responsabile ultimo della scelta.

un importante svolta dopo la pubblicazione di Quine del saggio sui due dogmi dell'empirismo: W. L. Quine, *Da un punto di vista logico. Saggi logico-filosofici*, Raffaello Cortina, Milano 2004. Inoltre, di particolare rilievo è l'introduzione di Luca Fonnesu a P. Foot, *La natura del bene*, Carocci, Roma 2007.

23 Cfr. H. Arendt, *The Human Condition*, The University of Chicago, U.S.A. 1958; tr. it. di S. Finzi, *Vita Activa. La condizione umana*, Bompiani, Milano 2014.

24 In merito a questa criticità etica, un'ipotesi di soluzione per un *ethics by design* – quindi per una progettazione etica a monte – torna utile la soluzione che rinvia all'impiego di team interdisciplinari (sociologici, antropologi, eticisti) che supportino la programmazione. Ideale sarebbe l'introduzione di un consulente etico aziendale per l'IA che svolga un doppio ruolo attivo: (a) di consulenza diretta e (b) di formazione del personale sulle problematiche etiche legate all'impiego dell'IA.

c) definisca ruoli di consulenza, informazione e formazione evitando la matematizzazione dell'etica e forme di algocentrismo.

d) utilizzi modelli dinamici e flessibili e non modelli statici.

Modelli dinamici risultano essenzialmente combinati con le situazioni di incertezza e complessità, evitando generalizzazioni e mettendo in luce l'agire umano come modello privilegiato in quanto capace di operare in condizioni ad alto rischio e creando nuove soluzioni, facendo dell'incertezza un punto di forza: il "carburante dell'evoluzione umana"[25].

Una concezione euristica dell'intelligenza e dell'agire permette di evitare il pericolo di un'inversione logica. Se, infatti, l'intelligenza artificiale nasceva come tentativo di imitazione di quella umana, oggi si assiste a un capovolgimento che tenta di modellare l'intelligenza umana sulla computazione, pensandola secondo i criteri di ottimizzazione e velocità di calcolo. Non a caso, il modello HIP della mente umana, la considera come un computer che processa le informazioni. Mentre il modello COGNET (*Cognition as a network of tasks*), utilizzato proprio in ambito militare, è utile per la gestione di più compiti simultaneamente[26]. Esso risulta essere più dinamico e realistico rispetto al modello HIP in quanto preserva il fattore complessità evitando riduzionismi e configurandosi come un modello adottabile nell'ambito dell'interazione uomo-macchina. I modelli, tuttavia, anche se dinamici, rimangono delle formalizzazioni estremamente approssimative della complessità: "Models are opinions embedded in mathematics"[27]. Un'arma autonoma o semi-autonoma programmata su un *dataset* che prevede delle sovra o sotto rappresentazioni (es. persone di colore o con determinati tratti caratteristici che possono essere oggetto di discriminazione comportano una trasformazione del modello in *bias*.

In generale, il bilancio sulle armi autonome, premesso e definito anche la questione della matematizzazione dell'etica, comporta delle valutazioni di rischio non sottovalutabili:

a) In merito alle norme del diritto internazionale umanitario che possono essere violate;

b) In merito al mantenimento della catena delle responsabilità;

c) In merito alla dignità umana.

La supervisione umana e il modello di interazione uomo-macchina tornano utili sia in fase di progettazione che di applicazione dell'IA, soprattutto se consideriamo quanto osservato in precedenza sulla possibile presenza di rumori che distorcono la percezione dell'IA nel riconoscimento degli obiettivi.

3. Interrompere l'umano

25 G. Gigerenzer, *op. cit.*, p. 120.
26 Cfr. McFarlane, Latorella, *op. cit.*, p. 39.
27 C. O'Neil, *op. cit.*, p. 21.

Affidare le operazioni belliche all'IA significa, dunque: (a) tenere conto del principio di autonomia e progettare la macchina in modo che questa risulti non vincolante per le scelte umane; (b) evitare la generalizzazione di scenari complessi a modelli stabili e ripetibili. Per evitare questo rischio, il modello di McFarlane prevede alcune condizioni:

a) considerare le caratteristiche individuali del destinatario dell'intervento;
b) definire le modalità di coordinamento tra destinatario e sistema;
c) significato e finalità dell'intervento;
d) effetti complessivi dell'intervento[28].

A queste si aggiunge la possibilità dell'utente di poter ignorare tutti i consigli offerti e intervenire in maniera adatta alla situazione, garantendo così un'assunzione di responsabilità[29]. Il paradigma HCI presenta operazioni biunivoche nell'interazione fra i due sistemi: le macchine interrompono l'essere umano[30] e l'uomo interrompe la macchina. Il rilievo etico che emerge dal lavoro di McFarlane è la misurazione del limite di interruzione dell'essere umano[31]. Bisogna partire dall'estirpazione del *bias* di automazione, così da evitare che la valutazione della scelta ottimale della macchina (piano descrittivo di uno scenario stabile) si imponga sulla prudenza della scelta umana (piano prescrittivo in uno scenario instabile).

Nel modello proposto da McFarlane e Latorella, il primo elemento da considerare è la consapevolezza dell'interruzione[32]. La capacità umana di simbolizzazione e, soprattutto, il maggiore valore conferito all'agire umano rispetto alla macchina, permetterebbe di bloccare il famoso drone che, a causa di un disturbo, potrebbe decidere di sganciare una bomba non avendo riconosciuto un pulmino scolastico nelle vicinanze scambiandolo per uno struzzo, e quindi calcolandolo come un obiettivo sacrificabile nel contesto di un'operazione bellica. Grazie alla supervisione umana e all'intervento dell'operatore, non solo sarà salvo il pulmino, ma sarà salva anche la responsabilità. La limitazione dell'autonomia dell'IA significa maggiore riconoscimento di responsabilità per l'essere umano. L'ONG britannica 'Article 36' ha individuato la questione delle armi autonome riferendola all'elemento umano e all'interazione uomo-macchina, con utilizzo della locuzione "controllo umano significativo"[33].

Tuttavia, come analizza McFarlane, la questione dell'interruzione chiama in gioco il processo compito-interruzione-ripresa del compito, in cui l'interruzione comporta:

(i) Diminuzione velocità dell'azione;
(ii) Maggiore possibilità di commettere errori;

28 Cfr. D. C. McFarlane, *Human computer Interruption*, cit.
29 Il principio è aggiunto sempre in L. Floridi, *Etica dell'intelligenza artificiale. Sviluppo, opportunità, sfide*, cit.
30 Cfr. D. C. McFarlane, K. A. Latorella, *op., cit.,* p. 19.
31 Cfr. Ivi, pp. 14-15.
32 Ivi, p. 15.
33 Tamburrini, *op. cit.*, p. 105.

(iii) Riduzione dell'efficienza delle persone;
(iv) Aumento del fattore stress[34].

Tutti elementi che – se valutati nel paradigma HCI e delle armi semi-autonome – potrebbero indurre a uno sganciamento della macchina dall'operatore incentivando la produzione e la diffusione di armi autonome. Qui torna la questione delega-esonero. L'esonero è apertura alla deresponsabilizzazione e incentiva una tendenza: generare non solo l'interruzione ma la potenziale sostituzione. Una decisione 'proporzionale' che calcola costi e benefici di una certa operazione bellica permette di distinguere quando dal piano tale operazione si passa al livello dell'agire morale. L'una rispondente a un comando esterno, l'altra a un comando interno a cui, comunque, si è addestrati secondo criteri di un senso comune dell'umanità condivisa: "un bilancio preventivo di questo tipo può a sua volta comportare l'uso di capacità cognitive ed emotive, di competenze sociali ed esperienziali che non sono attualmente alla portata di un sistema di IA e della robotica"[35].

Il modello HCI guarda all'interruzione come potenziale rischio nelle applicazioni di intelligenza artificiale, e formula un "adeguato intervento contestualizzato in ragione del destinatario"[36] e che si concretizza nel raggiungere il "giusto livello di perturbazione, rispettando al contempo l'autonomia tramite le opzioni che offre"[37]. In senso costruttivo, il supporto dell'IA non è solo esecutivo, ma anche *percettivo* e *consultivo*. Percettivo perché, di fatto, superando i limiti del sistema di percezione umana riesce a individuare un *target* anche ad ampio raggio e in tempi minori – come scrive McFarlane: "computers are sometimes built to control physical processes that people cannot or should no control directly"[38]. È inoltre consultivo, in quanto la macchina permette di calcolare velocemente quelle che sono le azioni eseguibili ottimizzando i tempi di valutazione, soprattutto in situazioni, come uno scenario bellico, in cui il tempo della deliberazione è spesso ridotto e la capacità di giudizio annebbiata. Questo sempre in una situazione di arma semi-autonoma con controllo da remoto. Alle possibilità offerta dall'IA all'operatore bisogna sempre aggiungere quella che è una decisione personale *oltre* le scelte offerte dalla macchina. Solo così vengono rispettati i principi di responsabilità, autonomia e libertà dell'azione.

Riprendendo McFarlane, Floridi scrive che la "contestualizzazione trova fondamento nelle informazioni relative a capacità, preferenze e finalità degli utenti, e nelle circostanze in cui l'intervento avrà luogo"[39]. Questo ci pone di fronte a un'annosa questione relativa all'esercizio della libertà umana e che, ancora, si interseca con il *bias* di automazione. Per quanto si possa progettare eticamente un'arma semi

34 D. C. McFarlane, K. A. Latorella, p. 25.
35 Ivi, p. 88.
36 L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 239.
37 *Ibidem.*
38 D. C. McFarlane, A. Latorella, *op. cit.*, p. 38.
39 Ivi, p. 239.

autonoma, seguendo i principi HCI, rimane vincolante il 'principio (in)formazione', in quanto l'operatore umano deve essere a conoscenza:

(i) delle possibilità offerte dal sistema di IA.

(ii) del fatto che il supporto decisionale dell'IA non è vincolante, in quanto non è una scelta necessariamente corretta.

(iii) del fatto che possa scegliere diversamente da quanto suggerito dall'IA.

4. Le buone pratiche: verso un modello sostenibile di armi autonome

Alla luce dei problemi precedentemente evidenziati, emergono alcune *best practice* in merito all'utilizzo bellico dell'IA. Partiamo dal principio della *supervisione umana*, da intendersi come rispetto della sfera di autonomia dell'agente morale e utile a mantenere un livello minimo di controllo umano significativo (CUS) in riferimento a una maggiore personalizzazione dell'agire. La validità di tale principio la si riferisce:

1. alla 'condizione Petrov'[40], per cui l'essere umano deve intervenire come sistema ausiliario di salvaguardia per impedire che il malfunzionamento o comportamento imprevisto del sistema sfoci in attacco diretto contro i civili e i suoi beni o provocando danni collaterali eccessivi;

2. al suo essere catalizzatore di responsabilità;

3. alla garanzia che esso pone del rispetto dignità umana, in momenti in cui questi sono sottoposti a condizioni di vita o di morte;

4. alla garanzia del diritto alla vita, non-negoziabile, e che chiama in causa il *principio di distinzione* della Convenzione di Ginevra[41] definendo i limiti dello *ius in bello*.

5. al *principio di proporzionalità* definito sempre dalla Convenzione di Ginevra, per cui se i costi sono troppo alti, anche per un fine buono, si rende necessario evitare l'operazione bellica. Chiaramente qui il sotto-problema sta nel definire il limite di tale costo.

Il principio della supervisione permette di assumere il modello HCI come schema e teoria dell'interazione a cui ricondurre i limiti di applicabilità dell'arma autonoma considerando

40 La derivazione è dal nome di Stanislav Petrov, responsabile della sicurezza OKO di allarme preventivo contro attacchi nucleari. Questi, nel 1963, si rende conto che c'è un problema di captazione errata di attacco nucleare. Il valore dell'intelligenza euristica e delle funzioni vicarianti, gli permisero quell'esercizio della *phronesis* utile a evitare un contrattacco evitando un conflitto nucleare. Petrov "non nutriva una fiducia completa nel sistema OKO, entrato da poco in servizio e ancora poco noto per pregi e difetti; [inoltre] la sua formazione scientifica, e non esclusivamente militare, lo aveva messo nella condizione di non accettare come dato incontrovertibile il responso della macchina" [G. Tamburrini, *op. cit.*, p. 109].

41 ICRC – INTERNATIONAL COMMITEE OF THE RED CROSS (1977), *Protocol Additional to the Geneva Conventions of 12 August 1949,* ICRC, Geneva.

1. il modello di Latorella dell'IMSM (*Interruption Management Stage Model*)[42], basato su principi teorici e dati empirici dell'interruzione umana in contesti complessi. L'IMSM interpreta l'interruzione dividendola in fasi:

i) Percezione dell'annuncio di interruzione;

ii) Interpretazione dell'annuncio;

iii) integrazione dell'interruzione nel compito in corso d'opera;

iv) ripresa del compito in corso[43].

Il modello torna utile, ancora una volta, per individuare in scala quelli che sono i punti di discrimine fra l'agire autonomo dell'essere umano e il momento in cui questo è oggetto di interruzione.

2. Le modalità di interruzione individuate da McFarlane e che possono scandire il rapporto con la macchina:

(i) interruzione immediata;

(ii) interruzione negoziata;

(iii) interruzione mediata;

(iv) interruzione programmata[44].

Individuare le modalità di interazione uomo-computer con una scala modulata è già un primo modo per definire il limite di interruzione, per preservare l'autonomia decisionale dell'operatore umano, ma soprattutto per evitare la completa sostituzione. Questo non solo destituisce l'agente dalla sua dignità etica, ma anche il paziente morale: la depersonalizzazione priva entrambi gli attori morali della loro condizione umana e dei loro diritti. È il caso dell'interruzione immediata del compito che non lascia margine di scelta all'essere umano con conseguenze etiche disastrose e un alto rischio di irreversibilità degli effetti prodotti. Per questo, come propone Floridi in riferimento alla progettazione etica e partendo dagli studi di McFarlane: "I designer di AI4SG dovrebbero costruire sistemi decisionali in dialogo con gli utenti che interagiscono con questi sistemi e ne sono influenzati; sulla base della comprensione delle caratteristiche degli utenti, delle modalità di coordinamento, delle finalità e degli effetti di un intervento; e nel rispetto del diritto degli utenti di ignorare o modificare gli interventi"[45].

In merito al *principio responsabilità*, possiamo invece considerare alcuni interventi fondamentali, partendo dall'analisi di Tamburrini e O'Neil. A monte, risulta sempre ottimale una limitazione dell'autonomia della macchina, e quindi dell'interruzione umana, considerando il contesto di applicazione dell'IA. Su questo tema sono stati firmati molti appelli sin dal 2012, ad esempio quelli di chimici e fisici contro lo sviluppo di armi batteriologiche[46]. I diversi sistemi di *Governance* è

42 D. C. McFarlane, A. Latorella, *op. cit.*, p. 15.

43 Cfr. Ivi, p. 17.

44 Ivi, p. 25.

45 L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 239.

46 Nel 2013, ad esempio, una ONG lancia la campagna SKR (*Stop Killer Robots*) per arrivare alla stesura di un trattato internazionale contro le armi autonome. Ancora, nel 2015 una lettera aperta chiede di fermare lo sviluppo delle armi autonome all'*International Joint Conference of*

anche necessario che sensibilizzino e avviino una formazione sui temi etici, eventualmente sostenendo la costituzione di team interdisciplinari che cooperino alla progettazione etica dell'IA in ambito bellico (e non solo). Ne deriva, un'*ethics by design* che definisca le responsabilità, in riferimento a intenzione, conoscenza e imprudenza – elementi necessari ad attribuire responsabilità in riferimento a un crimine di guerra[47].

Sotto un profilo più pratico-operativo il sistema *blockchain* è una strategia utile a fissare e individuare la responsabilità, controllando e registrando ogni momento della catena decisionale con sistemi di connessioni fra i dati (passaggi decisionali) crittografati e irreversibili. Inoltre, sempre in riferimento alla responsabilità, riduce o annulla totalmente il livello di confusione e permette riconoscimenti univoci evitando il 'problema delle molte mani'. Non da ultimo, tale sistema permette di evitare un altro subdolo meccanismo di deresponsabilizzazione morale, in quanto evita di 'scaricare' accuse nei confronti della complessità delle tecnologie utilizzate.

Le ipotesi di soluzione individuate costituiscono delle buone pratiche funzionali a regolamentare l'utilizzo delle armi autonome, evitando estremismi che le demonizzino o che, al contrario, ne prevedano un utilizzo indiscriminato. Le raccomandazioni europee, che non possiedono una forza giuridica vincolante, costituiscono una fase di passaggio funzionale a considerare i rischi e i danni eticamente irreversibili che comporterebbe l'uso incontrollato dell'IA. Si rende necessario un approccio interdisciplinare, in modo che aspetti più tecnici come il modello HCI possano entrare in dialogo con le riflessioni etiche (normative, applicate, descrittive). Tale indirizzo di lavoro è utile ad arginare:

1. macro-rischi relativi all'applicazione dell'IA e che si riferiscono al pericolo generale di matematizzazione dell'etica, attuando forme di riduzionismo;

2. meso-rischi, qui legati all'ambito bellico, in cui potrebbero non essere rispettati principi di proporzionalità, distinzione e il diritto alla vita;

3. micro-rischi per preservare la dignità etica individuale, garantendo autonomia e responsabilità nella relazione intersoggettiva che intercorre fra i due soggetti di uno scenario bellico.

Di fatto, il rischio della de-legittimazione etica aperta dal riduzionismo e dal *bias* di automazione presenta la responsabilità non come un onere etico. Al contrario, essa sostanzia la soggettività, valorizza la condizione umana in quanto capace di accogliere, gestire e rispettare lo scenario della complessità nelle nuove interazioni con gli agenti artificiali.

Artificial Intelligence [IJCAI]. Su questo tema la ricostruzione di Tamburrini è molto interessante: G. Tamburrini, *op. cit.*, pp. 77-82.

47 Cfr. G. Tamburrini, *op. cit.*, p. 91.